Final Technical Report

Project Title:	"A Multi-scale, Multi-Model, Machine-Learning Solar Forecasting Technology"
Covering Period:	Feb 1st, 2013 to Feb. 28th 2017
Approved Period:	Feb 1st, 2013 to Feb. 28th 2017
Submission Date:	May 31st, 2017
Recipient:	IBM TJ Watson Research Center P.O. BOX 218 YORKTOWN HEIGHTS NY 10598-0218
Website (if available)	www.research.ibm.com
Award Number:	DE-EE0006017
Project Team:	National Renewable Energy Lab (NREL) Argonne National Laboratory ISO New England (ISO-NE) California ISO (CAISO) Tucson Electric Power (TEP) Green Mountain Power (GMP) Juwi Solar, Inc. 3Tier, Inc./Vaisala Florida Gulf Coast University (FGCU) Northrop Grumman (NG) Northeastern University (NE)
Principal Investigator:	Dr. Hendrik F. Hamann; Distinguished Researcher Phone: 914-945-2430 Email: hendrikh@us.ibm.com
Submitted by: (if other than PI)	Susan Scott; Senior Project Manager Phone: 720 300 5863 Email:suescott@us.ibm.com
DOE Project Team:	DOE Program Manager - Guohui Yuan DOE Project Officer – Thomas Rueckert DOE Award Administrator – Fania Kate Gordon FOA Manager – Dr. Anastasios Golnas
Business Contact:	Carl E. Taylor**;** Senior Government Contracts Manager Phone: 713 797 4625 Email: cetaylor@us.ibm.com
Technology Manager:	NA
Project Officer:	NA
Grant Specialist:	NA
Contracting Officer:	NA

1. **Executive Summary**

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.[1,2] At these high levels of penetration, solar power forecasting will become very important for electricity system operations because it is the *least expensive* way to integrate larger amount of solar energy into the electric grid. However, solar forecasting is a very difficult task with different challenges for transmission and distribution networks and inaccuracies can result in substantial economic losses and power system reliability issues because electric grid operators must continuously balance supply and demand.

The goal of the project was the development and demonstration of a significantly improved solar forecasting technology (short: *Watt-sun*), which leverages new big data processing technologies and machine-learnt blending between different models and forecast systems. The technology aimed demonstrating major advances in accuracy as measured by existing and new metrics which themselves were developed as part of this project. Finally, the team worked with Independent System Operators (ISOs) and utilities to integrate the forecasts into their operations.

The technical thrust of the work lies in the idea of injecting state-of-the-art big data machine-learning to the field of meteorology and solar forecasting. To put the achievements of this project into perspective, numerical weather prediction (NWP) models have been improving forecasting accuracies by (only) ~6% *per decade* (basically by refining the physics of the forecasting models as well as improved data assimilation techniques[3]. Key accomplishments of this project are:

- A full suite of metrics (including economic and reliability ones) for measuring the accuracy of solar forecasts was established, which enables grid operators to assess the accuracy of different forecasting systems in a consistent and scientific sound manner.[4,5]
- Methods for deriving “baseline” and “target” values for those metrics were developed, which provide guidance to system operators on what forecasting accuracies can be expected from a standard as well as state-of-the art forecasting system.[6,7]
- A new method (*Watt-sun*) for solar forecasting was developed, which leverages big data technologies and a novel machine-learning approach (called situation-dependent, multi-model blending).[8]
- Demonstrated with the *Watt-sun* forecasting system improved forecasting accuracies in average by more than 100% over baseline (or by > 30% compared to the next best forecast system/model) at multiple locations for point, regional and continental forecasts as measured by the suite of metrics (for all forecast horizons from 15 mins to 48 hours ahead).[9]
- A “open” replicate of the *Watt-sun* forecasting system was created at the National Renewable Energy Laboratory (NREL) ensuring that it can be continued to be used for the larger public good.
- Operational day-ahead forecasts in various forms were provided to the ISO-New England and Green Mountain Power throughout the last two years of the project.
- Team won the 2017 Utility Variable-Generation Integration Group (UVIG) Achievement Award “For major contributions to advancing the state-of-the-art of solar energy forecasting.

Contents

1. Executive Summary 2
2. Project Objectives 3
3. Project Results and Discussion 6
3.1 Metrics for Assessing the Accuracy of Solar Forecasting 6
3.1.1 Development of Basic Metrics 6
3.1.2 Baseline and Target Values 9
3.1.3 Baseline and Target Metrics Values for Test Sites 12
3.1.4 Reliability Metrics 14
3.2 The Watt-sun Forecasting System 15
3.2.1 Overview of the Watt-sun Forecasting Method 16
3.2.2 Categorization of Weather Situations 17
3.2.3 Machine Learning Models 21
3.2.4 Forecasting Error Reduction 23
3.2.5 Results from 5 Test Sites 28
3.2.6 Comparison to ECMWF 31
3.2.7 Nation-wide Solar Forecasting 31
3.2.8 3D Radiative Transfer 34
3.2.9 Short-term forecasting 37
3.3 Integration of the Watt-sun Forecasting System 38
4. Significant Accomplishments and Conclusions 40
5. Inventions, Patents, Publications 41
5.1 Full papers 41
5.2 Conferences 41
5.3 Patents 44
5.4 Press (Selected) 44
5.5 Awards 45
6. Path Forward 45
7. References 46
8. Glossary 48

2. Project Objectives

Solar forecasting will become an integral part of the energy future as increasingly renewable energy is becoming online. Therefore, the project will have significant impacts to the national goals of clean energy progression of the US.

1. The project yielded the *first* consistent set of methods for measuring, comparing and assessing the accuracy of a solar power forecasts, which is not only critically important as solar forecasting information is being integrated in power system operations but also for gauging the technical progress in this field.
2. A novel approach for solar forecasting was invented by combining traditional forecasts with state-of-the-art big data machine learning. The validity of the approach was demonstrated and piloted with utilities and ISOs. The improvements in forecasting accuracy will enable much more cost-effective operations of the power grid.

The project had three main tasks with the following objectives:

1. Task#1: To develop a suite of metrics (statistical, uncertainty quantification, ramp characterization, economic, and reliability ones) for assessing the accuracy of solar forecasting for the industry and to evaluate the performance of these metrics; this included developing methods for determining proper baseline and target values for such metrics.
2. Task#2: To develop a new approach to solar forecasting (*Watt-sun*) which improves accuracy of solar forecasting by >100% above a baseline; this task included evaluation of *Watt-sun* at five test sites using the metrics as developed in task #1.
3. Task#3: To integrate these forecasts into the operations of at least one ISO and one utility and demonstrate benefits to these end-users.

Table 1 provides a more detailed view of the milestones and deliverables of this project organized by the tasks (task#1 to 3 are in red, blue and green respectively). Go/No-Go milestones are in bold. Table 1 also shows the budget period (BP). More prescriptive information about the different tasks can be obtained from the quarterly reports and the Statement of Project Objectives (SOPO).

BP	Task	Short Description
1	1.1.1.2	Development of Deterministic Metric Suite
1	1.1.1	A suite of generally applicable, value-based metrics
2	1.1.2A	Develop a detailed plan/process how to quantify the benefits of the *Watt-sun* system
2	1.2.2A/1.3.2A	Baseline and target values for each metrics including economic ones
2	1.1.2A	Report on the benefits of the *Watt-sun* system to the ISO, utility, and energy producer
2	1.1.2B	White paper and submission to a peer-reviewed journal on metrics development
2	1.2.2B/1.3.2B	White paper and submission to a peer-reviewed journal article on target/baseline values.
3	1.1.1 / 1.1.2	Present results from a simulation study of a high penetration solar in the FESTIV modeling environment to evaluate reliability and economic impacts of better solar power forecasts.
3	1.1.1 / 1.1.2	Demonstrate and quantify measurable improvements in power system reliability metrics (ACE, AACEE, CPS2 scores) and reserve levels (economic metric) to maintain reliability levels from improved solar power forecasts in high penetration solar scenarios
1	2.1.1	Complete infrastructure of *Watt-sun* system
1	2.1.1	Demonstration of operational forecasting
1	2.2.1	Trained categorization/machine learning algorithms,
1	**2.2.1**	**Demonstrate at least 33 % forecasting improvements**
1	2.2.1	Provide initial feedback to NOAA
1	2.3.1	Identified at least two different, geographically diverse test sites
2	**2.1.2/2.2.2/2.3.2**	**Demonstrate at least 50 % forecasting improvements**
2	2.1.2/2.2.2/2.3.2	A report describing the architecture of the 2nd Gen *Watt-sun* system
2	2.1.2/2.2.2/2.3.2	Detailed DoE deep dive webinar/presentation on the *Watt-sun* architecture
3	2.1.2.3	Demonstrate superiority of the developed 3D radiative transfer model vs. a state-of-art 1D radiative transfer model using NAM inputs and SurfRad, ISIS and ARM validation data and using metrics developed as part of Activity A. Validation will be performed with > 6 months of data for each individual validation site and results presented for each site separately.

3	2.3.1	Provide publically available irradiance 0 to 48 hour ahead forecast (accessible via a web page) from the *Watt-sun* system. The forecasts will have a spatial resolution of 0.05 degrees and cover the entire continental US with a temporal resolution of 1 hour.
3	2.3.2	Replicate the *Watt-sun* system in a public cloud environment and train/enable NREL personnel to provide forecasting to all five test sites (GMP, Smyrna, TEP, CAISO and ISO-NE). This milestone includes showing replicability and scalability of the *Watt-sun* system without proprietary technologies. The replication includes all components of the *Watt-sun* system and will be a "stand-alone" system.
3	2.1.3	Demonstrate at least 100 % improved forecasting (towards all base target metrics as developed in activity A during 2nd budget period) and (in addition) less than 8 % normalized root mean square error for all time horizons at all five different test sites (GMP, Smyrna, TEP, CAISO and ISO-NE) using 3rd Gen *Watt-sun* system. Demonstration includes providing forecasts, validation and verification. Forecasts horizons will range from at least 15 minutes to 48 hours ahead with an interval of 15 minutes or shorter. The 3rd Gen *Watt-sun* system has updated modules of (1) a big data bus, (2) a radiative transfer module, (3) a radiance to power module, (4) an information blending module, and (5) a categorization/ machine learning module. Progress towards target metrics will be measured as relative improvement ((A-B)/(T-B) with A as the achieved, B the baseline, and T the target value for a given base metric as developed in activity A). The performance of the *Watt-sun* system will be also evaluated using enhanced metrics, which will be developed in this budget period including economic and reliability metrics as well as compared to analog ensemble forecasts used by the forecasting industry.
3	2.3.4	Demonstrate more than 30 % improvements over "corrected" ECMWF based solar forecasts (including the ECMWF forecasts in our blend) and by more than 15 % without the ECMWF for all time horizons and metrics for all test sites (GMP, Smyrna, TEP, CAISO and ISO-NE). For individual cases (point forecasts) we will demonstrate more than 35 % improvements over "corrected" ECMWF based solar forecasts.
3	2.1.2.6	Publication or detailed report describing the architecture of the 3nd Gen *Watt-sun* system including how to interface (input/output) using standard, open data formats (GRIB2, netCDF, HDF, XML etc.) so that other models can be incorporated.
3	2.1.2.7	Detailed DoE deep dive webinar/presentation on all the aspects of the *Watt-sun* architecture, machine-learning, and all other associated elements. This includes how to interface (input/output) with the *Watt-sun* system using standard, open data formats (GRIB2, netCDF, HDF, XML etc.) so that other models can be incorporated and/or the system can be customized. This deep dive will be structured such that anyone viewing this webinar/presentation will be able to gather the necessary knowledge to reconstruct the *Watt-sun* architecture, and be able to create forecasts upon feeding various model data to *Watt-sun*. This deep-dive will be recorded and made available for public dissemination.
3	2.3.3	Comprehensive publication of the architecture and all methods and procedures used in the *Watt-sun* system; this includes detailed results from the field tests.
1	3.1.1	Comprehensive set of use cases for the integration
2	3.1.2	Fully working instance of the *Watt-sun* technology
3	3.1.3	Successful integration of *Watt-sun* at the ISO-New England and Green Mountain Power and other stakeholder with the forecasts being used in operations for more than 12 months providing tangible benefits. Success is gauged by public feedback from the utility and ISO partner(s). This includes showing tangible improvements to the load forecasts for the ISO-New England of more than 20% per unit solar penetration (for example, this means that we will demonstrate 2% relative load forecast improvements for 10% solar penetration). Furthermore, benefits are measured by the set of economic metrics (see task 1)
3	3.1.4	Development of detailed business plan/strategy for the *Watt-sun* technology to ensure that the *Watt-sun* will be further developed and maintained after the project has ended.

Table 1: Summary of all milestones and deliverables organized by budget period (=BP).

3. Project Results and Discussion

3.1 Metrics for Assessing the Accuracy of Solar Forecasting

3.1.1 Development of Basic Metrics

A key gap in developing solar forecasting models was the unavailability of a consistent and robust set of metrics to measure and assess forecasting accuracy. Previously, each person (forecast provider, system operator etc) used its own metrics to describe the forecasting accuracy. Furthermore, it was not clear that the existing metrics (such as mean absolute error) were very suitable for power system operators considering that the predictability of large events (e.g., ramps) is much more relevant to the electric grid than mean deviations. To develop a consistent set of metrics addressing the needs of power system operations three workshops were held, where feedback and guidance from stakeholders was obtained: (i) 93rd American Meteorological Society Annual Meeting: Solar Forecasting Metrics Workshop, Austin, Texas (2013); (ii) UVIG Workshop on Variable Generation Forecasting Applications to Power System Planning and Operations: Solar Forecasting Metrics Workshop, Salt Lake City, Utah (2013); (iii) UVIG Workshop on Variable Generation Forecasting Applications to Power System Planning and Operations: Solar Forecasting Metrics Workshop, Tucson, AZ (2014).Table 2 shows a summary of the metrics developed in this project, which includes statistical, uncertainty quantification, ramp characterization and economic ones, which are now discussed in more detail.

3.1.1.1 Statistical Metrics

The distribution of forecast errors is a graphical representation of the raw forecasting error data, which provides a good overview of the performance of forecasts for longer time periods. In addition, interval forecasts of solar power can help determining 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.[5,10,11] In this project, the distribution of solar power forecast errors was estimated using the KDE method.

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) rttot distribution.

The Kolmogorov-Smirnoff integral (KSI) and OVER (part of the KSI which integrates above (over) the Kolmogorov-Smirnov critical value) metrics were originally proposed by others.[12] 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 cumulative distribution functions (CDF). Instead of comparing forecast error 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 root mean square error (RMSE), mean absolute error (MAE), maximum absolute error (MaxAE), and mean bias error (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 (i.e., KSIPer = 100*KSI and OVERPer = 100*OVER).

Type	Metric	Description/Comment
Statistical Metrics	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

Table 2: Suite of metrics for solar power forecasting. A smaller value indicates a better forecast for most of the metrics, except for Pearson's correlation coefficient, skewness, kurtosis, distribution of forecast errors, and swinging door algorithm.

3.1.1.2 Metrics for Uncertainty Quantification and Propagation

Two metrics were used 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 were proposed based on the use of concepts from information theory, which can utilize all the information present in the forecast error distributions.[13,14] This information entropy

approach based on Rényi entropy was adopted here to quantify the uncertainty in solar forecasting, where generally, a larger value of Rényi entropy indicates a higher uncertainty in the forecasting.

3.1.1.3 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.[15] Naturally, different temporal and spatial scales influence the severity of up- or down-ramps in solar power output. In this project, the swinging door algorithm was used to identify ramps over varying time frames because of its flexibility and simplicity.[16,17]

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". The width of the door directly characterizes the threshold sensitivity to noise and/or insignificant fluctuations to be specified. With a smaller door, many small ramps will be identified; with a larger door, only a few large ramps will be identified.

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
Corr. 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	2,711.31	1,488.28
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	4,136.96	2,476.55
NRMQE	0.33	0.26	0.20	0.13	0.16	0.09	0.06	0.04
MaxAE (MW)	84.10	74.33	2,260.94	1,304.73	3,380.28	1,735.24	17,977.53	16,127.32
MAE (MW)	14.81	11.34	286.65	191.17	413.11	256.69	1,973.90	1,064.52
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	1,497.29	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
Std dev. (MW)	21.65	39.57	418.00	282.62	599.94	376.20	2,260.09	1,482.44
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 % (MW)	50.59	39.57	990.66	637.45	1,394.85	838.27	5,652.60	3,079.32
Capacity (MW)	100.00	100.00	3,463.00	3,463.00	6,088.00	6,088.00	6,4495.00	6,4495.00

Table 3: Metrics values by using an entire year of WWSI-2 data (see explanation in the text).

3.1.1.4 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 period 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.

3.1.1.5 Evaluation and testing of metrics

The suite of metrics as summarized in Table 2 was first tested using a data set 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.[18,19] This study included solar data based on a 1-minute interval using satellite-derived, 10-km x 10-km gridded, hourly irradiance data as well as 60-minute solar power plant output data. The solar power output data comprised distributed generation rooftop photovoltaic, utility-scale photovoltaic, and concentrating solar power with thermal storage. In addition, the WWSIS-2 data included day-ahead solar forecasts, which were produced by 3TIER based on NWP simulations. The 1-hour-ahead forecasts were synthesized using a 1-hour-ahead persistence-of-cloudiness approach.

Four scenarios were analyzed: (1) for a single solar power plant with a 100-MW capacity; (2) 46 solar power plants near Denver, Colorado, with an aggregated 3,463-MW capacity; (3) 90 solar power plants in the state of Colorado with an aggregated 6,088-MW capacity; (4) solar power plants in the entire Western Interconnection in the United States, including 1,007 solar power plants with an aggregated 64,495-MW capacity. The evaluation included a sensitivity analysis, e.g. how would the metrics change if the forecasting accuracy would increase: (i) uniform improvements excluding ramping periods; (ii) ramp forecasting magnitude improvements (iii) ramp forecasting threshold changes. By way of example, using the WWSIS-2 data, the values for different metrics are reported in Table 3 for the four geographical scenarios. Uncertainty metrics for the four geographical scenarios are shown in Table 4.

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
4.83	4.64	4.24	4.63	4.33	4.73	4.47	4.01

Table 4: Uncertainty metrics for the four geographical scenarios.

The main impact (for the major forecasting improvement) and the total impact (for all forecasting improvements) for each metric is 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 improvement (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.

3.1.2 Baseline and Target Values

To properly gauge the quality of solar forecasts it is important to establish a baseline as well as an appropriate target, which can be expected from such an improved forecast. Evidently, this is not a trivial task given that the accuracy for forecasting is highly dependent on location, time of

year, forecasting horizon, spatial extent and other factors. Generally, a baseline model is used for comparison, which is selected from: (i) persistence models [20,21,22] ;(ii) numerical weather prediction (NWP) models without bias correction[23,24] ;and (iii) NWP models with bias correction.[25,26]

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 5: Sensitivity analysis of metrics to three types of forecasting improvements (see text for details).

