Air Force Institute of Technology

AFIT Scholar

5-2020

Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data

Christil Pasion
Air Force Institute of Technology

Torrey J. Wagner
Air Force Institute of Technology

Clay Koschnick
Air Force Institute of Technology

Steven J. Schuldt
Air Force Institute of Technology

Jada B. Williams
Air Force Institute of Technology

See next page for additional authors

Recommended Citation

Authors

Christil Pasion, Torrey J. Wagner, Clay Koschnick, Steven J. Schuldt, Jada B. Williams, and Kevin Hallinan

Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data

Christil Pasion [1], Torrey Wagner [1,*] , Clay Koschnick [1], Steven Schuldt [1] , Jada Williams [1] and Kevin Hallinan [2]

[1] Graduate School of Engineering and Management, Air Force Institute of Technology, Wright-Patterson AFB, OH 45433, USA; christil.pasion.1@us.af.mil (C.P.); clay.koschnick@afit.edu (C.K.); steven.schuldt@afit.edu (S.S.); jada.williams.ctr@afit.edu (J.W.)

[2] Department of Mechanical and Aerospace Engineering, University of Dayton, Dayton, OH 45469, USA; kevin.hallinan@udayton.edu

* Correspondence: torrey.wagner@afit.edu

Abstract: Solar energy is a key renewable energy source; however, its intermittent nature and potential for use in distributed systems make power prediction an important aspect of grid integration. This research analyzed a variety of machine learning techniques to predict power output for horizontal solar panels using 14 months of data collected from 12 northern-hemisphere locations. We performed our data collection and analysis in the absence of irradiation data—an approach not commonly found in prior literature. Using latitude, month, hour, ambient temperature, pressure, humidity, wind speed, and cloud ceiling as independent variables, a distributed random forest regression algorithm modeled the combined dataset with an R^2 value of 0.94. As a comparative measure, other machine learning algorithms resulted in R^2 values of 0.50–0.94. Additionally, the data from each location was modeled separately with R^2 values ranging from 0.91 to 0.97, indicating a range of consistency across all sites. Using an input variable permutation approach with the random forest algorithm, we found that the three most important variables for power prediction were ambient temperature, humidity, and cloud ceiling. The analysis showed that machine learning potentially allowed for accurate power prediction while avoiding the challenges associated with modeled irradiation data.

Keywords: photovoltaics; solar panels; power prediction; machine learning; random forest

1. Introduction

Power generation from solar photovoltaics (PV) is expected to grow 30% in the next five years, and much of this growth is anticipated to be in the form of distributed solar PV systems [1]. Distributed PV can be advantageous to residential customers and commercial/government facilities—both in urban settings as well as more disperse settings (e.g., remote military installations)—where there may be limitations on building large, centralized PV arrays. The challenge of intermittency for solar energy is well-established and highlights the critical function of forecasting solar PV power output—especially in a distributed environment.

Solar PV power forecasting has been studied extensively. Lorenz et al. (2014) provided an overview [2], and Raza et al. (2016) discussed recent advances [3]. Often, solar power forecasting studies are based on predicting irradiance or using historical power output. Yang et al. (2015) used exponential smoothing to improve predictions of horizontal irradiance [4]; Gueymard (2008) studied irradiance forecasting for surfaces of any angle [5]. Lorenz et al. (2010) used regional weather data to forecast irradiance, which was then converted to power [6]. Various studies have considered predicting irradiance or power using weather and prior power output data [7,8]. Additionally, previous

studies forecasting solar irradiance or power output are often based on data from a limited number of locations [8–11].

The novel aspect of this research is the quantification of the ability of machine learning to predict photovoltaic power output in the absence of irradiation data, using collected data from a range of climate zones. This is motivated by challenges with available irradiance data; it is conceptually a reliable predictor of solar power output, and irradiance is found to be the most important factor in predicting solar panel power output in two photovoltaic studies that utilize modeling [12,13]. However, irradiation data can be time-consuming to measure at a specific site, and prediction of irradiation can generate forecast errors, may be unsuitable for accurate PV performance analysis, and may contain 8–25% uncertainty if modeled [7,9,14,15]. Additionally, this work studied forecasting power output for horizontal PV arrays for the following reasons:

1. Many entities do not have space available to install large solar arrays; thus, horizontal, distributed arrays, such as building rooftops, can broaden the opportunities to implement solar energy.
2. Many models have been developed for latitude-tilted applications [16]. While latitude-tilted solar panels possess the ability to capture more direct solar irradiation, horizontal solar panels have been found to perform better under diffuse irradiation conditions [17–20].

Accordingly, we performed our data collection and analysis of horizontal photovoltaics in the absence of irradiation data. This tested the hypothesis that accurate power prediction can result from the combination of advances in machine learning and avoided irradiation uncertainty. The objective of the work was to quantify the ability of this approach.

The approach used in this work was based on the following selection of input variables and the type of photovoltaic panel. There are several factors identified in prior research that affect both the irradiation that reaches the panel and the panel's ability to convert the irradiation to usable energy:

- Cloud Ceiling: the presence of clouds above a panel will scatter solar irradiance and decrease the amount of irradiation a panel receives; the cloud ceiling is measured at the altitude where at least 5/8ths of the sky above the weather station is covered by clouds [17–25];
- Latitude: the latitude of each location will dictate the sun deflection angle; this will affect the amount of sunlight the panel receives [12,21–23,25,26];
- Month: when the sun rises and sets and how high it will appear in the sky at any location on the earth is determined (in part) by the time of year at that location [13,21];
- Hour: the time of day determines how high the sun is in the sky—or whether or not it is present at all. Hour controls for the sun's position in relation to the time of day [21];
- Humidity: water affects incoming sunlight through refraction, diffraction, and reflection. Indirectly, humidity also affects dust build-up on panels due to the formation of dew increasing coagulation of dust [27]; conversely, dew formation on the surface of a panel may increase performance when compared to a humid air condition [28];
- Temperature: the efficiency of a solar panel will generally decrease with an increase in panel temperature [29,30]. Including temperature as an explanatory variable for power output has led to increased predictability [12,13,31–33];
- Wind speed: the temperature of the panel may be affected by the speed of the wind surrounding the panel [34,35]. Increased wind speed can also clean the dust off of the panel surface or stir up dust, thereby affecting the irradiance that reaches the panel [36];
- Visibility: this variable is a measurement of the distance at which a light can be seen and identified [37]. Visibility will primarily affect how much irradiation reaches the panel and can have a negative effect on power output if visibility is low during daylight hours;
- Pressure: Pressure may have an effect on power output predictability by indicating a weather occurrence—such as a storm [38]; this variable has not been extensively explored in solar panel power output literature;

- Altitude: there is less atmosphere for the sun to travel through at locations with higher altitudes; this results in a higher level of irradiation at locations farther above sea level.

Monocrystalline and polycrystalline silicon PV panels comprise nearly 90% of the world's photovoltaics and achieve efficiencies of 15–25% and 13–16%, respectively [39]. Polycrystalline panels were selected for this analysis as they are more widely installed than monocrystalline photovoltaics and have a lower cost, making them well-suited for distributed PV settings.

Prior researchers have predicted photovoltaic power output or efficiency utilizing multiple input factors, such as irradiation, temperature, humidity, solar elevation angle, wind speed, wind direction, month, and others [37–40]. Table 1 summarizes the key characteristics from four photovoltaic studies; the present work was also included for comparison. The table highlights that numerous variables have been studied for use in photovoltaic modeling over various timeframes, depending on the research objectives. In Table 1, short is defined as having an effect within a day, *medium* is on the order of months, and *long* is an effect that takes a year or more to impact the power output.

Busquet et al. (2018) primarily studied the medium- and long-term effects of factors, such as aging and soiling; panel age is not commonly used by other studies [35]. Aging describes the amount of time the panel has been installed and exposed to the elements, and soiling describes the dust build-up of the panel's surface. Kayri et al. (2017) and Lahouar et al. (2017) forecasted solar panel power output and used short-term factors, such as solar elevation angle and wind direction. However, they did not include longer-term factors, such as aging [12,13]. Mekhilef et al. (2012) conducted a medium-timeframe review primarily interested in the effects of dust, humidity, and air velocity, including the contribution of water droplets trapped inside the cell and dew-induced dust accumulation [27].