In this project and as shown in Table 6, we used two different methods to establish a baseline for short-term and long-term forecasts, respectively. For the short-term, we adopted a smart persistence approach[6,7,27], while for the long-term a NWP model (here the North American Mesoscale Forecast System (NAM)[28] was used to obtain the atmospheric conditions. The reason for using the NAM model only for longer term-forecasts is due to the fact that NWP models rarely achieve useful skill at lead times smaller than a few hours because of the (spin-up) period they require to achieve numerical stability. The output from NAM was fed to a two-streamer Radiative Transfer Model (RTM)[29] and the PVLib tool box[30] to derive the solar power forecasts. To remove substantial bias errors, a first order machine learning (linear regression model) model is applied based on the data from the previous three days.

Forecast Horizon	Weather Information	Irradiance Forecasts	Power Forecasts
15-min-ahead, 1-hour-ahead, and 4-hour-ahead	Persistence	Streamer RTM	Persistence of cloudiness
Day-ahead up to 48 hours	NAM	Streamer RTM	(1) PVLib + linear regression; or (2) linear least square fit (if no PV specifications available)

Table 6: Overall approach to determining baseline forecasts at different forecast horizons.

The forecasting is divided into two parts: non-ramping period and ramping period. The target values for solar forecasting metrics are derived by the following procedure: (i) for the non-ramping period, applying uniform forecasting improvements by *x%* based on the baseline forecasting; (ii) for the ramping period, applying ramp forecasting improvements by *y%* based on the baseline forecasting; and (iii) deriving a complete set of target metrics. The values of *x%* and *y%* are determined based on the economic impacts of improved solar power forecasting (i.e., a reduction of 25% in reserve levels. This level was confirmed by our ISO and utility partner.

3.1.2.1 Flexibility Reserves for 15MA, 1HA, 4HA, and DA Forecasting

The reduction in the amount of reserves whether this is for 15 minutes ahead (15MA), 1 hour ahead (1HA), 4 hour ahead (4HA) or day ahead (DA) that must be carried to accommodate the uncertainty of solar power output is anticipated to be one of the significant cost savings associated with improved solar power forecasting. Following previous work[31,32], improved forecasting (on average) reduces the amount of reserves that must be held. More specifically, the various types of flexibility reserves are defined by:

For 15MA, 1HA, and 4HA solar power forecasting, spinning reserves are used to derive the target solar forecasting values. Spinning reserves represent the online capacity that can be deployed very quickly (seconds to minutes) to respond to variability. The spinning reserve for 0- to 4-hours-ahead forecasting (R_s^{HA}) is defined as the 95% confidence interval ($\emptyset_{95}$) of solar power forecast errors (e^{HA}) at the 15MA, 1HA, or 4HA horizon. $R_s^{HA} = \emptyset_{95}(e^{HA})$

For DA solar power forecasting, both spinning and non-spinning reserves are used to derive the solar forecasting target. Non-spinning reserves represent the off-line or reserved capacity, or load resources (interruptible loads), capable of deploying within 30 minutes for at least 1 hour. The spinning reserve for the DA forecasting (R_s^{DA}) is defined as the 70% confidence interval ($\emptyset_{70}$) of the DA solar power forecast errors (e^{DA}).[31] The non-spinning reserve (R_{ns}^{DA}) is defined by the difference between a 95% confidence interval ($\emptyset_{95}$) and a 70% confidence interval ($\emptyset_{70}$) of the DA solar power forecast errors (e^{DA}): $R_s^{DA} = \emptyset_{70}(e^{DA})$ and $R_{ns}^{DA} = \emptyset_{95}(e^{DA}) - \emptyset_{70}(e^{DA})$. To estimate the economic benefits it was assumed that the cost of non-spinning reserve per MW (C_{ns}^{MW}) is twice the cost of spinning reserve per MW (C_s^{MW}) $C_{ns}^{MW} = 2 \times C_s^{MW}$, which includes (i) start-up costs of two types of generators used for spinning and non-spinning reserves (gas turbine and oil turbine); and (ii) heat rates and fuel costs of four fuel types (biomass, nuclear, coal, and combined cycle). The costs were selected according to the ISO-New England system.[7]

Figure 1: Locations of the three point and two regional test sites.

3.1.2.2 Test Sites: System Operators, Utilities, and Energy Producers

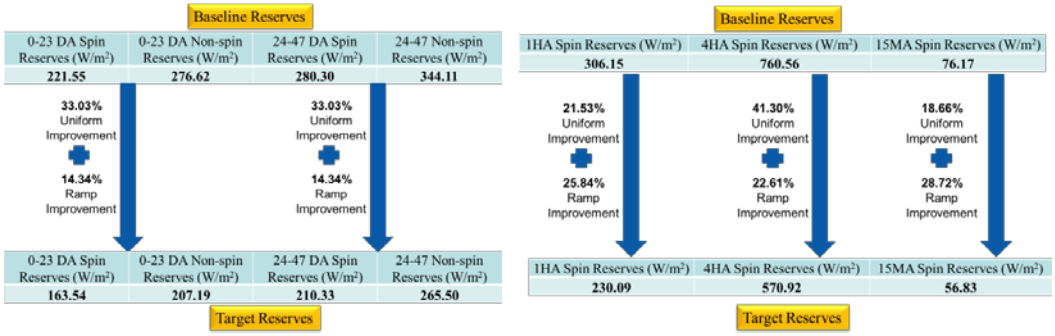
Three PV plants were chosen among hundreds of sites available for which the Watt-sun system is forecasting (see Figure 1): Smyrna, Green Mountain Power (GMP), and Tucson Electric Power (TEP). In addition, two regional test cases, ISO-New England (ISO-NE) and California-ISO (CAISO) were chosen to cover two distinct atmospheric conditions: a cloudier and more humid climate for the ISO-NE region, in contrast to relatively drier climate in the CAISO region.

3.1.3 Baseline and Target Metrics Values for Test Sites

The above mentioned five test sites were chosen to determine baseline and target metrics. Further below we will also compare the forecasting results from the *Watt-sun* syste. For brevity, we show here only the results for baseline and target values for 2 of those 5 test sites (ISO-NE and GMP). Additional results from the other three test sites were reported in the detail elsewhere.[27]

Test sites	Role	Forecast Horizon		Validation	Evaluation Period
		15MA, 1HA, and 4HA	Day-ahead		
ISO-NE	System operator	Persistence	NAM	GHI –12 MesoWest sites	03/05/13 – 30/10/13
CAISO	System operator	Persistence	NAM	Aggregated Power	04/05/13 – 30/10/13
GMP	Utility	Persistence	NAM	Direct Power measurements	03/05/13 – 30/10/13
TEP	Utility	Persistence	NAM	Direct Power measurements	02/06/13 – 30/10/13
Smyrna	Energy producer	Persistence	NAM	Direct Power measurements	03/05/13 – 30/10/13

Table 7: Test sites of system operators, utilities, and energy producers.



(a) Day-ahead forecasts at ISO-NE (b) 15MA, 1HA, and 4HA forecasts at ISO-NE

Figure 2: Target reserves values based on uniform and ramp forecasting improvement (ISO-NE)

3.1.3.1 ISO-NE Baseline and Target Metrics Values

For ISO-NE, solar generation is mostly behind the meter and interconnected to the distribution system. Therefore, the value of an improved forecast technology will lead to improved net load (i.e., the load minus the PV) forecasts – especially for the day-ahead unit commitment process. Therefore, the metrics were calculated based on solar irradiance instead of power. The baseline and target metrics for ISO-NE are summarized in Table 8. The capacity used for normalization is 1000 W/m^2. DA (both 0-23 and 24-47 hours ahead) baseline forecasts performed better than the 4HA baseline forecasts, which can be partially attributed to the cloudy weather of ISO-NE region; the cloud movement significantly affects the persistence forecast. It is important to note that for the ISO-NE case, the irradiance is calculated by averaging a set of sites. However, since the available sites are closely located in a small geographical region, their irradiance values are correlated.

Figure 2 shows the baseline and target reserves values (in terms of irradiance) at different forecast horizons. To achieve the target reserves, there is more ramp forecasting improvement required than uniform improvement for DA and 4HA forecasts, and there is more uniform improvement than ramp forecasting improvement for shorter timescale forecasts (1HA and 15MA) required. Figures 3(a) and 3(b) illustrate the distributions of solar power forecast errors for ISO-NE baseline and target forecasting, respectively.

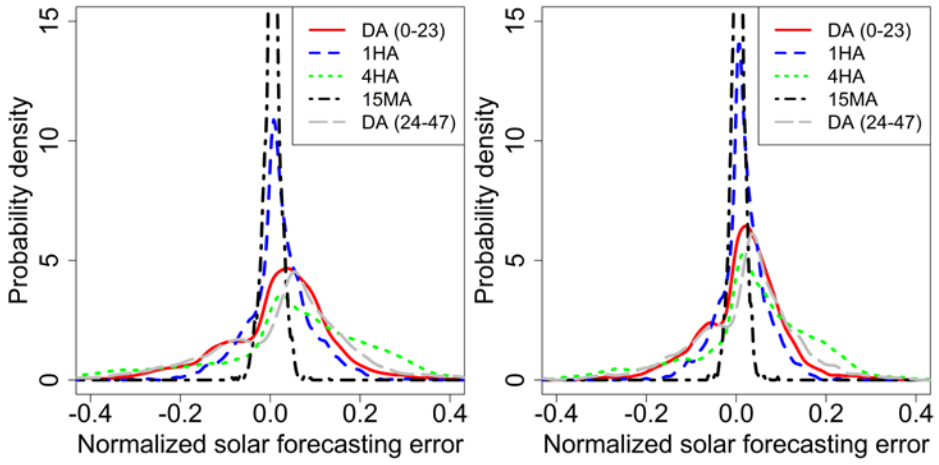


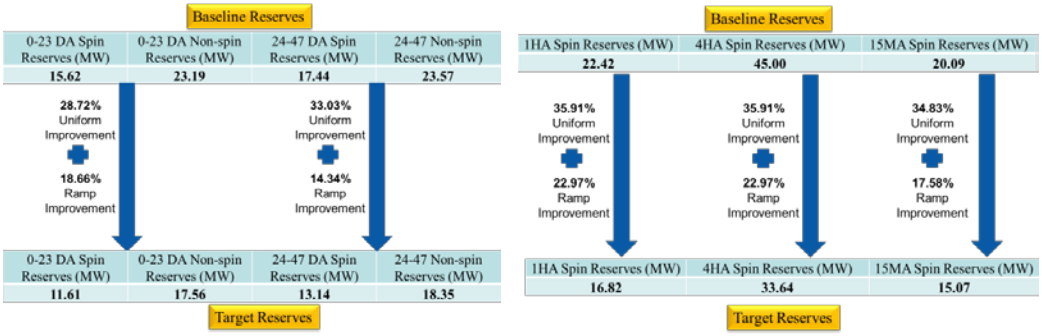
Figure 3: Distribution of baseline and target solar power forecast errors at DA, 4HA, 1HA, and 15MA forecast horizons (ISO-NE)

Metrics	DA (24-47) Baseline	DA (24-47) Target	DA (0-23) Baseline	DA (0-23) Target	4HA Baseline	4HA Target	1HA Baseline	1HA Target	15MA Baseline	15MA Target
Correlation coefficient	0.80	0.90	0.86	0.93	0.73	0.85	0.96	0.97	1.00	1.00
RMSE (W/m^2)	152.55	115.75	122.27	89.73	192.21	143.12	73.25	55.19	18.32	13.68
NRMSE by capacity	0.15	0.12	0.12	0.09	0.19	0.14	0.07	0.06	0.02	0.01
MaxAE (W/m^2)	617.54	528.96	513.11	369.03	715.12	553.44	357.36	265.00	129.05	91.99
MAE (W/m^2)	119.13	88.80	92.58	67.83	147.58	107.52	52.99	40.12	13.02	9.81
MAPE by capacity	0.12	0.09	0.09	0.07	0.15	0.11	0.05	0.04	0.01	0.01
MBE (W/m^2)	24.05	19.62	15.98	14.80	52.87	39.39	18.57	14.12	4.60	3.51
KSIPer (%)	170.83	148.47	147.98	121.29	144.25	131.54	60.92	47.50	20.74	17.21
OVERPer (%)	89.12	67.55	68.99	47.07	75.54	62.44	7.85	4.48	0.00	0.00
Std. dev. (W/m^2)	150.69	114.11	121.25	88.52	184.85	137.63	70.87	53.36	17.73	13.23
4RMQE (W/m^2)	212.81	165.39	175.62	128.06	261.63	198.49	107.36	80.52	28.17	20.91
N4RMQE by capacity	0.21	0.17	0.18	0.13	0.26	0.20	0.11	0.08	0.03	0.02
95th percentile (W/m^2)	315.68	234.01	254.17	184.18	380.34	286.38	154.16	116.37	38.41	28.55
Renyi entropy	5.29	5.11	5.18	5.16	5.22	5.10	4.74	4.80	4.42	4.48
NRMSE by clear sky irradiance	0.28	0.22	0.22	0.16	0.30	0.23	0.12	0.09	0.03	0.02
MAPE by clear sky irradiance	0.22	0.17	0.17	0.12	0.23	0.17	0.09	0.07	0.02	0.02

Table 8: Baseline and target metrics values for ISO-NE at different forecast horizons.

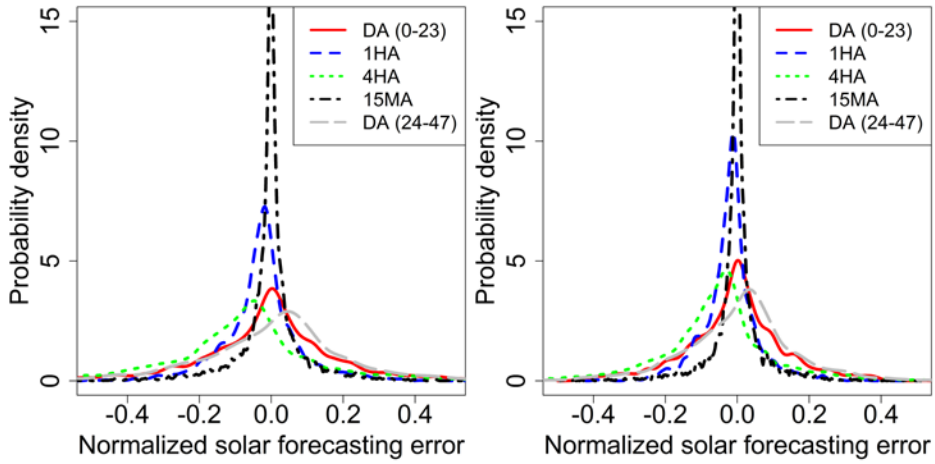
3.1.3.2 GMP Baseline and Target Metrics Values

GMP has relatively high solar penetration. At the time of the study, there was approximately 47 MW PV installed behind the meter, which represents about 5% of peak load in the GMP region. Table 9 summarizes the baseline and target values at different forecast horizons. Figure 4 shows the baseline and target reserves. For all forecast horizons, there are more uniform improvements than the ramp forecasting improvements required. Figures 5(a) and 5(b) illustrate the distributions of solar power forecast errors for baseline and target forecasting, respectively. The 4HA forecast tends to under forecast the power generation compared to other forecast horizons, which might be due to morning clouds in the region and the shading by mountains.



(a) Day-ahead forecasts at GMP (b) 15MA, 1HA, and 4HA forecasts at GMP

Figure 4: Target reserves values based on uniform and ramp forecasting improvement (GMP).



(a) Baseline (b) Target

Figure 5: Distribution of baseline and target solar power forecast errors at DA, 4HA, 1HA, and 15MA forecast horizons (GMP).

3.1.4 Reliability Metrics

In addition to the metrics discussed above, the team worked on metrics to quantify the practical reliability benefits of forecasts enhancements. Towards that end, a methodology was developed which utilized a multi-timescale power system operation model (Flexible Energy Scheduling Tool for Integration of Variable Generation (FESTIV)[33-35]) to calculate the area control area (ACE), the absolute area control error in energy (AACEE), the standard deviation of the area control error (σ_{ACE}) and the North American Electric Reliability Corporation Control Performance Standard 2 (CPS2) score based on the unit commitment, economic dispatch, and automatic generation control processes. In addition, a new integrated reliability metric, namely the Expected Synthetic Reliability (ESR) was developed, which quantifies the reliability performance from the improved solar power forecasts as

$ESR = {}^{1}/_{4}\,[CPS2_{score} - AACEE \quad - \sigma_{ACE} - N_V\ ]\pi r^2$, where N_V is the number of violation periods.

For this work a representative IEEE 118-bus system was adopted to simulate different scenarios with different levels of improvements, locations, forecast horizons, and solar penetration levels. The results are presented further below.

Metrics	DA (24-47) Baseline	DA (24-47) Target	DA (0-23) Baseline	DA (0-23) Target	4HA Baseline	4HA Target	1HA Baseline	1HA Target	15MA Baseline	15MA Target
Correlation coefficient	0.67	0.82	0.72	0.85	0.66	0.80	0.91	0.95	0.94	0.97
RMSE (MW)	9.44	7.19	8.63	6.47	10.87	8.02	5.21	3.83	4.29	3.23
NRMSE by capacity	0.20	0.15	0.18	0.14	0.23	0.17	0.11	0.08	0.09	0.07
MaxAE (MW)	38.10	30.06	30.05	24.45	45.43	35.00	29.10	22.42	31.16	23.81
MAE (MW)	7.03	5.35	6.21	4.69	7.89	5.74	3.64	2.64	2.42	1.73
MAPE by capacity	0.15	0.11	0.13	0.10	0.17	0.12	0.08	0.06	0.05	0.04
MBE (MW)	0.07	0.21	-0.36	-0.21	-3.76	-2.68	-1.25	-0.90	-0.07	-0.04
KSIPer (%)	138.02	147.06	108.81	119.86	213.1	148.51	79.73	57.25	10.06	12.87
OVERPer (%)	63.21	67.18	32.46	42.56	126.4	62.94	13.99	1.19	0.00	0.00
Standard dev. (MW)	9.45	7.19	8.63	6.47	10.20	7.56	5.06	3.73	4.29	3.23
4RMQE (MW)	13.42	10.25	12.28	9.17	15.91	11.98	8.00	5.99	7.77	6.09
N4RMQE by capacity	0.28	0.22	0.26	0.19	0.34	0.25	0.17	0.13	0.16	0.13
95th percentile (MW)	20.38	15.31	19.51	14.51	23.70	17.32	11.38	8.55	10.04	7.53
Renyi entropy	5.33	5.24	5.34	5.31	4.95	4.86	4.56	4.45	3.40	3.17
NRMSE by clear sky power	0.34	0.26	0.35	0.27	0.41	0.30	0.21	0.15	0.18	0.13
MAPE by clear sky power	0.25	0.19	0.26	0.19	0.29	0.21	0.14	0.10	0.10	0.07

Table 9: Baseline and target metrics values for GMP at different forecast horizons.