Solar irradiance is one common factor that the four studies used. The present work differentiated itself from prior work by predicting horizontal solar panel power output only using readily available data—such as position, time, and weather, while not including irradiation. If the power output of a solar panel can be reasonably predicted without including irradiation as an input, then it becomes easier to assess the cost-effectiveness of a PV array at any global location.

Table 1. Comparison of independent variables.

Characteristics		Present Work	Busquet et al. [35]	Kayri et al. [12]	Lahouar et al. [13]	Mekhilef et al. [27]
Model type		Multiple machine learning algorithms	Linear regression	Linear regression Random forest Artificial neural network	Random forest Forecasting	Case study
Type of panel		Polycrystalline	Many	Unknown	Unknown	Many
Orientation		Horizontal	20 degree tilt	Unknown	Unknown	Many
Locations		12 in the United States	Hawaii	Turkey	Australia	6 in Asia
Output		Power	Daily energy	Power	Power	Efficiency
Factors	**Timeframe**					
Hour of day	Short	x			x	
Month	Medium	x		x	x	x
Ambient temperature	Short	x	x	x	x	
Wind speed/air velocity	Short	x	x	x	x	x
Visibility	Short	x				
Atmospheric pressure	Short	x				
Cloud ceiling	Short	x				
Altitude	Long	x				
Latitude	Long	x				
Soiling (dust)	Medium		x		x	x
Aging	Long		x			
Solar elevation angle	Short			x		
Solar irradiation	Short		x	x	x	x

2. Materials and Methods

This section presents the procedures and processes used in this study. A description of the test equipment used to gather the data, how the data was processed for predictive modeling, model development, and validation methods are provided.

2.1. Materials and Equipment

The test systems used in the study were designed and manufactured as part of a previous research effort and were distributed to global United States Air Force (USAF) installations [40]. The test systems were comprised of the following equipment:

- Renogy 50-watt, 12-volt, polycrystalline PV panels;
- Raspberry Pi 3, model B, version 1.2 computer systems;
- Waterproof Pelican cases;
- CAT cables, power cables, and SD cards.

The Raspberry Pi computer system was used to record the following information at 15-min time intervals: panel power output, temperature, humidity, date, and time. The SD card in the computer was retrieved by the site monitors and downloaded every month, and the dataset was sent to the researchers. Site monitors at each location were given instruction to clean off the panel whenever dust or snow cover was observed. Although this was performed daily for some locations, others were cleaned less frequently. The unknown frequency of panel cleaning at some locations was a known limitation of this research.

2.2. Data Description

Data collected from 12 locations were utilized within this study—the data is available for further analysis [41]. The collection locations were selected from a larger dataset of all Department of Defense (DoD) installations located within 25 regions [40]. Using this dataset, along with a recognized climate classification matrix, a Pareto analysis was performed to determine the locations of test sites within climate regions [40]. While reviewing the collected data, the project team discovered that only a subset of locations collected reliable data. After post-processing the data, the team chose to limit collection data to the northern hemisphere. This decision was motivated by seasonal differences between hemispheres and selecting collection sites in close proximity to National Oceanic and Atmospheric Administration (NOAA) weather stations.

The test systems at each location provided the ambient temperature, relative humidity, timestamp, and power output for each panel. Altitude, latitude, and four weather variables from the NOAA were also added to the dataset. The weather stations that recorded the NOAA wind speed, cloud ceiling, visibility, and atmospheric pressure data were located at airports no more than five miles from each test system [42]. The cloud ceiling data measured the lowest cloud layer with 5/8ths or greater opacity, and a value of 22 km indicated a lack of cloud cover.