3.2 The *Watt-sun* Forecasting System

The main research theme in this task was to explore how accurate and scalable (thus low cost) forecasting may be enabled by blending multiple forecasting models using a novel machine learning approach. The foundation of renewable energy forecasting is physical modeling including NWP models[24,36] as well as models based on the advection of total sky imager[37,38] and satellite images[39,40]. Moreover, the accuracy of these models can be boosted by statistical post-processing. Established methods include model output statistics (MOS)[41,42], multi-model averaging[43-45], and the dynamic integrated forecast (DICast)[46,47] approach. DICast effectively combines MOS and model averaging - several MOS forecasts are averaged using weights optimized using typically a few days of history. More recently, aided by progresses in computation and the advent of *Big Data*[48,49] which enables convenient retrieval and processing of large volumes of historical data, more sophisticated machine learning techniques[50,51] begin to be employed[52-55]. In contrast to MOS or DiCast, which largely are based on linear regression, state-of-the-art machine learning promises the correction of forecasting errors which are nonlinear to the input variables or which are dependent on the interactions between the variables. Typically, machine learning is used to train a regression between historical measurements (e.g., solar power) as the response variable and historical forecasts (e.g., solar power, solar angles, temperature, etc.) as the predictor variables. The trained regression is then applied for future forecasts.

More specifically, we explored an approach of machine learning based, situation-dependent, multi-model blending for renewable energy forecasting. The salient feature is that a set of appropriately chosen parameters are used to create different weather situation categories in which the input models exhibit different error characteristics. Historical data are binned into different weather situations, and machine learning models for statistically correcting the forecasts are trained separately for each situation. This practice avoids some common pitfalls of the machine learning as it will become clear in the discussion below.

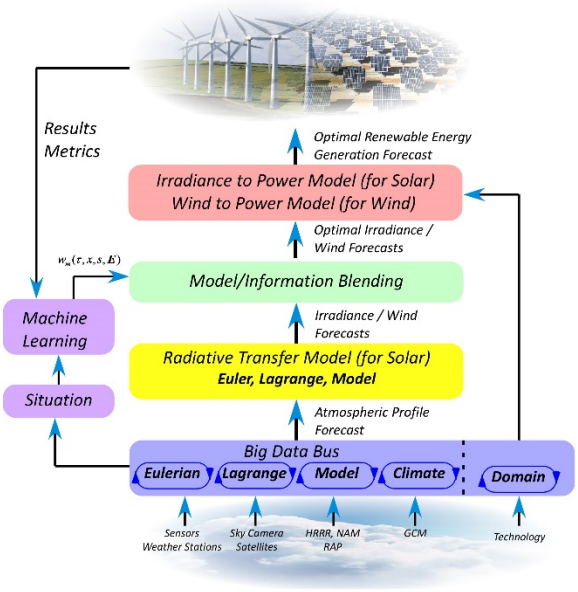


Figure 6: Architectural view of situation-dependent, machine learning based multi-model blending.

3.2.1 Overview of the *Watt-sun* Forecasting Method

In its simplest form, the situation-dependent multi-model blending method can be represented by the following equation yielding an optimal forecast (C_{blend}) for a given parameter (wind, irradiance, etc.) using a linear combination of models,

$$C_{blend}(\tau, x, s(E)) = \sum_m w_m(\tau, x, s(E)) C_m(\tau, x)$$

where τ is the forecast horizon, *x* is the spatial extent of the forecast, and *s* is the weather situation defined by a set of parameters ***E***. C_m is a forecast associated with an input physical model (e.g., an NWP model) and w_m is its respective machine-learnt weighting coefficient, which is a function of forecast horizon, location, and weather situation *s*. The index *m* corresponds to different physical models and forecast systems.

Figure 6 provides a simplified architectural overview of the system applied to renewable energy forecasting assuming historically measurements are available for the training targets. A “big data” bus provides forecasts of atmospheric conditions (such as temperature, wind speed, cloud properties, etc.) from various input forecasting models. A radiative transfer model module converts forecasted atmospheric conditions first into irradiance and then an irradiance to power model determines the generated solar power forecasts. If the method is applied to wind power forecasting a wind to power model would have to be deployed. The different power forecasts are blended by the information blending module. A categorization module classifies the weather situation and a machine learning module provides the blending for each weather situation as we will discuss in much more detail below. Initially the system is trained on historical data, but as new measurements become available it continually retrains. Next, a typical implementation of situation categorization and machine learning is represented, as summarized in Figure 7, including the rationale behind it, and a display of exemplary results.

For the analysis which we are discussing below, global horizontal irradiance (GHI), diffusive normal irradiance (DNI), surface temperature at 2 m height (T2m), and wind speed at 10 m above

ground (W10m) measurements were taken from seven stations of the surface radiation (SurfRad) network[56]. GHI and DNI forecasts are calculated from the vertical atmospheric and cloud profiles (temperature, pressure, humidity, cloudy liquid water, and ice content) and surface albedo forecasted by the NWPs using a plane-parallel multi-layer radiative transfer model [29]. T2m and W10m forecasts are taken from the NWPs directly. Daily 18h Coordinated Universal Time (UTC) runs of the North American Mesoscale (NAM)[57] model (resolution is 5 km) and the Global Forecast System (GFS)[58] (resolution is 0.5 deg), and 15h UTC run of the Short Range Ensemble Forecast (SREF)[58] (40 km resolution with the advanced research Weather Research Forecast (WRF) core, central member)[59] are used to extract the day-ahead forecasts of GHI, DNI, T2m, and W10m. Forecasts of 12 to 36 hours ahead for NAM (18z run) and GFS and 15 to 39 hours for SREF (15z run) ahead are extracted and validated against the measurements. The validation time is one year from 2015-03-01 to 2016-02-28. The training of the RF model and situation-dependent blending is performed on data from 2013-03-01 to 2015-02-28.

3.2.2 Categorization of Weather Situations

The categorization of weather situations starts with analyzing how the systematic errors of the individual forecast models depend on the atmospheric state parameters including forecasted ones. To illustrate the process, we show here the results for GHI forecasts from the NAM model. The GHI forecasting errors are quantified using measurements from the SurfRad station in Bondville (BND), Champagne, IL.[56] The forecasting error dependences on atmospheric parameters, such as Direct Normal Irradiance (DNI), cloud liquid water and cloud ice contents, cloud base and top heights, surface temperature at 2 m (T2M), surface pressure, etc., are derived from the daily run of the NAM model at 18 UTC hour using Functional Analysis of Variance (FANOVA).

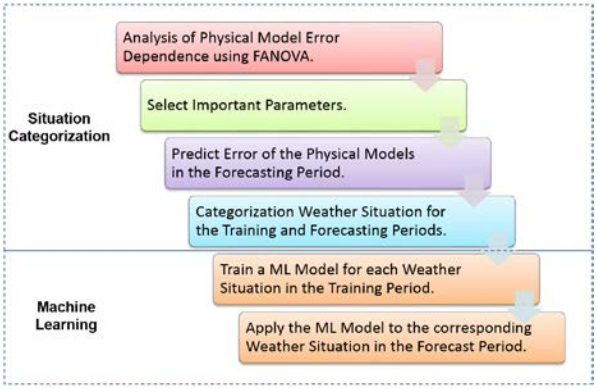


Figure 7: Flow chart showing the steps of situation categorization and machine learning.

3.2.2.1 Functional Analysis of Variance

Feeding the input data from the forecasts into a quantile regression forest[60] learning model, one fits of the relationship between the GHI forecast error and the forecasted parameters,

$E_{GHI}=F(x_1, x_2, \ldots, x_n)$,

where $E_{GHI} = GHI_{forecasted} - GHI_{measured}$, is the GHI forecast error, x_1 is $GHI_{forecasted}$, and x_2, …, x_n are the additional forecasted parameters.

Limited by the size of the available training dataset and high dimensionality of the input parameters, such statistical fitting of forecast error is noisy. To counter the limitation, the errors of the forecasts are then broken to its 0^{th}, 1^{st}, 2^{nd} … order dependence on individual input parameters

using FANOVA[61]. , FANOVA decomposes the overall error into mean bias, the dependence on individual parameters, the interaction between two-parameter pairs, etc. The zeroth order term f_0 is the mean bias error of a forecast. The first order term f_i provides the error dependence on x_i only, while the effects of all other parameters are averaged out (with zeroth order term removed). The second order term $f_{i,j}$ provides the error dependence on x_i and x_j (with zeroth and first order terms removed): $F = f_0 + \sum_i f_i(x_i) + \sum_{i \neq j} f_{i,j}(x_i, x_j) + \ldots$ with

$$f_0 = \int F(x_1,...,.x_n)dx_1...dx_n$$

$$f_i = \int F(x_1,...,.x_n)dx_1...dx_{i-1}dx_{i+1}dx_n - f_0$$

$$f_{i,j} = \int F(x_1,...,.x_n)dx_1...dx_{i-1}dx_{i+1}....dx_{j-1}dx_{j+1}..dx_n - f_i(x_i) - f_j(x_j) - f_0$$

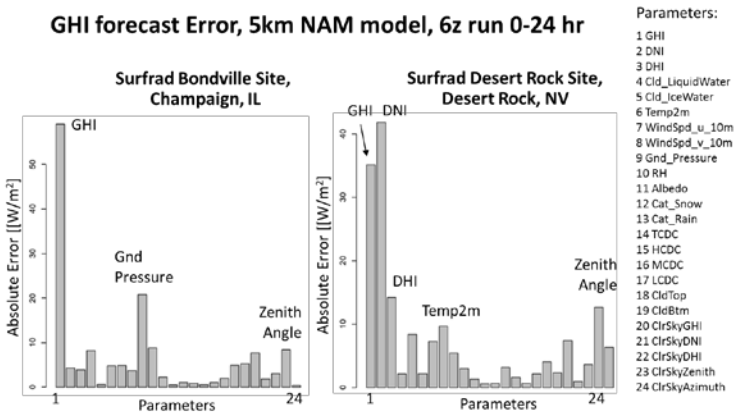


Figure 8: Significance of Input Parameters on FANOVA 1st Order Error

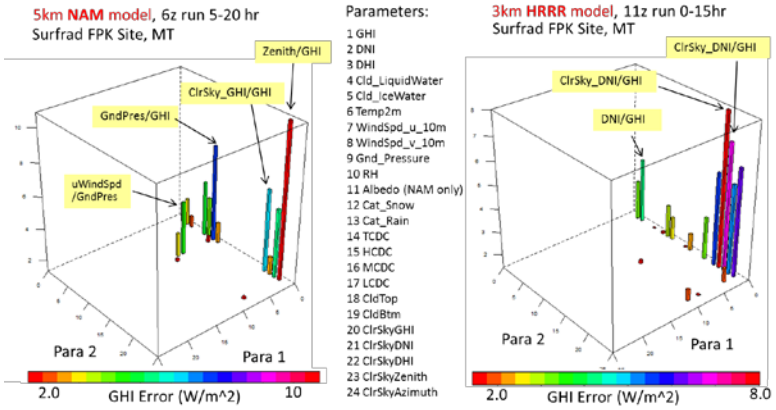


Figure 9: Significance of Input Parameters on FANOVA 2nd Order for the NAM and the High Resolution Rapid Refresh (HRRR) model.

3.2.2.2 Parameter Selection

For any FANOVA term (0th,1st, 2nd order, or beyond) as shown in Figure 10, we computed the variance of the error. Generally, a large variance means the FANOVA term has a large

dependence on the parameter(s). Thus, it is important to include the parameter(s) for situation categorization. As shown in Figure 8 and 9 for 24 parameters, the importance of the parameters towards situation categorization is *site* and *model* dependent thus important parameters and weather situation categorization need to be determined using the training dataset on a case-by-case basis. The importance of a given parameter is quantified by summing up the variance of all the 1st and 2nd order FANOVA terms relating to it. All parameters derived from the NWP models are ranked and the top parameters of importance beyond a threshold (1 W/m^2 for GHI forecasting) are selected. The max number of parameters is limited by the size of the training data, as a practical rule of thumb, about 1/100 of the number of training labels available.

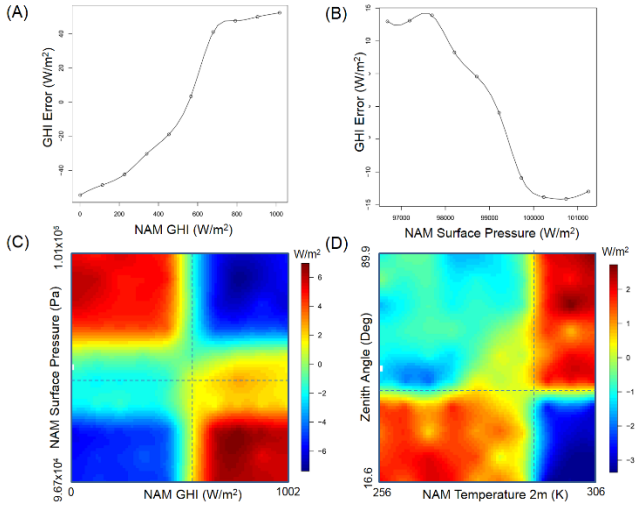


Figure 10: Salient examples of NAM GHI forecast error obtained from FANOVA. (A) and (B) show 1st order dependence on NAM forecasts of GHI and surface pressure, respectively. (C) and (D), respectively, show salient 2nd order GHI forecast error dependences on GHI and surface pressure (C) and on 2 m temperature and zenith angle (D).

3.2.2.3 Examples for Situation Categories

Figure 10 shows illustrative examples of salient the FANOVA estimated 1st and 2nd order NAM GHI forecasting error dependences. Additional examples and discussions are included further below. Figures 10A and 10B show the 1st order error dependence on GHI and surface pressure, respectively. A negative 1st order error (under-prediction) occurs for small GHI or large surface pressure, while a positive 1st order error (over-prediction) occurs for large GHI or small surface pressure. Two examples of 2nd order GHI error dependence on input parameters are shown in Figures 10C and 10D. We observe that the forecasting error vs. GHI and surface pressure forecasts (Figure 10C) can be roughly divided into four regions or situations. For small (large) GHI and small (large) pressure, the 2nd order forecasting error is negative, otherwise the 2nd order error is positive. Similarly, a strong interaction between forecasts of 2 m temperature and zenith angle is observed in Figure 10D. More examples from the FANOVA analysis are shown below.

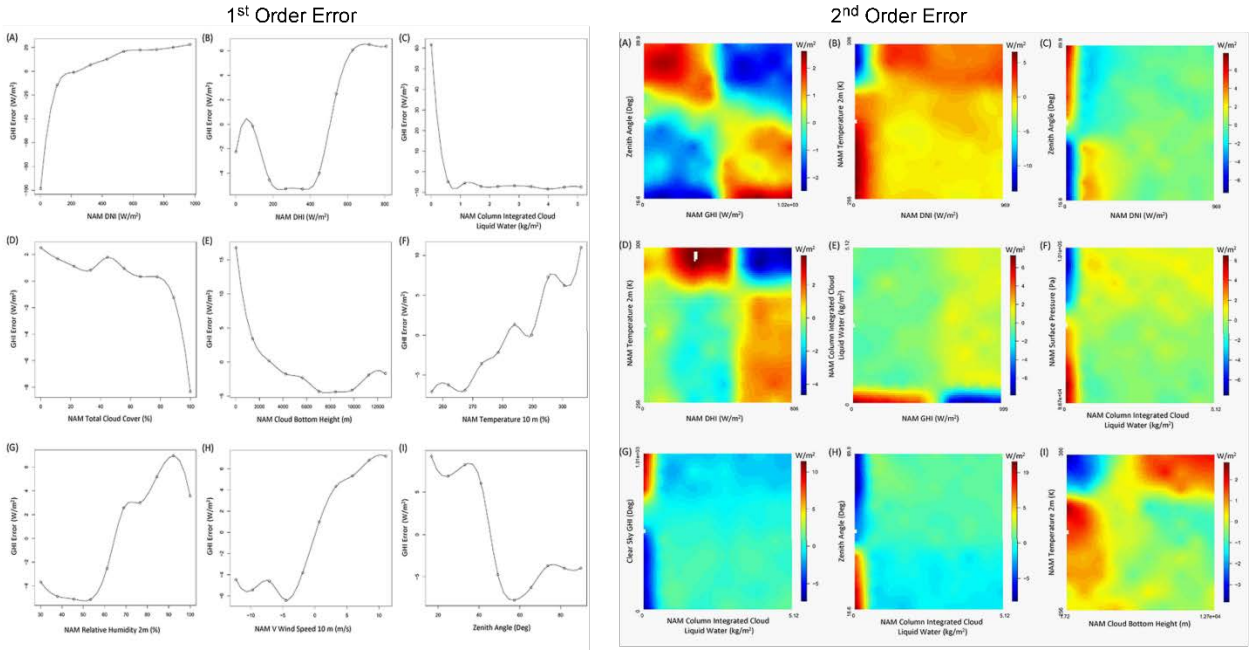


Figure 11: Additional examples of 1st order and 2nd order NAM GHI forecast error dependence at the Surfrad BND station in Champagne, IL.

Such error dependences are high-order systematic errors of the forecasting model (mean bias error is the 0th order systematic error). While it is of separate interest to investigate their underlying causes for improving the different forecasting models, which was done throughout the project with Stan Benjamin from the Earth System Research Laboratory (ESRL) of the National Oceanic and Atmospheric Agency (NOAA), here for machine learning aimed at statically minimizing error, their implication is two-fold. First, the error dependence provides information on selection of important parameters carrying information to improve model accuracy. Second, the pattern of error dependences on the important parameters suggests that one may divide the entire space into subspaces (i.e., situation categories) based on the expected model error, as illustratively marked by dashed lines in Figures 10C and 10D. Such situation categorization ensures the forecasting error of an input physical model is similar in the same category. This enables the more effective forecasting error reduction using machine learning because forecasts can be trained using data of similar nature.

When multiple input models are involved, the dimensionality of the space (formed by the important parameters from the multiple models) increases. For such situation categorization, since we are ultimately concerned about combining the different weather models so that their errors can be reduced, an intuitive way is to categorize according to the expected errors of the individual models, which in turn is linked to the important parameters via the FANOVA derived error dependences. For simplicity of visualization, Figure 12 gives an example of the GHI forecast error for the BND SurfRad site using NAM and GFS models. Figure 13 shows a three-model situation categorization. An unsupervised classification learning algorithm, Gaussian mixture models[62], is used to classify situation categories. The color of each point visualizes the resulting categories. We constrain the maximum number of situation categories to be up to ten, while the optimal number of categories is determined by the Bayesian Information Criteria. It is worth noting that such situation categorization, even though entirely data driven (using model error), it nevertheless is correlated with empirically defined meteorological weather situations (such as clear-sky, partially cloudy, overcast, etc.) as discussed below. For a given forecasting data point, we first

compute the expectation of the error of the individual models using FANOVA by summing up the error dependences on all important parameters, and then use the trained Gaussian mixture model[63] to categorize the data points.

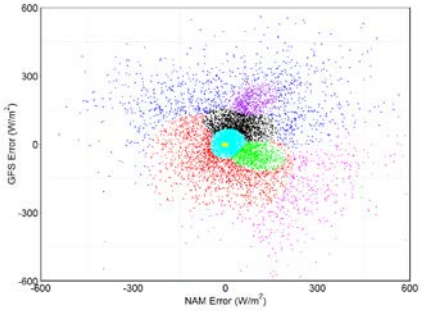


Figure 12: Color of the dots shows the situation categories created by the error of NAM and GFS forecasts for the BND SurfRad site.

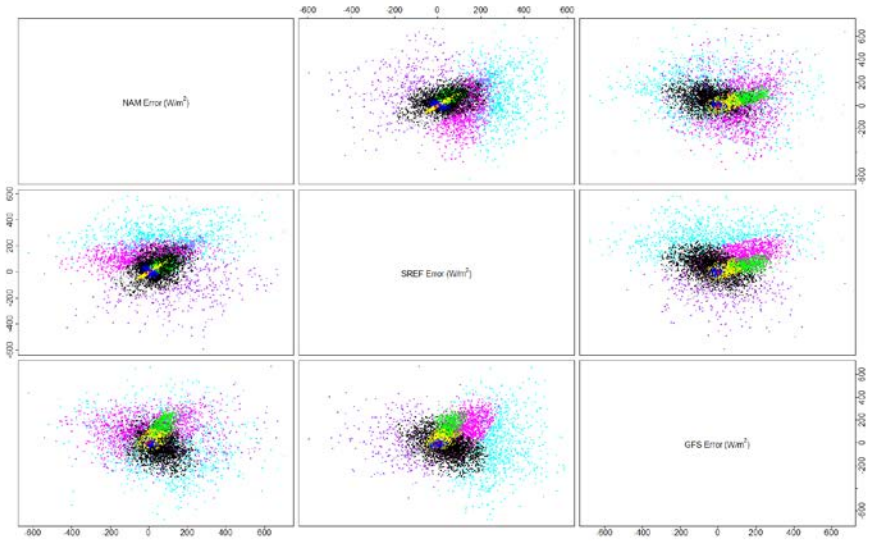


Figure 13**:** Color of the dots shows the situation categories created by the error of NAM, GFS, and SREF forecasts for the BND Surfrad site.

3.2.3 Machine Learning Models

For each situation category within given periods of training data and forecast data, a supervised machine learning model is independently trained on the training time period (establishing a regression between the predictor variables and the response variables by minimizing a certain cost function = metrics) and applied to the forecast time period. Generally, the response variables are the measurements of the quantity of interest such as solar irradiance. In the simplest form, predictor variables are the forecasts of the quantity of interest by different NWPs or other physical models. Including predictions of selected important parameters such as temperature, pressure, etc. often leads to better accuracy. Testing an array of supervised learning algorithms, we found that significant forecasting error reduction with respect to the best model can be achieved.

The situation categorization ensures that the input models have relatively similar error values in the same situation category, so that a learning model can more effectively reduce the error of the forecast. Take for example the Random Forest (RF) model[64], which is commonly used for forecasting when there is sufficient data. RF is a type of bagging method[65]. It averages an ensemble of over-fitted tree models, each model fitted on a subset of training data and using a subset of predictor variables, which makes the RF model robust - even if the training data contain predictor variables which are irrelevant or highly correlated with each other, RF performance does not degrade significantly. Such benefit of a RF model, however, at the same time can cause difficulty for accurate prediction of infrequent cases. Since there is relatively small number of such training data, most trees in the ensemble will not see them, the averaged prediction of the ensemble is thus biased towards the "mean" (i.e., the fitting of the common cases). The situation categorization helps RF by grouping the common cases and the infrequent cases into the training data in different situation categories, thus mitigates the "bias towards mean" problem. In addition, the categorization also prevents a few erroneous outlier training data points from significantly impacting the forecasting performance as these data points, due to their different error characteristic, tend to be classified as a separate situation category.

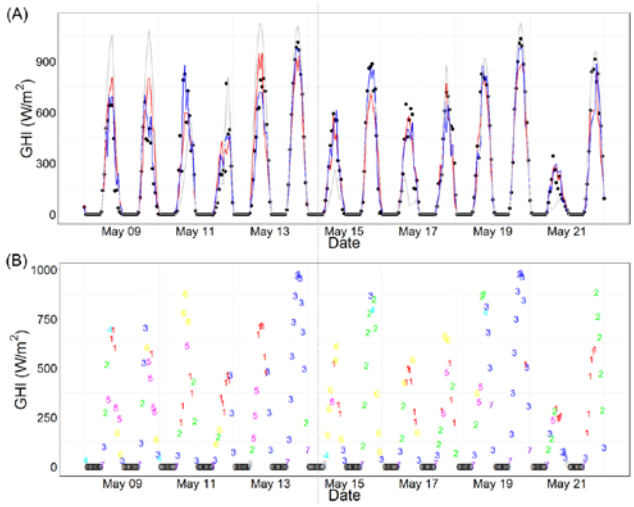


Figure 14: (A) Exemplary GHI forecasted day-ahead by GFS (gray line), random forest learning without situation categorization (red line), and situation dependent blending (blue line) compared to measurements (black squares). The data is for SurfRad BND station in Champaign, Illinois 05/08/15 to 05/21/15. Forecasts were issued at 18h UTC for the 12 to 36 hours ahead. The three individual models used to create the blending are NAM, GFS, and SREF. (B) repeats the situation dependent blending forecasts (blue line in A) with the situation categories represented by the color of each data points.

In addition to RF, Linear Model (LM) and Support Vector Machine (SVM)[66,67] are also used for the forecasting. LM provides explicit situation-dependent blending weight coefficients, thus helps evaluating the performance of the different models in different situations. LM is also favored when the training data set becomes excessively large given its lower computing cost[68]. SVM (with radius basis function kernel) using selected predictor variables provides a comparable accuracy to RF, but is often more accurate when only a short period of training data is available. Given no single machine learning algorithm covers accuracy, robustness, and flexibility, a multi-expert learning system combining them is one way to achieve the best overall performance. A multi-expert learning system which dynamically selects an individual learning algorithm from a set of competing ones according to recent performance is employed. By using a set of learning algorithms of different complexities (thus of training data requirement), the multi-expert learning mitigates common troubles associated with training data (for example changes in the NWP model

or missing training data). It thus improves the robustness of the forecasting system in absence of manual intervention, which in turn improves the scalability.

3.2.3.1 Multi-Expert Learning

Robustness of the forecasts is of importance for providing operational forecasts at low cost. Methods need to be scalable to many sites and different forecasting targets with the least involvement of a human expert. Some of the hurdles include the upgrade or temporary unavailability of an individual models, missing or undetected erroneous measurement data, and events that cannot be reliably predicted such as unplanned maintenance at a solar farm. A multi-expert learning approach mitigates their impact. The tradeoff is the cost of more computation which is becoming increasingly affordable. For multi-expert learning, a dozen machine learning models are set-up. The individual machine learning algorithms are run in parallel, the algorithm provides the best accuracy for the last two days is selected for future forecasts. To automatically choose the best performing model setting, we varied the following configurations: (1) the selection of the input models, (2) the selection of the training data size, (3) the maximum number of situation categories, and (4) the machine learning algorithm. A typical mix of machine learning models for 12 to 36 hours ahead forecasting as shown in Table 10. The different selection of input models deals issues relating to model change or unavailability of a model data. The different training data size deals with issues relating to model changes, data availability, or potential changes in site specification (such as degradation of efficiency). In such cases, it is advantageous to exclude specific models or old training data. We also note that the accuracy for some sites benefits from including only training data from the same season. While overall the situation categorization improves forecasting accuracy, accuracy may reduce if forecasting data points are misclassified. This problem is mitigated by varying the maximum number of situation categories (including the limiting case of no categorization). The combination of RF, SVM, and LM enables both high accuracy and robustness as discussed. Additionally, two “no-learning” experts taking directly GFS model output and persistence are also included to handle rare case of no data availability or certain unpredictable events (such as PV plant electrical failures).

	Input models	Training data size	Situation categories	Learning Algorithm
1	NAM/SREF/GFS	All data, hourly	Yes	RF
2	NAM/SREF/GFS	All data, hourly	Yes	SVM
3	NAM/SREF/GFS	All data, hourly	Yes	LM
4	NAM/SREF/GFS	All data, hourly	No	RF
5	NAM/SREF/GFS	Same season, hourly	Yes	RF
6	NAM/SREF/GFS	3 months, hourly	Yes	SVM
7	NAM/SREF/GFS	1 months, hourly	No	SVM
8	GFS	1 month, hourly	No	RF
9	NAM	1 month, hourly	No	RF
10	NAM	1 week, hourly	No	LM
11	GFS	None	No	None
12	Persistence	None	No	None

Table 10: A typical mix of machine learning models for multi-expert learning.

3.2.4 Forecasting Error Reduction

Figure 14A shows an exemplary period of GHI measurements (black squares) at BND station versus forecasts by GFS (gray line), conventional machine learning using RF model (red line),

and situation dependent blending (blue line). Here the RF model is used to establish a baseline, as after testing different learning models we found that it reflects the best accuracy of a conventional machine learning approach without situation categorization have to offer. While the RF model (red line) reduced the forecasting errors, compared to individual models (such as the GFS forecasts, gray line), its short-coming is a tendency towards the mean irradiance value. For the two clear sky days, 05/13/15 and 05/19/15, the RF forecasts are below the measurements, while for the cloudy days 05/08/15 and 05/09/15, the RF forecasts are higher than the measurements. This indeed reflects the bias towards "mean" of the RF learning algorithm as elucidated previously. Using situation categorization, the forecasts (blue line) are improved for both clear sky and cloudy days. The situation categorization based forecasts are shown in Figure 10B. It is observed that the two clear sky days belong mostly to one category (blue) while the cloudy days belong to other categories. Note that the situation categories were created using FANOVA predicted forecasting errors of the individual models without explicitly dealing with clear sky or cloudy. The clear sky days are nevertheless put into the same category (blue), presumably because clear sky days have distinct error characteristics. As the situation categorization enables different learning models to be fitted for clear sky *vs.* cloudy days, a reduction of overall error follows.

Figure 15 summarizes the mean absolute error of the four parameters predicted by the NAM, GFS, SREF models as well as by RF model and situation-dependent blending. The MAE of the GHI forecasts (Fig. 15A) by the uncorrected NAM, GFS, and SREF models are respectively 94, 115, and 103 W/m^2 (red bars). The MAE of machine learning (RF model) without situation categorization is 80 W/m^2 (green bar). In contrast, the situation-dependent blending reduces the MAE to 72.5 W/m^2 (blue bar), a ~30% improvement with respect to the best individual model NAM, and a ~10% reduction with respect to RF learning. Similar degrees of improvement are also seen for DNI, T2m, and W10m as shown in Figure 15(B,C,D). The detailed GHI forecasting accuracy comparisons of individual models, RF learning, and situation-dependent blending using different metrics, are provided in the Table 11 for all Surfrad stations.

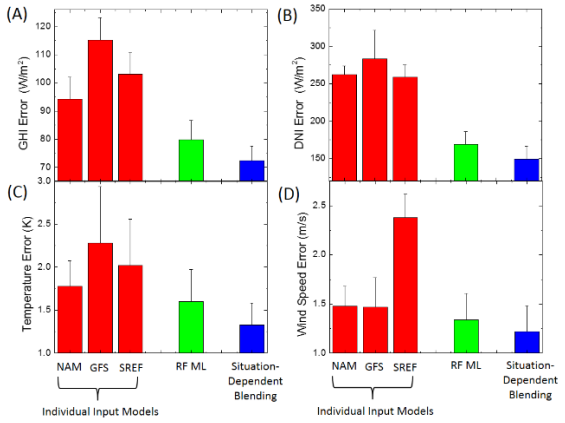


Figure 15 summarizes day-ahead forecast error of (A) global horizontal irradiance (GHI), (B) direct normal irradiance (DNI), (C) temperature at 2 m above ground (T2m), and (D) wind speed at 10 m above ground (W10m) using different methods. Red: uncorrected NAM, GFS, and SREF. Green: conventional machine learning using random forecast model without situation categorization. Blue: Situation-dependent model blending. The data shown are the average forecast error (day-time only) of the seven SurfRad stations from 02/28/15 to 02/28/16. The error bars show the standard deviation of the errors at the seven stations. Forecasts were issued at 18h UTC for the 12 to 36 hours ahead. The three individual models used to create the blending are NAM, GFS, and SREF.

Station Bondville (BND), Champaign, IL

BND Metrics	NAM	GFS	SREF	RF ML	Situation-Dependent Blending
Capacity (W/m^2)	1000	1000	1000	1000	1000
Correlation Coefficient	0.863	0.836	0.842	0.905	0.914
RMSE (W/m^2)	144	176	166	115	110
NRMSE	0.144	0.176	0.166	0.115	0.11
MaxAE (W/m^2)	947	846	812	667	631
MAE (W/m^2)	90.9	118	106	77.6	72.6
MAPE	0.0909	0.118	0.106	0.0776	0.0726
MBE (W/m^2)	23.4	37	55.6	-7.01	-4.89
KSIPer	4.784	6.57	9.435	4.685	4.292
Std Deviation (W/m^2)	142	173	157	115	110
Skewness	0.38	0.298	0.696	0.459	0.175
Kurtosis	4.32	2.78	3.46	3.31	3.45
RMQE_4 (W/m^2)	237	274	265	181	175
NRMQE_4	0.237	0.274	0.265	0.181	0.175
Percentile95(W/m^2)	276	359	368	186	180

Station Table Mountain (TBL), Longmont, CO

TBL Metrics	NAM	GFS	SREF	RF ML	Situation-Dependent Blending
Capacity (W/m^2)	1000	1000	1000	1000	1000
Correlation Coefficient	0.846	0.836	0.81	0.887	0.906
RMSE (W/m^2)	173	176	176	132	120
NRMSE	0.173	0.176	0.176	0.132	0.12
MaxAE (W/m^2)	875	846	848	607	562
MAE (W/m^2)	107	118	117	91.5	79.7
MAPE	0.107	0.118	0.117	0.0915	0.0797
MBE (W/m^2)	63.3	37	22.3	-15.3	-6.51
KSIPer	10.042	6.57	5.656	6.37	5.173
Std Deviation (W/m^2)	161	173	175	131	120
Skewness	0.779	0.298	0.251	0.236	-0.00909
Kurtosis	3.54	2.78	2.71	2.32	2.89
RMQE_4 (W/m^2)	276	274	273	198	187
NRMQE_4	0.276	0.274	0.273	0.198	0.187
Percentile95(W/m^2)	388	359	355	216	196

Station Fort Peck (FPK), Poplar, MT

FPK Metrics	NAM	GFS	SREF	RF ML	Situation-Dependent Blending
Capacity (W/m^2)	1000	1000	1000	1000	1000
Correlation Coefficient	0.887	0.836	0.878	0.913	0.922
RMSE (W/m^2)	136	176	134	103	97.7
NRMSE	0.136	0.176	0.134	0.103	0.0977
MaxAE (W/m^2)	837	846	781	649	648
MAE (W/m^2)	88.1	118	84.8	68.6	63.8
MAPE	0.0881	0.118	0.0848	0.0686	0.0638
MBE (W/m^2)	54.2	37	42.7	-5.13	-2.34
KSIPer	9.466	6.57	7.881	4.171	3.299
Std Deviation (W/m^2)	124	173	127	103	97.7
Skewness	0.404	0.298	0.599	0.798	0.637
Kurtosis	5.92	2.78	5.63	4.36	4.24

RMQE_4 (W/m^2)	225	274	226	169	160
NRMQE_4	0.225	0.274	0.226	0.169	0.16
Percentile95(W/m^2)	276	359	272	178	159

Station Goodwin Creek (GCM), Goodwin Creek, Mississippi

GCM Metrics	NAM	GFS	SREF	RF ML	Situation-Dependent Blending
Capacity (W/m^2)	1000	1000	1000	1000	1000
Correlation Coefficient	0.837	0.836	0.835	0.887	0.907
RMSE (W/m^2)	167	176	191	131	120
NRMSE	0.167	0.176	0.191	0.131	0.12
MaxAE (W/m^2)	866	846	899	677	743
MAE (W/m^2)	105	118	118	87.4	77.2
MAPE	0.105	0.118	0.118	0.0874	0.0772
MBE (W/m^2)	37.1	37	85.2	1.27	5.14
KSIPer	7.23	6.57	13.656	6.077	5.625
Std Deviation (W/m^2)	163	173	171	131	120
Skewness	0.554	0.298	1.44	0.9	0.628
Kurtosis	4.05	2.78	2.9	3.81	4.2
RMQE_4 (W/m^2)	273	274	306	213	197
NRMQE_4	0.273	0.274	0.306	0.213	0.197
Percentile95(W/m^2)	331	359	458	240	219

Station Sioux Falls (SXF), Garretson, SD

SXF Metrics	NAM	GFS	SREF	RF ML	Situation-Dependent Blending
Capacity (W/m^2)	1000	1000	1000	1000	1000
Correlation Coefficient	0.856	0.836	0.833	0.895	0.908
RMSE (W/m^2)	150	176	163	116	109
NRMSE	0.15	0.176	0.163	0.116	0.109
MaxAE (W/m^2)	902	846	758	714	675
MAE (W/m^2)	90.3	118	103	77.2	70.9
MAPE	0.0903	0.118	0.103	0.0772	0.0709
MBE (W/m^2)	41	37	39.1	-6.75	-4.97
KSIPer	6.447	6.57	7.54	4.327	4.128
Std Deviation (W/m^2)	144	173	158	116	109
Skewness	0.455	0.298	0.483	0.604	0.57
Kurtosis	5.66	2.78	3.66	3.77	3.98
RMQE_4 (W/m^2)	254	274	262	186	176
NRMQE_4	0.254	0.274	0.262	0.186	0.176
Percentile95(W/m^2)	301	359	341	203	189

Station Desert Rock (DRA), Desert Rock, NV

DRA Metrics	NAM	GFS	SREF	RF ML	Situation-Dependent Blending
Capacity (W/m^2)	1000	1000	1000	1000	1000
Correlation Coefficient	0.895	0.836	0.885	0.916	0.922
RMSE (W/m^2)	145	176	141	119	114
NRMSE	0.145	0.176	0.141	0.119	0.114
MaxAE (W/m^2)	899	846	812	732	655
MAE (W/m^2)	84.5	118	87	77.6	69.6
MAPE	0.0845	0.118	0.087	0.0776	0.0696
MBE (W/m^2)	52	37	21.6	-6.64	-5.15
KSIPer	8.371	6.57	5.738	5.423	3.389

Std Deviation (W/m^2)	136	173	139	119	114
Skewness	1.15	0.298	1.74	1.14	0.253
Kurtosis	6.9	2.78	4.99	5.11	5.24
RMQE_4 (W/m^2)	255	274	242	199	193
NRMQE_4	0.255	0.274	0.242	0.199	0.193
Percentile95(W/m^2)	305	359	315	215	178

Station Penn State University (PSU), Pennsylvania Furnace, PA.