A graphical depiction of the 12 locations is provided in Figure 1; there were two sites in Colorado that appeared as a single red dot due to their proximity. Additionally, Table 2 provides the latitude, longitude, and Köppen–Geiger climate region of each location. Note, all latitudes were north, and all longitudes were west. Note, seven different climate regions were represented in the dataset, indicating a diverse range of locations.
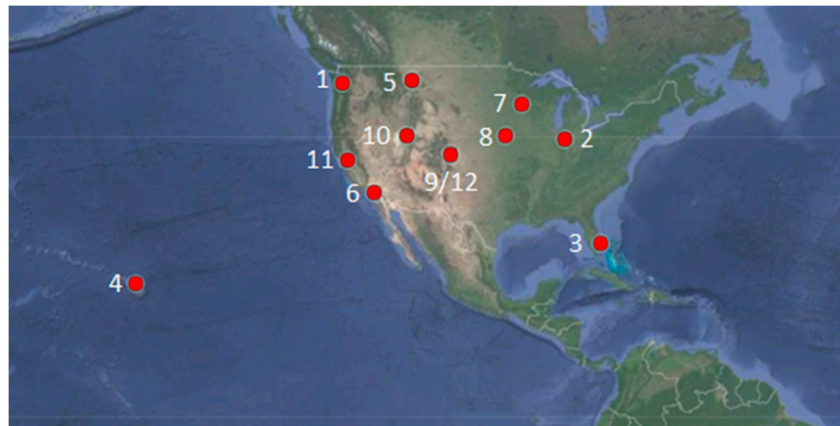
Figure 1. Geographic locations of data collection sites.

Table 2. Name and coordinates of data collection sites.

Site	State	Latitude (deg)	Longitude (deg)	Köppen–Geiger Climate Region [40]
1. Camp Murray	Washington	47.11	122.57	Csb
2. Grissom	Indiana	40.67	86.15	Dfa
3. JDMT	Florida	26.98	80.11	Cfb
4. Kahului	Hawaii	20.89	156.44	Af
5. Malmstrom	Montana	47.52	111.18	BSk
6. March	California	33.9	117.26	Csa
7. MNANG	Minnesota	44.89	93.2	Dfa
8. Offutt	Nebraska	41.13	95.75	Dfa
9. Peterson	Colorado	38.82	104.71	BSk
10. Hill Weber	Utah	41.15	111.99	Dfb
11. Travis	California	38.16	121.56	Csa
12. USAFA	Colorado	38.95	104.83	BSk

JDMT: Jonathan Dickinson Missile Tracking Annex; MNANG: Minnesota Air National Guard; USAFA: U.S. Air Force Academy.

Descriptive statistics for each numeric variable are shown in Table 3; hour and month were not listed as they were described as categorical variables in the model.

Table 3. Descriptive statistics for numeric variables.

Variable	Units	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Power output	Watts	0.3	6.4	13.8	13.0	18.9	34.3
Latitude	Degrees	20.89	38.16	38.95	38.12	41.15	47.52
Humidity	Percent	0	17.5	33.1	37.1	52.6	100
Ambient temp	Celsius	−20.0	21.9	30.3	29.3	37.5	65.7
Wind speed	km/h	0	9.7	14.5	16.6	22.5	78.9
Visibility	km	0	16.1	16.1	15.6	16.1	16.1
Pressure	Millibars	781	845	961	925	1008	1029
Cloud ceiling	km	0	4.3	22	15.7	22	22
Altitude	m	0.3	0.6	140	244	417	593

2.3. Data Pre-Processing

The initial dataset was filtered to only include the time window of 10:00–15:45 to avoid modeling periods of darkness and reduced sunlight. This restriction also helped mitigate possible obstructions from both natural and man-made objects when the sun was low in the sky. Next, the pairwise correlation coefficients for all numeric variables across all sites were calculated—the results are presented in