PSU Metrics	NAM	GFS	SREF	RF ML	Situation-Dependent Blending
Capacity (W/m^2)	1000	1000	1000	1000	1000
Correlation Coefficient	0.839	0.836	0.838	0.891	0.901
RMSE (W/m^2)	153	176	170	121	116
NRMSE	0.153	0.176	0.17	0.121	0.116
MaxAE (W/m^2)	779	846	840	652	746
MAE (W/m^2)	96.7	118	109	81.1	75.7
MAPE	0.0967	0.118	0.109	0.0811	0.0757
MBE (W/m^2)	26.3	37	61.3	-4.85	-0.9
KSIPer	7.694	6.57	11.145	5.541	4.881
Std Deviation (W/m^2)	151	173	158	121	116
Skewness	0.128	0.298	0.881	0.305	0.287
Kurtosis	3.68	2.78	3.11	3.13	3.79
RMQE_4 (W/m^2)	245	274	269	190	187
NRMQE_4	0.245	0.274	0.269	0.19	0.187
Percentile95(W/m^2)	292	359	378	209	196

Table 11: Comparison of GHI day-ahead (12 - 36 hours) forecasting error for the seven NOAA Surfrad stations using the suite of metrics developed in this project. Forecasts are generated using uncorrected NAM, GFS, SREF, random forecast model, and situation-dependent model blending. The data are for time period 02/28/15 to 02/28/16.

The underlying reason for such performance improvement may be understood by revisiting the situation-dependent error of the forecasts. Statistical post-processing corrects the systematic error of one or more forecast models. The conventional multi-model averaging method largely reduces important parameters in the predictor variables, machine learning approaches are capable of correcting higher order errors dependences, thus achieving an overall improved forecast accuracy. For instance, recalling the NAM model GHI forecasting has a significant 2nd order error dependence on 2m Temperature and zenith angle ranging from -3.5 to 3.5 W/m^2 (Figure 10C). The corresponding 2nd order error plots of GHI forecasts from conventional machine learning (RF model) and situation-dependent blending are shown in Figure 16. Both RF learning and situation-dependent blending reduce the 2nd order error to a range of -2.5 to +2.5 W/m^2. Furthermore, even compared to RF learning (Figure 16A), the situation-dependent blending (Figure 16B) apparently has further reduced the 2nd order error dependence on 2m Temperature and zenith angle. (The variances of 2nd order error plots are 0.45 and 0.92, respectively.) In situation-dependent blending, the weather situations are categorized according to the forecasting errors of the input models that are linked to the important parameters via the FANOVA derived error dependences exemplified by Figure 10. Such observed reduction of high order error indicates that machine learning becomes more effective for error reduction in each situation individually, which is at the core of the better overall accuracy of situation-dependent blending compared to conventional approaches.

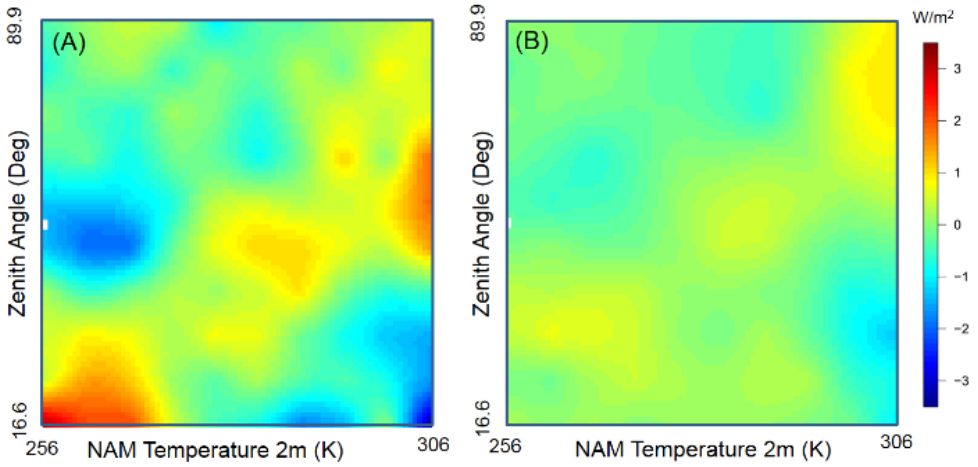


Figure 16: Second order GHI forecast error dependences on NAM temperature 2 m forecast and zenith angle. (A) is the result from forecast using random forest model without situation categorization and (B) is from situation-dependent model-blending.

3.2.5 Results from 5 Test Sites

Site 504 TEP FRV

Time horizon	24 to 47 hr	0 to 23 hr	4 hr	1 hr	15 min
BaseLine	3.56	4.03	4.09	2.14	1.11
Target	2.67	2.94	3.017	1.618	0.78
Status	2.41	1.6	2.151	1.695	0.976
Improvement	129.21%	222.94%	180.65%	85.22%	40.61%

Site 101 Smyrna

Time horizon	24 to 47 hr	0 to 23 hr	4 hr	1 hr	15 min
BaseLine	0.13	0.12	0.14	0.122	0.03
Target	0.1	0.09	0.11	0.087	0.02
Status	0.099	0.088	0.106	0.088	0.028
Improvement	103.33%	106.67%	113.33%	97.54%	18.00%

Site 901 GMP PostRoad

Time horizon	24 to 47 hr	0 to 23 hr	4 hr	1 hr	15 min
BaseLine	7.03	6.21	7.89	6.64	2.42
Target	5.35	4.69	5.74	4.816	1.73
Status	4.83	4.54	4.936	4.463	1.95
Improvement	130.95%	109.87%	137.40%	119.34%	68.12%

Site 2000 CAISO

Time horizon	24 to 47 hr	0 to 23 hr	4 hr	1 hr	15 min
BaseLine	98.56	98.91	111.97	93.98	22.24
Target	71.74	72.68	85.35	70.95	15.45
Status	90.3	82.7	83.500	81.300	19.3
Improvement	30.80%	61.80%	105.950%	55.059%	43.30%

Site 12201 ISONE SEMA

Time horizon	24 to 47 hr	0 to 23 hr	4 hr	1 hr	15 min
BaseLine	119.13	92.58	147.58	52.99	13.02
Target	88.8	67.83	107.52	40.12	9.81

Status	52.6	59	48.4	46.500	10.9
Improvement	219.35%	135.68%	247.58%	50.427%	66.04%
AVERAGE					
Time horizon	**24 to 47 hr**	**0 to 23 hr**	**4 hr**	**1 hr**	**15 min**
Ave. for 5 sites	**122.73%**	**127.39%**	**157.18%**	**81.52%**	**47.21%**

Table 12: Summary of the MAE (in MW) forecasting results for the 5 test cases. The validation time period is one year from 2015-03-01 to 2016-02-28. The training of the RF model and situation-dependent blending is performed on data from 2013-03-01 to 2015-02-28.

Detailed results from the *Watt-sun* system for the 5 test sites have been reported at UVIG 2016 DoE Solar Forecasting workshop in Denver, in previous reports and other publications. Here we focus on the summary of the results for the MAE and RMSE metrics as shown in Table 12 and Table 13, respectively. Improvements are defined as ((A-B)/(T-B) with A as the achieved (status), B the baseline, and T the target value.

Site 504 TEP FRV

Time horizon	**24 to 47 hr**	**0 to 23 hr**	**4 hr**	**1 hr**	**15 min**
BaseLine	5.3	5.82	5	3.12	2.04
Target	3.99	4.21	3.68	2.34	1.55
Status	3.71	3.1	3.46	2.52	1.75
Improvement	121.37%	168.94%	116.67%	76.92%	59.18%

Site 101 Smyrna

Time horizon	**24 to 47 hr**	**0 to 23 hr**	**4 hr**	**1 hr**	**15 min**
BaseLine	0.17	0.17	0.19	0.1	0.05
Target	0.13	0.12	0.14	0.07	0.04
Status	0.12	0.115	0.142	0.072	0.047
Improvement	125.00%	110.00%	96.00%	93.33%	30.00%

Site 901 GMP PostRoad

Time horizon	**24 to 47 hr**	**0 to 23 hr**	**4 hr**	**1 hr**	**15 min**
BaseLine	9.44	8.63	10.87	5.21	4.29
Target	7.19	6.47	8.02	3.83	3.23
Status	7.2	6.6	6.42	3.72	3.32
Improvement	99.56%	93.98%	156.14%	107.97%	91.51%

Site 2000 CAISO

Time horizon	**24 to 47 hr**	**0 to 23 hr**	**4 hr**	**1 hr**	**15 min**
BaseLine	168.39	150.54	184.62	119.91	29.01
Target	120.05	110.82	149.17	90.75	21.42
Status	145.3	120.23	145.2	105.2	24.56
Improvement	47.77%	76.31%	111.20%	50.45%	58.63%

Site 12201 ISONE SEMA

Time horizon	**24 to 47 hr**	**0 to 23 hr**	**4 hr**	**1 hr**	**15 min**
BaseLine	152.55	122.27	192.21	73.25	18.32
Target	115.75	89.73	143.12	55.19	13.68
Status	79	86	74.8	57.2	16.2
Improvement	199.86%	111.46%	239.17%	88.87%	45.69%

AVERAGE

Time horizon	24 to 47 hr	0 to 23 hr	4 hr	1 hr	15 min
Ave. for 5 sites	**118.71%**	**112.14%**	**143.84%**	**83.51%**	**57.00%**

Table 13: Summary of the RMSE (in MW) forecasting results for the 5 test cases. The validation time period is one year from 2015-03-01 to 2016-02-28. The training of the RF model and situation-dependent blending is performed on data from 2013-03-01 to 2015-02-28.

There are two key results: First, the *Watt-sun* system very successfully enables the forecasting accuracy to reach target value (>100% improvement) for longer forecasting time horizon (over 4 hours). Second, the 1-hour-ahead forecasting performance improvement is close to target, while the 15-minute-ahead is close to the mid-point between baseline and target This trend is not surprising because for the shorter forecasting time horizon, the smart persistence forecast is used as the baseline. The numerical weather models, due to the time required to spin up with data assimilation and uncertainties in the initial boundary conditions, turn out to be often less accuracy compared to persistence. Despite the accuracy improvement enabled by blending the models, the results, though easily outperform smart persistence, still have difficulty reaching target accuracy.

As an example, for the improvement in reliability, the absolute area control error in energy (AACEE) is being reported in Table 14 showing in most cases significant improvements in reliability using the *Watt-sun* forecasting system.

		5.08% penetration			15.24% penetration			25.40% penetration		
		Baseline	Target	Watt-sun	Baseline	Target	Watt-sun	Baseline	Target	Watt-sun
GMP	Value [MWh]	1785	926	908	1861	1387	1339	2569	2034	1832
2DA	Improvement	-	**48.12**	49.13	-	**25.47**	28.05	-	20.83	28.69
GMP	Value [MWh]	1632	925	762	1771	1381	1339	2476	1903	1526
1DA	Improvement	-	43.32	53.31	-	22.02	24.39	-	23.14	38.37
GMP	Value [MWh]	1957	1356	1225	2024	1507	1473	2674	2437	2229
4HA	Improvement	-	30.71	37.40	-	25.54	27.22	-	8.86	16.64
GMP	Value [MWh]	851	756	926	1339	1179	1429	1586	1477	2141
1HA	Improvement	-	11.16	-8.81	-	11.95	-6.72	-	6.87	-34.99
TEP	Value [MWh]	1446	1029	689	1541	1256	992	2317	1982	1381
2DA	Improvement	-	28.84	52.35	-	18.49	35.63	-	14.46	40.41
TEP	Value [MWh]	1563	1117	589	1662	1341	871	2404	2012	1345
1DA	Improvement	-	28.53	**62.32**	-	19.31	**47.59**	-	16.31	**44.05**
TEP	Value [MWh]	1306	849	879	1342	1153	1161	2164	1682	1962
4HA	Improvement	-	34.99	32.71	-	14.08	13.49	-	22.27	9.33
TEP	Value [MWh]	826	607	699	1069	975	1055	1614	1369	1612
1HA	Improvement	-	26.51	15.38	-	8.79	1.31	-	15.18	0.12
Smyr	Value [MWh]	1804	1543	1427	2043	1857	1832	2696	2212	1984
2DA	Improvement	-	14.47	20.91	-	9.11	10.33	-	17.95	26.41
Smyr	Value [MWh]	1765	1499	1188	1949	1852	1668	2603	2025	1946
1DA	Improvement	-	15.07	32.69	-	4.98	14.42	-	22.21	25.24
Smyr	Value [MWh]	1916	1677	1488	2093	1931	1832	2727	2393	1988
4HA	Improvement	-	12.47	22.34	-	7.74	12.47	-	12.25	27.11
Smyr	Value [MWh]	1402	943	1254	1821	1378	1751	1982	1691	1959
1HA	Improvement	-	32.74	10.56	-	24.33	3.84	-	14.68	1.16
ISO-	Value [MWh]	2199	1483	1774	2209	1683	1784	2673	1841	2353
2DA	Improvement	-	32.56	19.33	-	23.81	19.24	-	**31.13**	11.97
ISO-	Value [MWh]	1693	1186	1286	1758	1526	1665	2193	1794	1807
1DA	Improvement	-	29.95	24.04	-	13.21	5.29	-	18.19	17.61
CAIS	Value [MWh]	1985	1258	1061	2048	1901	1787	2625	2216	2079
2DA	Improvement	-	36.62	46.55	-	7.18	12.74	-	15.58	20.81
CAIS	Value [MWh]	1599	991	691	1941	1767	1334	2343	1885	1569
1DA	Improvement	-	38.02	56.79	-	8.96	31.27	-	19.55	33.03

Table 14: Overall results for AACEE under three solar power penetration levels.

3.2.6 Comparison to ECMWF

Besides comparing the performance of the *Watt-sun* system to the baseline and targets it was also compared to ECMWF[69] (European Centre for Medium-Range Weather Forecasts), which is a propitiatory forecasting model costing ~$250,000 annually and considered to be the "gold" standard. The *Watt-sun* forecast system is all applied to provide forecasting with and without the ECMWF model in the blending. ECMWF is updated only twice a day at 0-hour UTC and 12 hour UTC, which made it not useful for intra-day forecasting. The ECMWF also provides most accuracy advantage for longer forecasting time horizon (2DA and beyond). Figure 17 summarizes 2DA power forecast error (MAE) for the three-point test sites, Smyrna, TEP, and GMP Post road sites. The comparison is between 1st order learning corrected ECMWF model (red bar), situation-dependent model blending using NOAA public models without ECMWF (green bar), and with ECMWF (blue bar). The test time is 2014-12-1 to 2015-06-30 for the Smyrna site and 2015-1-1 to 2015-12-31 for TEP FRV and GMP Post road sites. The situation dependent blending excluding ECMWF provided 21% improvement (target 15%) upon 1st order corrected ECMWF. This shows the value of the blending methodology – by blending three public models, the forecast accuracy surpasses ECMWF. With ECMWF included in the forecasting itself, the blended forecasting shows 24% improvement (close to the target of 30%) upon 1st order corrected ECMWF.

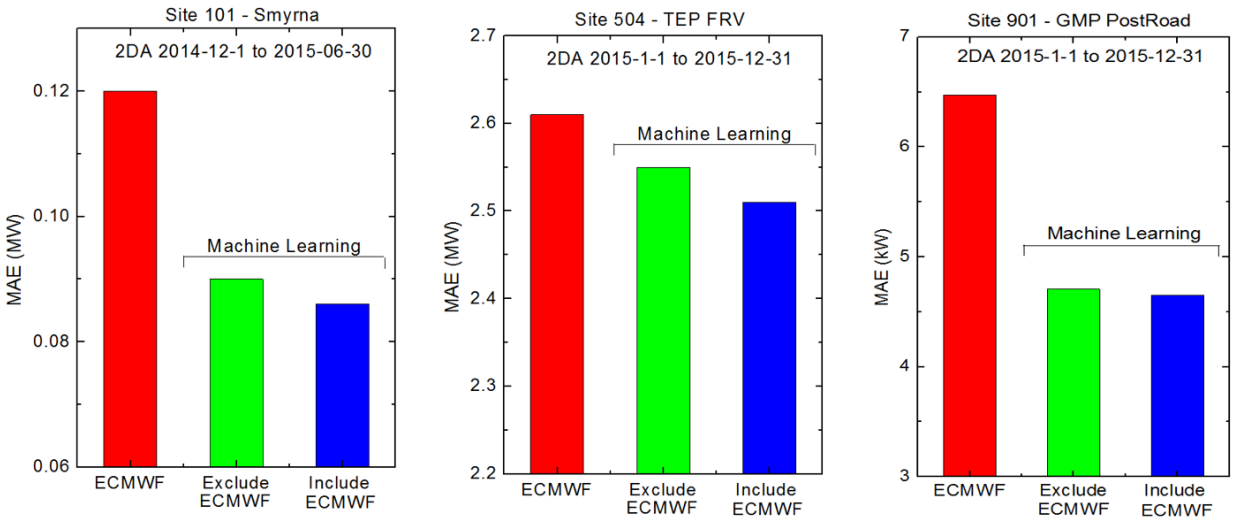


Figure 17: Summary of 2 day-ahead AC power forecast error (MAE) for the three "point-test" sites, Smyrna, TEP, and GMP Post road sites using different methods. Red: 1st order learning corrected ECMWF mode. Green: Situation dependent model blending using NOAA public models (NAM, GFS, and SREF) without ECMWF. Blue: Situation dependent model blending using NOAA public models and ECMWF.

3.2.7 Nation-wide Solar Forecasting

For forecasting for the entire US we leveraged the Remote Automatic Weather Stations (RAWS) system, which is a dense network (~1600 in the continental US) of weather stations run by the U.S. Forest Service and Bureau of Land Management.[70] RAWS provides hourly global horizontal irradiance measurements. Figure 18 shows next to the RAWS instrumentation a map of all RAWS stations, which we are using for developing a gridded forecast.

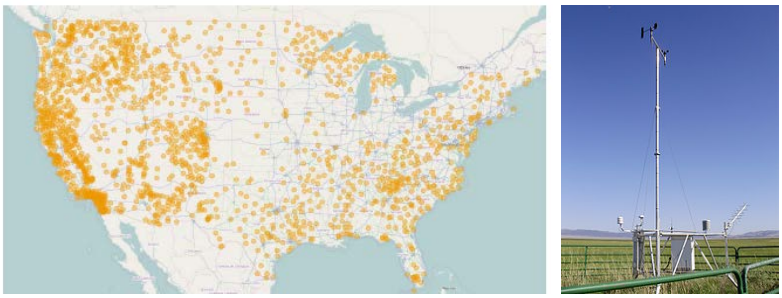


Figure 18: Location of the RAWS stations across the US (left) and RAWS instrumentation (right).

A critical step for developing a gridded forecast lies in the understanding whether or how proxy measurement sites can used for training / machine-learning (rather than actual co-located measurement sites). We note that this is – beyond the goal of developing gridded forecasts – a very important aspect of the *Watt-sun* system because if one could use proxy sites, the requirement of actual measurements from the forecast sites is not as stringent anymore and the applicability of *Watt-sun* would be clearly broadened.

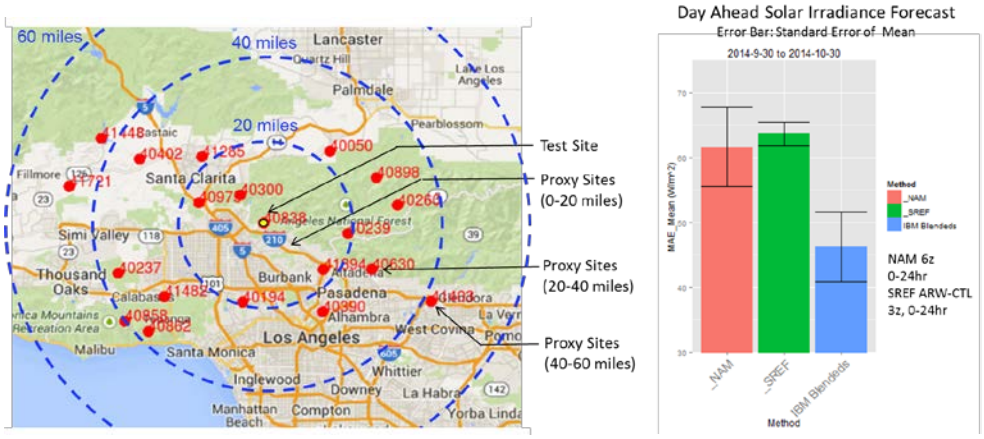


Figure 19: RAWS station in the Los Angeles area (left) and forecasting results (i.e., mean absolute error) for the target site just using proxy sites (right).

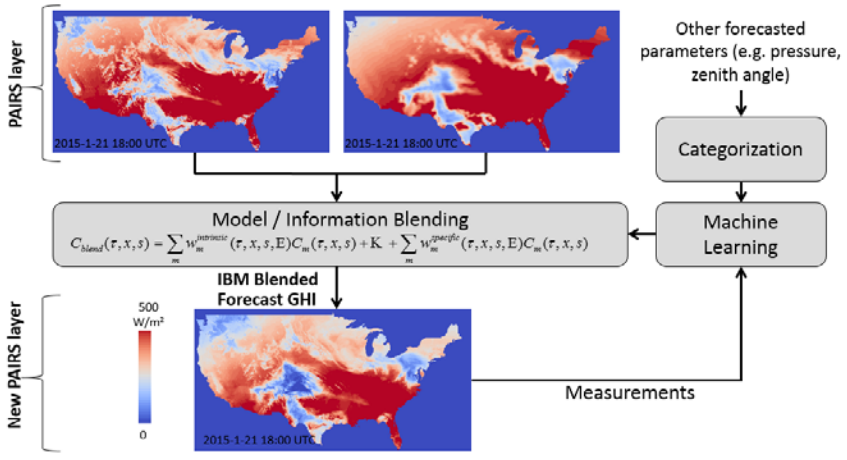


Figure 20: Illustration of the blending approach to develop a gridded forecast.

Figure 19 illustrates an example of the studies, which we have undertaken to understand the viability of the use of proxy sites for our machine-learning approach. In this Figure, we show different RAWS sites in the Los Angeles metro area with a "target" site in the center. We note that we have 5 proxy sites within 20 miles of the test site. We have an additional 10 RAWS stations within 40 miles and yet another 3 stations within 60 miles. Figure 19 shows on the right-side irradiance (GHI) forecast errors for the target site where we have used only the proxy sites within 40 miles. We compare the *Watt-sun* forecast error with the NAM and SREF forecast error demonstrating more than 25 % improved accuracy (less error).

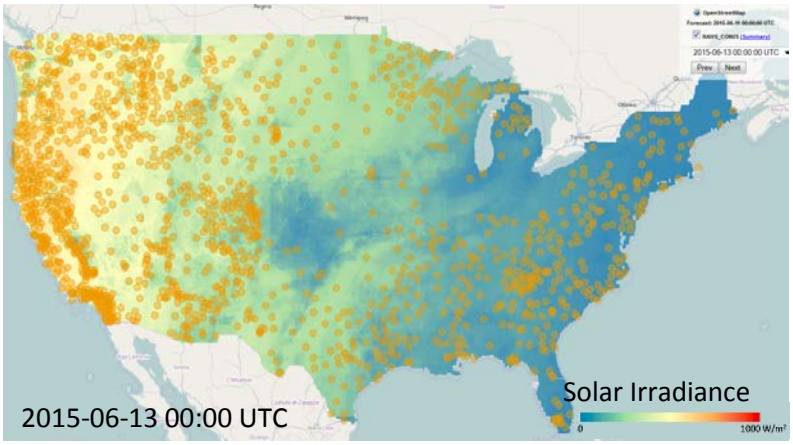


Figure 21: It shows a 48 hour ahead forecast of global horizontal irradiance (color scale) of contiguous US obtained via machine-learning based situation-dependent blending of two weather models – the north American mesoscale (NAM) model and the short-range ensemble forecast (SREF) model. This forecast is issued at 2015-06-11 00:00 UTC for 2015-06-13 00:00 UTC. The model blending is trained by historical forecasts and measurements at ~1600 remote automatic weather stations (RAWS) of the MesoWest network (yellow circles).

The results shown in Figure 19 certainly show how we can use our machine-learning approach to develop gridded forecasts. Clearly, as further away the proxy sites are from the actual test site the less viable the approach is but it is certainly not only distance. For understanding the validity of a proxy site, we studied systematically the correlations between the RAWS station across the whole continental US to develop a sophisticated map of weights. A paper is being prepared describing this approach. Figure 20 shows an example, where we have blended two GHI forecasts (NAM and SREF) using more than 1600 RAWS sites across the country.

Figure 21 shows a screenshot of the web interface, which was developed to share these continental US-wide irradiance forecasts. Each point (yellow circle) shows a RAWS station, where the performance of the *Watt-sun* system is being reported and compared with other weather models – using the metrics which were developed earlier in this project. In average, we show more than 30 % improvements across the 1600 sites. The data is available at https://pairs.res.ibm.com/.

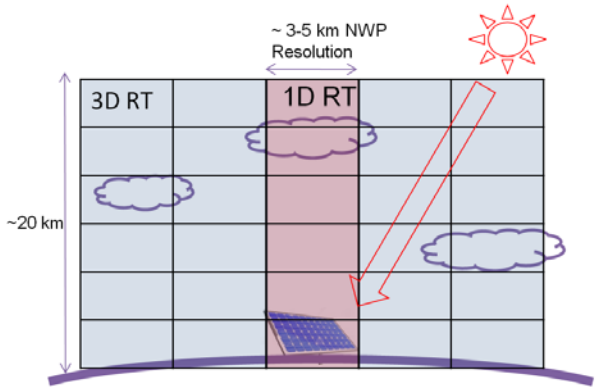


Figure 22: Schematic showing the necessity of a 3D radiative transfer model given the apparent deficiency of a 1D model in a partially cloudy scenario.

3.2.8 3D Radiative Transfer

A three-dimensional (3D) radiative transfer model was also developed for the *Watt-sun* forecast system. Towards that end, first we investigated the "potential" improvements of a 3D RTM versus a 1D or 2D model, which is what most forecasting systems are using today. Namely, the atmospheric states (pressure, temperature, cloud, and aerosol) are assumed to be uniform horizontally. As illustrated in Figure 22, such simplification has clear limitations in partial cloudy condition in which the solar irradiation may penetrate through "clear sky" between clouds and reach the solar panels. The 1D model cannot take full advantage of the high horizontal resolution of the latest cloud-resolving numerical weather prediction (NWP) models. To enhance forecasting accuracy, needed is a 3D model capable of handling horizontal distribution of the clouds as predicted by NWP models.

Our work is based upon (a) the existing 1D RT module and (b) the open source MCARATS (Monte Carlo Atmospheric Radiative Transfer Simulator). MCARATS is a general purpose 3D radiative transfer simulator.[71] MCARATS is essentially a Monte Carlo solver of the RT equations. It cannot read inputs from NWPs and it does not contain the necessary parameterization of wavelength dependent scattering or absorption by gas species, aerosols, and clouds etc. Thus, the existing 1D RT module has been modified to parse NWP inputs into 3D grid and supply the necessary scattering/absorption parameterization; then it calls the MCARATS to solve the RT equations, and finally read out the results from MCARATS. The accuracy of the 3D RT module with respect to the 1D module has been tested using 6 months of measurement data (2015-1-1 to 2015-6-30) on the NOAA SurfRad sites as summarized below. The NOAA 5 km NAM model (6z run 0 to 24 hour ahead) is used as the input NWP. For fair comparison, the output of both RT models are used as-is. There is no statistical correction applied.

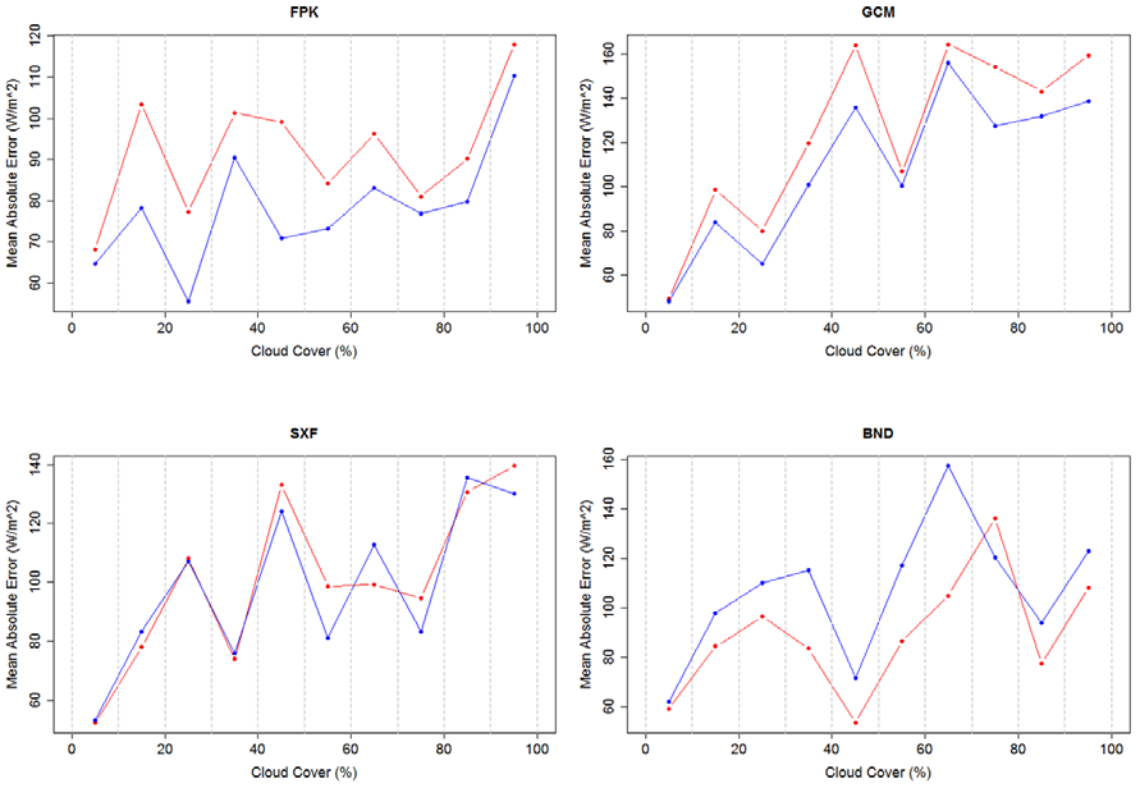


Figure 23: Comparison of 1D (red) and 3D (blue) radiative transfer models as a function of cloud cover percentage. The four panels show mean absolute error results for four SurfRad stations. Time period is 2015-1-1 to 2015-6-30. Inputs for the radiative transfer calculation are provided by the 5 km NAM model 06z run 0 to 24 hour ahead (except aerosol and surface albedo are provided MODIS).

Site	BND		TBL		FPK		GCM		PSU		SXF	
RT model	1D	**3D**	1D	**3D**	1D	**3D**	1D	**3D**	1D	**3D**	1D	**3D**
Capacity (W/m^2)	1000	**1000**	1000	**1000**	1000	**1000**	1000	**1000**	1000	**1000**	1000	**1000**
pcoe	0.86	**0.802**	0.808	**0.767**	0.896	**0.872**	0.807	**0.774**	0.83	**0.814**	0.837	**0.803**
RMSE (W/m^2)	153	**154**	204	**182**	147	**122**	206	**184**	174	**161**	170	**155**
NRMSE	0.153	**0.154**	0.204	**0.182**	0.147	**0.122**	0.206	**0.184**	0.174	**0.161**	0.17	**0.155**
MaxAE (W/m^2)	618	**586**	875	**698**	650	**599**	732	**741**	815	**834**	708	**730**
MAE (W/m^2)	93.9	**114**	133	**134**	100	**83**	141	**123**	110	**110**	109	**108**
MAPE	0.0939	**0.114**	0.133	**0.134**	0.1	**0.083**	0.141	**0.123**	0.11	**0.11**	0.109	**0.108**
MBE (W/m^2)	76.2	**2.02**	110	**2.88**	91.1	**-4.63**	116	**42**	83.5	**29.3**	87.5	**7.39**
KSIPer	15.758	**4.242**	17.391	**4.831**	15.453	**3.311**	19.595	**8.784**	14.861	**8.564**	15.122	**4.878**
StdDev (W/m^2)	133	**154**	172	**182**	115	**122**	171	**180**	153	**159**	146	**155**
Skewness	1.76	**0.697**	1.36	**0.64**	1.6	**0.856**	1.12	**0.683**	1.72	**1.31**	1.16	**0.74**
Kurtosis	3.45	**1.62**	2.39	**1.47**	3.48	**3.62**	1.43	**2.2**	4.16	**3.59**	3.21	**3.23**
RMQE_4 (W/m^2)	251	**227**	317	**265**	230	**194**	307	**283**	289	**265**	267	**246**
NRMQE_4	0.251	**0.227**	0.317	**0.265**	0.23	**0.194**	0.307	**0.283**	0.289	**0.265**	0.267	**0.246**
Percentile95(W/m^2)	382	**319**	486	**360**	329	**199**	451	**353**	383	**343**	374	**284**

Table 15: Comparison of 1D and 3D radiative transfer accuracy for six SurfRad stations under partially cloudy condition (cloud cover between 10% to 90%). Time period 2015-1-1 to 2015-6-30. Inputs for radiative transfer are provided by the 5 km NAM model 06z run 0 to 24 hour ahead (except aerosol and surface albedo are provided MODIS). TBL= Table Mountain, Boulder, Colorado, PSU=Penn State University, Pennsylvania

Figure 23 below shows the comparison of the two models as a function of NAM reported cloud cover percentage for four SurfRad sites. For near zero cloud cover, the 1D (red) and 3D (blue) models have comparable mean absolute error (MAE), which is not surprising since the 3D model essentially reduces to a 1D where there is no horizontal variation of cloud cover. For higher cloud cover, the 3D model has overall better performance. For example, as shown in Figure 23, the FPK, GCM sites shows a substantial error reduction for cloud cover greater than 10%. For SXF site the two models have comparable MAE. BND is an outlier for which the 1D model has smaller MAE.

The full suite of metrics [4,5] for the 1D/2D and 3D RT models is calculated for time period 2015-01-01 to 2015-06-30 and shown in Table 15. The comparison is done for six SurfRad sites in partial cloudy hours (cloud cover 10% to 90%), which usually have the larger irradiance forecast error compared to overcast (100% cloud cover) or clear sky conditions. Note that among the seven SurfRad stations, the DRA site located in Nevada desert is left out in this comparison since it does not have statistically significant number of cloudy hours in the 6-month time period.

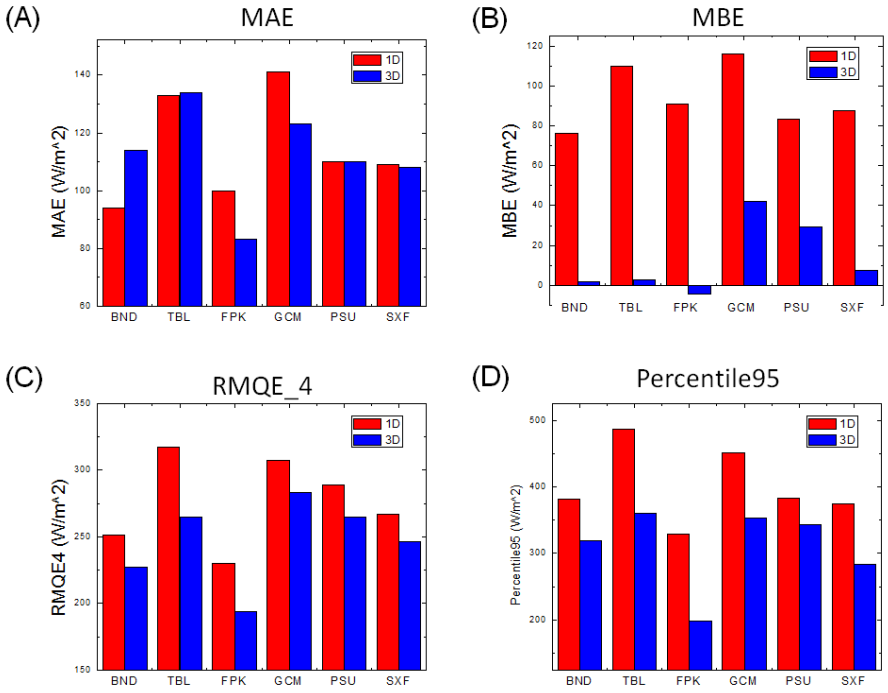


Figure 24: Comparison of the accuracy of 1D (red) vs. 3D (blue) radiative transfer models under partially cloudy condition (cloud cover between 10% to 90%). Comparison of four key metrics, (a) mean absolute error, (b) mean bias error, (c) root mean quartic error, and (d) the 95th percentile of forecast error is shown.

From Table 15, the comparison of four selected metrics is presented in Figure 24 below. Figure 24A shows the mean absolute error comparison. Among the six sites, two sites (FPK and GCM) have markedly reduced MAE using the 3D model, three sites (TBL, PSU, SXF) have comparable MAE using the 1D or the 3D model, while BND site has worse MAE using the 3D model. The average MAE is 114.5±18 W/m^2 and 112.0±17 W/m^2 for the 1D and 3D model, respectively. No statistically significantly reduction of MAE is achieved using the 3D model. In contrast to the MAE, however, the bias of the 3D model is significantly less than the 1D model (Figure 24B). The mean bias error (MBE) of the six sites on average is 94.0±16 W/m^2 for the 1D model and 13.2±18 W/m^2 for the 3D model.

Moreover, importantly for the use-cases of the forecasts, the metrics results (Table 15) shows that the 3D model significantly reduces the occurrence of the relatively large forecast error as characterized by the root mean quadratic error (RMQE) metrics as shown in Figure 24C. All six sites show substantial reduction of RMQE using the 3D RT model. The average RMQE is 276.8±33W/m^2 for the 1D model and 246.6±32W/m^2 for the 3D model, a reduction of over 10%. As well known, such large forecast errors are the most significant for economic value of the forecast. Towards this end we look at the 95th percentile of forecast errors (Percentile95), which represents the amount of non-spinning reserves required to compensate the forecast error. As shown in Figure 24D, all six sites show substantial reduction of Percentile95. The average Percentile95 is 400.8±57W/m^2 for the 1D model and 309.6±61W/m^2 for the 3D model, i.e. a reduction of non-spinning reserve by over 25%.