Figure 2. Only one pair of variables showed a high correlation coefficient: altitude and pressure. Altitude was subsequently removed since its value did not change with location, whereas pressure did have some degree variation for a location—i.e., power output.
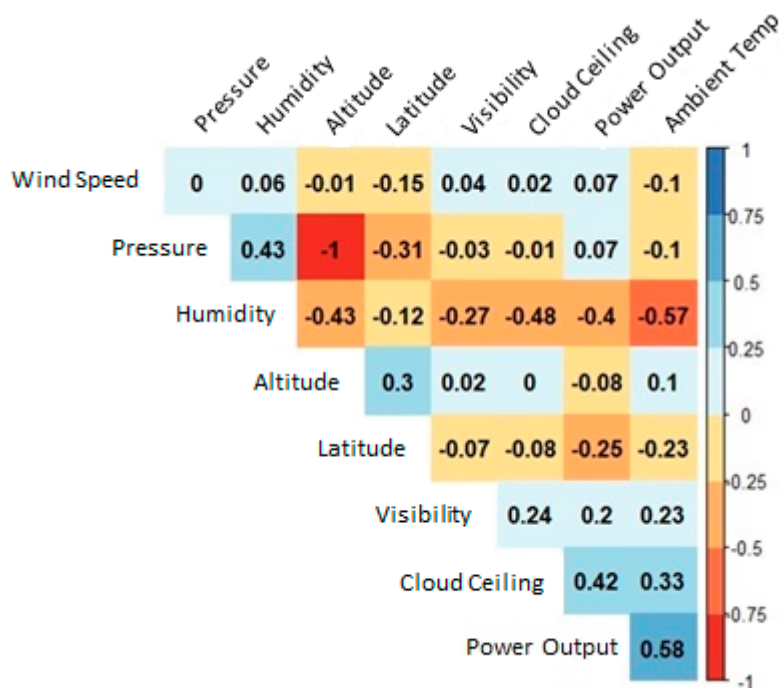

Figure 2. Correlation coefficients for numeric variables.

2.4. Machine Learning Modeling

H2O.ai is an open-source machine learning tool used in this study to compare various modeling algorithms to determine the best fit for power output; H2O.ai includes a tool called AutoML that automates the machine learning model building process through a graphical user interface [43,44]. For this research, algorithm accuracy was assessed using the entire dataset and using a cross-validation process, which divided the dataset into k bins, and then during each iteration of the model building process for a given algorithm, one bin was the validation set, and the other k-1 bins were the training set. Thus, k cross-validation models were built for each algorithm. For reproducibility, the number of folds was set to $k = 5$, the maximum runtime was limited to 8000 s, and other H2O.io input parameters were set by the software to the default values.

Six algorithms were compared in this research. The first five are the popular "base learner" algorithms [43]. The sixth algorithm (stacked ensemble build) is referred to as a "metalearner"; it creates an additional model, which is a combination of models from the other five algorithms. Descriptions of the six machine learning algorithms are provided below [45]:

- Deep learning is designed using the "multi-layer feedforward artificial neural network that is trained with stochastic gradient descent using back-propagation." This method provides understanding into network behavior based on altering the weights and biases;
- Gradient boosting machine (GBM) builds a model where regression trees are built in parallel. The generated leaf nodes are inputs into other models, such as the generalized linear model;
- The stacked ensemble build represents all of the models that are combined or stacked together using cross-validation folds;
- Generalized linear modeling (GLM) generates various distributions, including Gaussian, Poisson, Binomial, Multinomial, Gamma, Ordinal, and Negative Binomial regression, and estimates the regression. This algorithm can generate both classification and regression models;

- Distributed random forest (DRF) randomly selects a subset of the features and generates a single forest of regression or classification trees based on those features; this process is repeated—based on the number of trees specified—with a random subset on each iteration. The predictions are based on the average prediction of all of the trees in the forest;
- Distributed random forest extremely randomized trees (XRT) select thresholds differently when compared to the distributed random forest model. Thresholds from a random subset of features are chosen at random and ranked by the best threshold.

2.5. Impact of Input Variables

In the absence of irradiance data, understanding the importance of the input variables used to predict horizontal solar panel power output was important. Variable importance was determined by measuring how much each variable decreased the model mean squared error (MSE), defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{1}$$

In Equation (1), n is the number of validation data points, y_i is the actual response, and $\hat{y}_i$ is the predicted response. The MSE was calculated again after permuting each predictor variable and then subtracting the MSE of the validation dataset. The average change in MSE for each predictor variable permutation was then determined. This value was then scaled by dividing the MSE reduction by the variable's standard error.