3.2.9 Short-term forecasting

A new algorithm for short-term solar energy forecasting from a sequence of GOES satellite images was developed and implemented. Conventionally short-term forecasting algorithms[36,72] perform cloud advection using either (a) wind velocity field derived from numerical weather prediction (NWP) models or (b) optical flow analysis of a sequence of satellite images. In the former case, even the NWP model is perfectly accurate, the error in determining cloud height may lead to large error of the velocity field. In the latter case, the wind velocity field is assumed to be static and does not reflect the dynamics of the wind in the future.

	Products	Sets of Sites	Forecasted Parameters	Measurement Available
ISONE	GHI forecast 0 to 3 DA	RAWS Point Sites (15)	Hourly averaged GHI	Hourly averaged GHI
ISONE	POA Irradiance Forecast 0 to 3 DA	Solren Point Sites with Irradiance Measurements (19)	Hourly averaged GHI Hourly averaged POA Irradiance	Hourly averaged POA Irradiance
ISONE	Load Zone Forecast 0 to 3 DA	LoadZone Forecasts (8)	Hourly averaged PV normalized by nameplate AC	Hourly averaged PV normalized by nameplate AC from ~900 solren sites.
GMP	PV Forecast 0 to 1 DA	PV Point Sites (14)	Hourly averaged PV normalized by nameplate AC	Hourly averaged PV normalized by nameplate AC

Table 16: Forecasting products for stakeholders (POA=plane of array).

In contrast, the new algorithm combines optical flow and 2D Navier-Stokes Equation (NSE) which not only accurately captures the current wind velocity field but also, to a certain degree, the wind dynamics in the future. The algorithm is inspired by the following observation of satellite cloud images: the dynamics of clouds (represented by cloud optical depth (COD)) resembles the motion (transport) of a density in the fluid flow. This suggests that, to forecast the motion of COD images, a parametric model of the fluid flow can be “learned” from the COD images, observed in the past, to forecast the fluid flow. Hence, the learning phase of the algorithm is composed of the following two steps: (1) optical flow estimation: given a sequence of COD images, the snapshots of the optical flow based velocity fields are estimated from two consecutive COD images using standard optical flow techniques. (2) Navier-Stokes Equation (NSE) fitting: these snapshots are then assimilated into a NSE, i.e. an initial velocity field for NSE is selected so that the corresponding NSE’ solution is as close as possible to a sequence of optical flow snapshots of velocity fields. The prediction phase consists of utilizing a linear transport equation, which describes the transport of COD images in the fluid flow predicted by NSE, to estimate the future motion of the COD images. Using the algorithm, we demonstrated around 30% error reduction of irradiance forecasting on one-hour ahead time scale with respect to smart persistence and NWP models on typically partially cloudy days. Efforts are underway to implement the algorithm in C++ (current

implementation in Matlab) on a parallel computing platform and then to port it into the *Watt-sun* forecasting system.

3.3 Integration of the *Watt-sun* Forecasting System

An important task was to deliver operational forecasts to the ISO-NE and GMP. More specifically, the forecast products listed in Table 16.

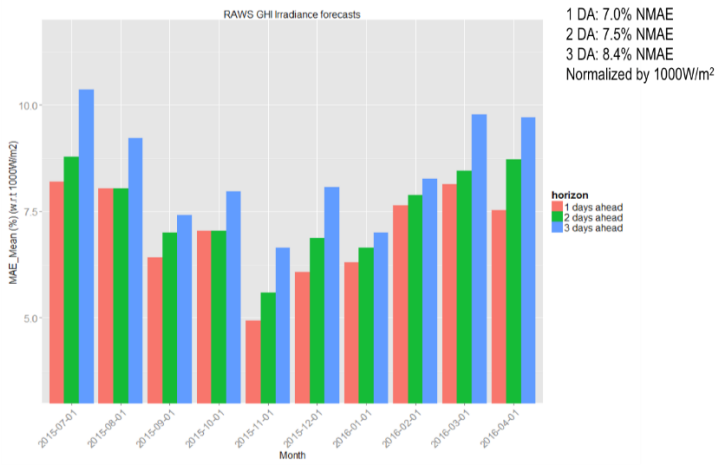


Figure 25: ISO-NE forecasting accuracy of GHI at RAWS point sites.

Figure 25 shows the ISO-NE GHI forecasting accuracy of the 1DA, 2DA, and 3DA GHI at RAWS point sites between July 2015 and April 2016. The average normalized forecasting MAE errors at the 1DA, 2DA, 3DA horizons are 7.0%, 7.5%, and 8.4%, respectively. For 1DA forecasts, the normalized MAE errors are varying between approximately 5% and 8%. Figure 26 shows the plane of array (POA) irradiance forecasting accuracy of the 1DA, 2DA, and 3DA GHI at Solren point sites between July 2015 and April 2016. The average normalized forecasting MAE errors at the 1DA, 2DA, 3DA horizons are 7.5%, 8.3%, and 9.4%, respectively. Figure 27 shows the forecasting accuracy of the 1DA, 2DA, and 3DA PV power at different load zones of ISO-NE. The MAE values are normalized by the nameplate power from ~900 Solren sites. The 1DA PV power forecasting MAE errors are varying between approximately 2% and 7% among different load zones. Most of the 2DA and 3DA forecasting MAE errors are below 10%. Figure 28 shows the forecasting accuracy of the 1DA PV AC power at different point sites at GMP. The corresponding nameplate PV capacities normalize the MAE values. The 1DA PV power forecasting MAE errors are varying between approximately 4% and 10.5% at different PV sites.

The accurate forecasting is an essential tool for facilitating the integration of solar photovoltaic (PV) power into the bulk power system. The quantification of the practical benefits of the forecasts from the perspectives of both reliability and economic value were performed using a multi-timescale power system operation model was discussed earlier. The representative IEEE 118-bus system has been adopted to simulate 400 scenarios with different levels of improvements, locations, forecast horizons, and penetration levels. The simulations show that: (i) *Watt-sun* forecasts perform better (compared to baseline forecasting) for most cases in terms of power system reliability performance; (ii) reliability benefits are gradually enhanced by the solar power forecasting improvement in multi-timescale power system operations; and (iii) benefits of improved forecasting increases drastically with higher penetration levels of solar energy.

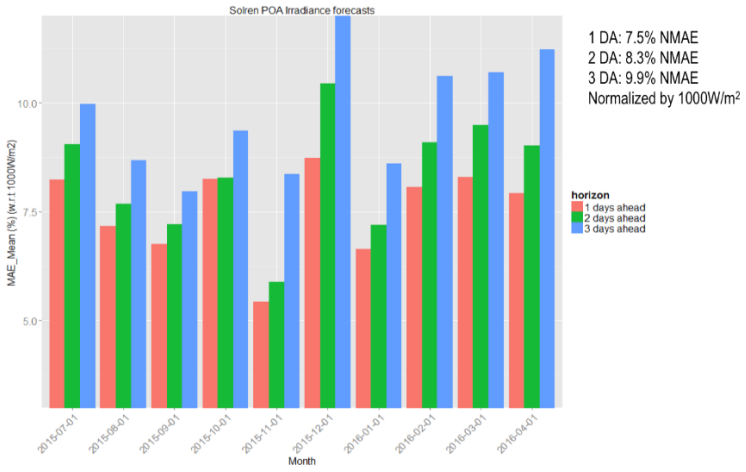


Figure 26: ISO-NE forecasting accuracy of POA irradiance at Solren point sites

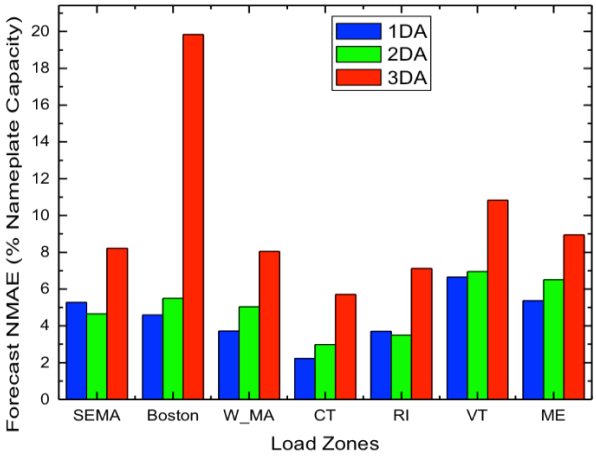


Figure 27: ISO-NE PV power forecasting accuracy at different load zones

Moreover, through interaction with ISO-NE, the team discovered that snow detection plays an important role in winter time. Snow cover induces low PV output (despite high irradiance) and can drastically reduce forecasting accuracy. The negative effect can be long lasting due to the contamination of training dataset. At the request of ISO-NE, a methodology for snow detection are developed and tested. A detailed manuscript is being prepared describing the details of this method.

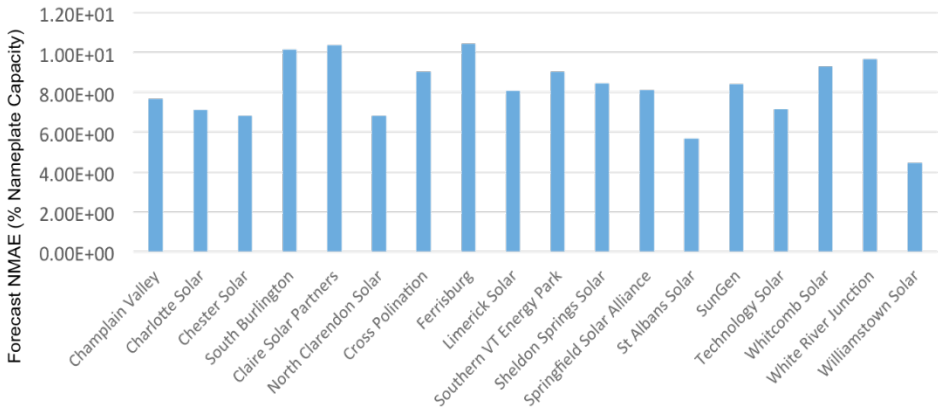


Figure 28: GMP- PV_{AC} power point forecast accuracy

4. Significant Accomplishments and Conclusions

The projected has resulted into several key accomplishments.

1. The work has led to a commonly accepted method and set of metrics how to measure the accuracy of solar forecasting, which is a very important step towards further developing improved forecasting methods. The metrics include statistical, uncertainty quantification, ramp characterization, economic and reliability metrics. Even more important a methodology was developed how to determine baseline and target values for these metrics, which can now be used to compare forecasts and set "expectations" of a standard and a state-of-art forecast, respectively.
2. Most importantly, the team advanced the state of forecasting significantly. The noticeable feature of the *Watt-sun* forecast system is that a set of appropriately chosen parameters is used to create different weather situation categories in which the input models exhibit different error characteristics. This approach (i.e., situation dependent, machine-learnt, multi-model blending has been demonstrated to advance the accuracy of the next best standard bias corrected model by more 30% and by more than 15% compared to a DiCast approach. These improvements are significant given the fact that traditional the accuracy of forecasting has only improved by ~6% each decade.[3] The *Watt-sun* forecasts were provided operational to the ISO-New England and Green Mountain Power throughout the project.
3. The *Watt-sun* forecasts met in average the target values for all 5 test sites and for all time horizons using the set of metrics, which were developed in this project. However, at short forecast horizons (1 hour ahead and 4 hour ahead) the *Watt-sun* forecast improvements have been less than for longer forecast horizons. This is at least partially due to the fact that "smart persistence" as a baseline is already quite accurate.
4. The Watt-sun system is being ported to the Physical Analytics Integrated Data Repository and Services (PAIRS)[49,73], which is completely scalable data and analytics platform and which may provide a smooth way for further commercializing the developed technology and to integrate it into other IBM offerings.
5. The team published 7 full papers, participated in 25 conferences, conducted 2 public webinars (one on the metrics and another one of the *Watt-sun* system), organized jointly

with the NCAR team and DoE 2 special sessions at annual UVIG conference, won 3 awards and had countless press mentioning. A replicate of *Watt-sun* forecasting system was also created at the National Renewable Energy Laboratory to ensure that the technology can be used for the public good.

5. Inventions, Patents, Publications

5.1 Full papers

1. A Methodology for Quantifying Reliability Benefits from Improved Solar Power Forecasting in Multi-Timescale Power System Operations
   M. Cui, J. Zhang, B.-M. Hodge, S. Lu, and H. F. Hamann
   submitted to IEEE Transactions on Smart Grid
2. Machine Learning based Situation-Dependent Multi-Model Blending for Enhancing Renewable Energy Forecasting
   S. Lu, Y. Hwang, I. Khabibrakhmanov, X. Shao, A. Florita, C. B. Martinez-Anido, B.-M. Hodge, J. Zhang, E. F. Campos, and H. F. Hamann
   submitted to Solar Energy.
3. The value of day-ahead solar power forecasting improvement
   C.B. Martinez-Anido, B. Botor, A. Florita; S. Lu; H.F. Hamann; B.-M. Hodge
   Solar Energy **129**, 192 (2016)
   doi:10.1016/j.solener.2016.01.049
4. On the usefulness of solar energy forecasting in the presence of asymmetric costs of error
   I. Khabibrakhmanov, S. Lu, H. F. Hamann, K. Warren
   IBM J. Res. & Dev. **60**, 7:1 (2016)
   doi:10.1147/JRD.2015.2495001
5. Baseline and Target: Road to a Better Solar Power Forecasting
   J. Zhang, S. Lu, B.-M. Hodge, H.F. Hamann, B. Lehman, J. Simmons, E. Campos
   Solar Energy **122**, 804 (2015)
   doi:10.1016/j.solener.2015.09.047
6. A Suite of Metrics for Assessing the Performance of Solar Power Forecasting
   J. Zhang, B.-M. Hodge, A. Florita, S. Lu, H.F. Hamann, V. Banunarayanan, A. Brockway
   Solar Energy **111**, 157 (2015).
   doi:10.1016/j.solener.2014.10.016
7. Recent Trends in Variable Generation Forecasting and Its Value to the Power System
   K. D. Orwig, M. Ahlstrom, V. Banunarayanan, M. Marquis, J. Sharp, J. Wilczak, J. Freedman, S. Haupt, J. Cline, O. Bartholomy, D. Bartlett, H.F. Hamann, Bri-M. Hodge, C. Finley, D. Nakafuji, J. Peterson, D. Maggio
   IEEE Transaction on Sustainable Energy **99**, 1 (2014).
   doi:10.1109/TSTE.2014.2366118

5.2 Conferences

1. Short-term Global Horizontal Irradiance Forecasting Based on Sky Imaging and Pattern Recognition (best paper award)
   C. Feng, M. Cui, M. Lee, J. Zhang, B.-M. Hodge, S. Lu, and H. F. Hamann
   IEEE Power & Energy Society General Meeting (2017) (accepted)

2. A machine-learning approach for regional photovoltaic power forecasting
Li, Yuan, Qian Sun, Brad Lehman, Siyuan Lu, Hendrik F. Hamann, Joseph Simmons
IEEE Power & Energy Society General Meeting (2016)
doi: 0.1109/PESGM.2016.7741991

3. Solar Irradiance Forecasting by Machine Learning for Solar Car Races
X. Shao, S. Lu, T.G. van Kessel, H.F. Hamann, L. Daehler, J. Cwagenberg, A. Li
IEEE Big Data (2016)
doi: 10.1109/BigData.2016.7840851

4. Solar radiation forecast with machine learning
X. Shao, S. Lu, H.F. Hamann
The 23rd International Workshop on Active-Matrix Flatpanel Displays and Devices, 19-206 (2016).
doi:10.1109/AM-FPD.2016.7543604

5. Smart Solar Field Instrumentation for Development of Site-Specific Irradiance to Power Models
J.C. Simmons, C. Bokrand, B.J. Potter, S.Lu, H.F. Hamann
5th PV Performance Modeling Workshop (2016).

6. Physical Analytics: Bringing big data together with physics (invited)
H.F. Hamann
Physics Informed Machine Learning (2016).

7. DoE Solar Forecasting Project: Progress in Short-Term Forecasting (invited)
H.F. Hamann
UVIG Forecasting Workshop (2016).

8. DoE Solar Forecasting Project: Overview of the IBM Project (invited)
H.F. Hamann
UVIG Forecasting Workshop (2016).

9. A Two-Dimensional Gridded Solar Forecasting System using Situation-Dependent Blending of Multiple Weather Models
S. Lu, Y. Hwang, I. Khabibrakhmanov, X. Shao, H.F. Hamann
American Geophysical Union Fall Meeting, A11H-0164 (2015).

10. Towards Gridded Foundational Solar Forecast of Enhanced Accuracy: Weather “Situation” Dependent Forecast Error and Machine-Learnt Multi-Model Blending
S. Lu, Y. Hwang, X. Xiao, H. F. Hamann
3rd International Conference Energy and Meteorology (2015).

11. Improvement of Solar Irradiance Forecast Using Machine Learning
S. Lu, X. Shao, Y. Hwang, I. Khabibrakhmanov, H. F. Hamann
9th Annual Machine Learning Symposium of the New York Academy of Sciences (2015).

12. Physical Analytics: An emerging field with real-world applications and impact (invited)
H.F. Hamann
American Physical Society March Meeting (2015).

13. Situation-dependent blending of multiple forecasting models based on machine learning (invited)
H.F. Hamann, S. Lu
SPIE Newsroom (2015).
doi:10.1117/2.1201510.00614

14. Baseline and target values for PV forecasts: Toward improved solar power forecasting
J.Zhang, B.M. Hodge, S.J. Simmons, S. Lu, E. Campos, B. Lehman, V. Banurarayan
Proceedings of IEEE Power & Energy Society General Meeting, 1 (2015).
doi:10.1109/PESGM.2015.7286239

15. Ensemble Solar Forecasting Statistical Quantification and Sensitivity Analysis
W. Cheung, J. Zhang, A. Florita, B.-M. Hodge, S. Lu, H. F. Hamann, Q. Sun, B. Lehman
5th Solar Integration Workshop: International Workshop on Integration of Solar Power into Power Systems, Brussels, Belgium (2015).

16. A Multi-faceted approach towards solar forecasting
E. Campos, E. Constantinescu, Y. Feng, J. Wang, Z. Zhou, A. Botterud, D. Cook, H.F. Hamann, S. Lu
95th Annual Conference of the American Meteorological Society, Phoenix (2015).

17. Machine Learning Based Multi-Physical-Model Blending for Enhancing Renewable Energy Forecast – Improvement via Situation Dependent Error Correction
S. Lu, Y. Hwang, I. Khabibrakhmanov, F. J. Marianno, X. Shao, J. Zhang, B. Hodge, H F. Hamann
95th Annual Conference of the American Meteorological Society, Phoenix (2015).

18. Situation Dependent Machine Learning based Multi-Model Blending for Enhancing Renewable Energy Forecasting (invited)
H.F. Hamann
2015 Forecasting Workshop, Denver, Colorado (2015).

19. Machine Learning Based Multi-Physical-Model Blending for Enhancing Renewable Energy Forecast – Improvement via Situation Dependent Error Correction (invited)
S. Lu, Y. Hwang, I. Khabibrakhmanov, F. J. Marianno, X. Shao, H.F. Hamann
European Journal of Control Conference, Linz, Austria (2015).
doi:10.1109/ECC.2015.7330558

20. A multi-scale solar energy forecast platform based on machine-learned adaptive combination of expert systems (invited)
H.F. Hamann
2014 Forecasting Workshop, Tucson, Arizona (2014)

21. Solar Forecast Improvement Project: A Public-Private Collaboration
M. Marquis, S. Benjamin, E. James, A. Heidinger, C. Molling, J. Michalsky, K. Lantz, V. Banunarayanan , S. Haupt, H. F. Hamann
4th International Workshop on Integration of Solar Power into Power Systems, Berlin, Germany (2014).

22. A multi-scale solar energy forecast platform based on machine-learned adaptive combination of expert systems
S. Lu, J. Lenchner, G. J. Tesauro, C. M. Corcoran, F. J. Marianno, J. Zhang, B.-M. Hodge, E. Campos, H. F. Hamann
American Meteorological Society 2014 Annual Meeting, Atlanta (2014).

23. Metrics for Evaluating the accuracy of solar power forecasting
J. Zhang, Bri-M. Hodge, A. Florita, S. Lu, H.F. Hamann, V. Banunarayanan
3rd International Workshop on Integration of Solar Power into Power Systems, London, UK (2013).

24. Creating a Standard Set of Metrics to Assess Accuracy of Solar Forecasts: Preliminary Results
V. Banunarayanan, A. Brockway, M. Marquis, S. Haupt, B. Brown, T. Fowler, T. Jensen, H.F. Hamann, S. Lu,B. Hodge,J. Zhang, A. Florita
American Geophysical Union Fall Meeting A13G-0306 (2013).

25. DoE Solar Forecasting Project: A Multi-scale, Multi-Model, Machine-Learning Solar Forecasting Technology(invited)
H.F. Hamann
American Meteorological Society 2013 Annual Meeting, Austin (2013).

5.3 Patents

1. US Patent: 9,471,884; H.F. Hamann, Y. Hwang, T.G. van Kessel, I.K. Khabibrakhmanov, S. Lu, R. Muralidhar
   Multi-model blending
2. US Patent Application: 20150186904; S. Guha, H.F. Hendrik, K.I. Klein, S.A. Bermudez Rodriguez
   System and Method for Managing and Forecasting Power From Renewable Energy Sources - pending
3. US Patent Application: 20140327769; H.F. Hamann, S. Lu
   Multifunctional Sky Camera System for Total Sky Imaging and Spectral Radiance Measurement - pending
4. US Patent Application: 20140324352; H.F. Hamann, S. Lu
   Machine Learning Approach for Analysis and Prediction of Cloud Particle Size and Shape Distribution - pending
5. US Patent Application: 20140324350; H.F. Hamann, S. Lu
   Machine Learning Approach for Analysis and Prediction of Cloud Particle Size and Shape Distribution - pending
6. US Patent Application: 20140320607; H.F. Hamann, S. Lu
   Multifunctional Sky Camera System for Total Sky Imaging and Spectral Radiance Measurement

5.4 Press (Selected)

1. Vu C (2015) Machine learning helps IBM boost accuracy of US Department of Energy solar forecasts by up to 30 percent.
2. Staff (2015) Better Solar and Wind Forecasting. (Energy Matters).
3. Staff (2015) IBM Boosts Accuracy of DOE Forecasts by 30%. (Solar Industry Magazine).
4. Staff (2015) Interview with Hendrik Hamann, Physical Analytics Manager at IBM Research. (AltEnergyMag).
5. Staff (2015) IBM Improves Solar Forecasts with Machine Learning. (Inside HPC).
6. Solomon DB (2015) Machine 'learners' compute cloud cover to balance power supplies. (Los Angeles Times).
7. Mearian L (2015) IBM's machine-learning crystal ball can foresee renewable energy availability. (Computer World).
8. Martin R (2015) Solar and wind forecasts are new wave of weather reporting. (Mashable).
9. Martin R (2015) Weather Forecasting Enters a New Era. (MIT Technology Review).
10. Hock L (2015) Machine Learning’s Impact on Solar Energy. (R&D Magazine).
11. Hall-Geisler K (2015) Solar Race Team Gets Help from a Superforecast. (Popular Science).
12. Glasner J (2015) IBM's Machine Learning Tech Takes on Solar Power's Flakiness. (Data Center Knowledge).

5.5 Awards

2017 Best Conference Paper ("Short-term Global Horizontal Irradiance Forecasting Based on Sky Imaging and Pattern Recognition") submitted to the 2017 Power & Energy Society General Meeting

2017 UVIG Achievement Award

2016 Industrial Physics Award from the American Institute of Physics

6. Path Forward

There are a couple of important directions how to move forward with this project

- Certain capabilities must be added for advancing the *Watt-sun* system: (1) The technology must be expanded to support automatic generation of probabilistic forecasts. For this, various techniques should be explored, for example whether to train for each quantile separately or using appropriate machine-learning algorithms to quantify directly the uncertainty for a given result. (2) The work with ISOs and utilities demonstrated that there must be a tighter integration between the forecasts and the actual use case for the end user. Examples for this would be to provide ramp forecasting products directly (rather than for example DNI etc) or perhaps regional solar load modification forecasts. (3) Another key area for additional research is to improve the accuracy of the *Watt-sun* system for short-term forecasting, which would require exploring even more and bigger data sources (see discussion below) and models such as improved cloud tracking mechanisms.
- One of the big lessons learnt from this project was that the data which is required for better and more advanced machine-learning, requires a more sophisticated compute infrastructure than originally envisioned. While the originally developed "big data bus" fulfilled all the requirements of this project, it will not be sufficient for the future (the data volume for the WRF models just used in this research has increased by more than 20x in the last 4 years). Towards that end, IBM has started to develop separately a very powerful big data platform for geo-spatial data processing and analytics (PAIRS= Physical Analytics Integrated Data Repository and Services). This platform has two very important features, by which it differentiates itself. First, the computation or processing is "independent" of data size because the computation is done without moving the data. Second, due to unique indexing scheme the platform also provides contextual data to support the various use cases for the power industry. One of the next technical tasks will be to "port" the *Watt-sun* system completely onto the PAIRS platform, which is also the conduit for commercializing the technology.
- Another very interesting area of research which has emerged from this work is the notion of "physics-informed machine-learning". Evidently, the work presented herein constitutes an interesting example how to combine physical models – attempting to model a very complex phenomenon – with big data analytics and statistics. The interesting research is whether this concept of situation-dependent, machine-learnt, multi-model blending can be further developed to a general framework to fuse domain knowledge and first-principle models with purely data-driven methods and whether this might be applicable for other applications where a complex physical phenomenon needs to be modeled.

7. References

1 Denholm, P. *et al.* The potential role of concentrating solar power in enabling high renewables scenarios in the United States. *Contract* **303**, 275-3000 (2012).
2 Margolis, R., Coggeshall, C. & Zuboy, J. SunShot vision study. *US Dept. of Energy* **2** (2012).
3 Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47-55 (2015).
4 Zhang, J. *et al.* A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy* **111**, 157-175 (2015).
5 Zhang, J. *et al.* in *3rd International workshop on integration of solar power into power systems, London, England.*
6 Zhang, J. *et al.* Baseline and Target Values for PV Forecasts: Toward Improved Solar Power Forecasting: Preprint. (2015).
7 Zhang, J. *et al.* Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting. *Solar Energy* **122**, 804-819 (2015).
8 Lu, S. *et al.* in *Proc. Euro. Control Conf.*
9 Hamann, H. F. & Lu, S. Solar & Alternative Energy Situation-dependent blending of multiple forecasting models based on machine learning.
10 Zhang, J., Chowdhury, S., Zhang, J., Messac, A. & Castillo, L. Adaptive hybrid surrogate modeling for complex systems. *AIAA journal* **51**, 643-656 (2013).
11 Juban, J., Siebert, N. & Kariniotakis, G. N. in *Power Tech, 2007 IEEE Lausanne.* 683-688 (IEEE).
12 Espinar, B. *et al.* Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations. *Solar Energy* **83**, 118-125 (2009).
13 Bessa, R., Miranda, V., Botterud, A. & Wang, J. 'Good'or 'bad'wind power forecasts: A relative concept. *Wind Energy* **14**, 625-636 (2011).
14 Hodge, B.-M. & Milligan, M. in *Power and Energy Society General Meeting, 2011 IEEE.* 1-8 (IEEE).
15 Mills, A. & Wiser, R. Implications of wide-area geographic diversity for short-term variability of solar power. Lawrence Berkeley National Laboratory, Environmental Energy Technologies Division. Sept. *escholarship. org/uc/item/9mz3w055* (2010).
16 Cui, M. *et al.* in *Power & Energy Society General Meeting, 2015 IEEE.* 1-5 (IEEE).
17 Florita, A., Hodge, B.-M. & Orwig, K. in *Green Technologies Conference, 2013 IEEE.* 147-152 (IEEE).
18 Lew, D., Brinkman, G., Ibanez, E., Hodge, B. & King, J. The western wind and solar integration study phase 2. *Contract* **303**, 275-3000 (2013).
19 Potter, C. W. *et al.* Creating the dataset for the western wind and solar integration study (USA). *Wind Engineering* **32**, 325-338 (2008).
20 Diagne, M., David, M., Lauret, P., Boland, J. & Schmutz, N. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews* **27**, 65-76 (2013).
21 Pelland, S., Remund, J., Kleissl, J., Oozeki, T. & De Brabandere, K. Photovoltaic and solar forecasting: state of the art. *IEA PVPS, Task* **14**, 1-36 (2013).
22 Inman, R. H., Pedro, H. T. & Coimbra, C. F. Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science* **39**, 535-576 (2013).
23 Perez, R. *et al.* Validation of short and medium term operational solar radiation forecasts in the US. *Solar Energy* **84**, 2161-2172 (2010).
24 Perez, R. *et al.* Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Solar Energy* **94**, 305-326 (2013).
25 Lorenz, E. *et al.* in *24th European photovoltaic solar energy conference.* 4199-4208 (Hamburg, Germany).
26 Mathiesen, P. & Kleissl, J. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. *Solar Energy* **85**, 967-977 (2011).
27 Zhang, J. *et al.* in *Power & Energy Society General Meeting, 2015 IEEE.* 1-5 (IEEE).

28 Mesinger, F. *et al.* North American regional reanalysis. *Bulletin of the American Meteorological Society* **87**, 343-360 (2006).
29 Key, J. R. & Schweiger, A. J. Tools for atmospheric radiative transfer: Streamer and FluxNet. *Computers & Geosciences* **24**, 443-451 (1998).
30 Stein, J. S. in *Photovoltaic Specialists Conference (PVSC), 2012 38th IEEE.* 003048-003052 (IEEE).
31 Hodge, B.-M., Florita, A., Sharp, J., Margulis, M. & Mcreavy, D. The value of improved short-term wind power forecasting. *National Renewable Energy Laboratory (NREL), Golden, CO* (2015).
32 Martinez-Anido, C. B. *et al.* The value of day-ahead solar power forecasting improvement. *Solar Energy* **129**, 192-203 (2016).
33 Ela, E., Milligan, M. & Kirby, B. Operating reserves and variable generation. *Contract* **303**, 275-3000 (2011).
34 Ela, E., Milligan, M. & O'Malley, M. in *Power and Energy Society General Meeting, 2011 IEEE.* 1-8 (IEEE).
35 Ela, E. & O'Malley, M. Studying the variability and uncertainty impacts of variable generation at multiple timescales. *IEEE Transactions on Power Systems* **27**, 1324-1333 (2012).
36 Mathiesen, P., Collier, C. & Kleissl, J. A high-resolution, cloud-assimilating numerical weather prediction model for solar irradiance forecasting. *Solar Energy* **92**, 47-61 (2013).
37 Marquez, R. & Coimbra, C. F. Intra-hour DNI forecasting based on cloud tracking image analysis. *Solar Energy* **91**, 327-336 (2013).
38 Chu, Y. *et al.* A smart image-based cloud detection system for intrahour solar irradiance forecasts. *Journal of Atmospheric and Oceanic Technology* **31**, 1995-2007 (2014).
39 Hammer, A., Heinemann, D., Lorenz, E. & Lückehe, B. Short-term forecasting of solar radiation: a statistical approach using satellite data. *Solar Energy* **67**, 139-150 (1999).
40 Perez, R. *et al.* A new operational model for satellite-derived irradiances: description and validation. *Solar Energy* **73**, 307-317 (2002).
41 Glahn, H. R. & Lowry, D. A. The use of model output statistics (MOS) in objective weather forecasting. *Journal of applied meteorology* **11**, 1203-1211 (1972).
42 Vislocky, R. L. & Fritsch, J. M. Improved model output statistics forecasts through model consensus. *Bulletin of the American Meteorological Society* **76**, 1157-1164 (1995).
43 Krishnamurti, T. *et al.* Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**, 1548-1550 (1999).
44 DelSole, T., Yang, X. & Tippett, M. K. Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological Society* **139**, 176-183 (2013).
45 Raftery, A. E., Gneiting, T., Balabdaoui, F. & Polakowski, M. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**, 1155-1174 (2005).
46 Haupt, S. E. & Kosovic, B. in *Computational Intelligence, 2015 IEEE Symposium Series on.* 496-501 (IEEE).
47 Mahoney, W., Wiener, G., Liu, Y., Myers, W. & Johnson, D. in *AGU Fall Meeting Abstracts.* 03.
48 Chen, Y. *et al.* in *Proceedings of the Industrial Track of the 13th ACM/IFIP/USENIX International Middleware Conference.* 1 (ACM).
49 Klein, L. J. *et al.* in *Big Data (Big Data), 2015 IEEE International Conference on.* 1290-1298 (IEEE).
50 Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* **27**, 83-85 (2005).
51 Tesauro, G., Gondek, D., Lenchner, J., Fan, J. & Prager, J. M. Simulation, learning, and optimization techniques in Watson's game strategies. *IBM Journal of Research and Development* **56**, 16: 11-16: 11 (2012).
52 Lu, S. *et al.* in *Control Conference (ECC), 2015 European.* 283-290 (IEEE).
53 Sharma, N., Sharma, P., Irwin, D. & Shenoy, P. in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on.* 528-533 (IEEE).
54 Aler, R., Martín, R., Valls, J. M. & Galván, I. M. in *Intelligent Distributed Computing VIII* 269-278 (Springer, 2015).

55 Li, Z., Rahman, S., Vega, R. & Dong, B. A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting. *Energies* **9**, 55 (2016).
56 Augustine, J. A., DeLuisi, J. J. & Long, C. N. SURFRAD--A national surface radiation budget network for atmospheric research. *Bulletin of the American Meteorological Society* **81**, 2341 (2000).
57 Rogers, E. *et al.* in *Preprints, 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction.*
58 Han, J. & Pan, H.-L. Revision of convection and vertical diffusion schemes in the NCEP global forecast system. *Weather and Forecasting* **26**, 520-533 (2011).
59 Grimit, E. P. & Mass, C. F. Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Weather and Forecasting* **17**, 192-205 (2002).
60 Meinshausen, N. Quantile regression forests. *The Journal of Machine Learning Research* **7**, 983-999 (2006).
61 Hooker, G. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics* (2012).
62 Rasmussen, C. E. in *NIPS.* 554-560.
63 Zivkovic, Z. in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on.* 28-31 (IEEE).
64 Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).
65 Dietterich, T. G. in *Multiple classifier systems* 1-15 (Springer, 2000).
66 Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**, 27 (2011).
67 Bennett, K. P. & Campbell, C. Support vector machines: hype or hallelujah? *ACM SIGKDD Explorations Newsletter* **2**, 1-13 (2000).
68 Hwang, Y., Lu, S. & Kim, J.-K. Bottom-up estimation and top-down prediction: Solar energy prediction combineing information from multiple sources. *Annals of statistics* **Submtted** (2016).
69 Bechtold, P. *et al.* Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society* **134**, 1337-1351 (2008).
70 Zachariassen, J., Zeller, K. F., Nikolov, N. & McClelland, T. A review of the forest service remote automated weather station (RAWS) network. (2003).
71 Deutschmann, T. *et al.* The Monte Carlo atmospheric radiative transfer model McArtim: Introduction and validation of Jacobians and 3D features. *Journal of Quantitative Spectroscopy and Radiative Transfer* **112**, 1119-1137 (2011).
72 Chow, C. W. *et al.* Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed. *Solar Energy* **85**, 2881-2893 (2011).
73 Lu, S. *et al.* in *Big Data (Big Data), 2016 IEEE International Conference on.* 2672-2675 (IEEE).

8. Glossary

15MA	15 minute ahead
1HA	1 hour ahead
4HA	4 hours ahead
AACEE	Absolute Area Control Error in Energy
ACE	Area Control Error
ARM	Atmospheric Radiation Measurement
BND	Bondville Surfrad Station

BP	Budget Period
CAISO	California Independent System Operator
CDF	Cumulative Distribution Function
CPS2	North American Electric Reliability Corporation Control Performance Standard 2
DA	day ahead
DiCast	Dynamic Integrated foreCast
DNI	Direct Normal Irradiance
DRA	Desert Rock Surfrad Station
ECMWF	European Centre for Medium-Range Weather Forecasts
ESR	Expected Synthetic Reliability
ESRL	Earth System Research Laboratory
FANOVA	Functional Analysis of Variance
FESTIV	Flexible Energy Scheduling Tool for Integration of Variable Generation
FPK	Fort Peck Surfrad Station
GCM	Goodwin Creek Surfrad Station
GFS	Global Forecasting System
GHI	Global Horizontal Irradiance
GMP	Green Mountain Power
Grib2	Gridded binary or general regularly-distributed information in binary form
HDF	Hierarchical Data Format
HRRR	High Resolution Rapid Refresh
ISIS	Integrated Surface Irradiance Study
ISO	Independent System Operator
ISO-NE	Independent System Operator of New England
KDE	Kernel Density Estimation
KSI	Kolmogorov-Smirnoff Integral
LM	Linear Model
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MaxAE	Maximum Absolute Error
MBE	Mean Bias Error
MOS	Model Output Statistics

NAM	North American Mesoscale
netCDF	Network Common Data Form
NOAA	National Oceanic and Atmospheric Administration
NREL	National Renewable Energy Laboratory
NWP	Numerical Weather Prediction
OVER	part of the KSI which integrates above (over) the Kolmogorov-Smirnov critical value)
PAIRS	Physical Analytics Integrated Data Repository and Services
POA	Plane of Array
PSU	Penn State University Surfrad Station
PV	Photovoltaic
PV	Photovoltaic
PVLib	PVLIB is a set of open source modeling functions that simulate PV system performance
RF	Random Forest
RMSE	Root Mean Square Error
RTM	Radiative Transfer Model
SOPO	Statement Of Project Objectives
SREF	Short Range Ensemble Forecast
SurfRad	Surface Radiation network
SVM	Support Vector Machine
SXF	Sioux Falls Surfrad Station
T2M	surface temperature at 2m above ground
TBL	Table Mountain Surfrad Station
TEP	Tuscon Electric Power
UVIG	Utility Variable-Generation Integration Group
W10M	wind speed at 10m above ground
Watt-sun	IBM's renewable and weather forecasting technology
WRF	Weather Research Forecast
WWSIS-2	Western Wind and Solar Integration Study Phase 2
XML	eXtensible Markup Language