2.6. Methodology Summary

Figure 3 below provides a flowchart of the analysis used for this study. While steps 2 and 3 were specified for the DRF algorithm, the general flow would apply to each algorithm assessed.
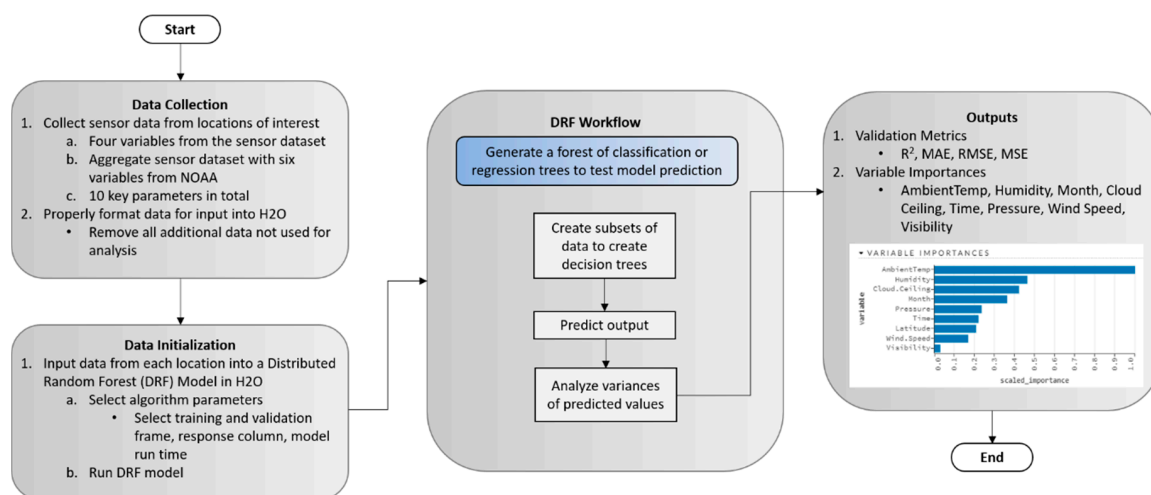

Figure 3. Nominal flowchart for distribution random forest (DRF) algorithm [46,47].

3. Results

The R^2, mean absolute error (MAE), and root MSE (RMSE) training data results for each algorithm are presented in Table 4; the DRF algorithm was the most accurate in modeling power output for the full dataset.

Table 4. H2O.ai and cross-validation modeling results.

Machine Learning Technique	H2O.ai			Cross-Validation
	R^2	MAE (W)	RMSE (W)	R^2
DRF—Distributed random forest	0.939	1.176	1.754	0.673
XRT—Extremely randomized trees	0.924	1.341	1.965	0.664
Stacked ensemble build	0.868	1.748	2.585	0.687
GBM—Gradient boosting machine	0.802	2.134	3.173	0.681
Deep learning	0.593	3.386	4.545	0.605
GLM—Generalized linear model	0.502	3.896	5.027	0.501

The primary methods of assessing the accuracy of the results were the R^2, MAE, and RMSE values presented in Table 4. During the training process, additional insight into model performance could be gained from the results in the cross-validation process, which are presented as the right column of Table 4 for the six algorithms. Cross-validation allowed for an efficient way to test the predictive capability of an algorithm on data not included in training the model. Based on the cross-validation results (using five folds), the stacked ensemble build and gradient boosting machine methods performed slightly better than the DRF method—a 2.1% and 1.2% increase in R^2, respectively. Based on the results in Table 4 and the commonality of the distributed random forest algorithm (DRF) with our comparison studies, we conducted further analysis on the DRF model.

Random forest regression is an ensemble method that aggregates a series of individual regression trees in order to reduce model variance. The random forest model consisted of a number of decision trees and a separate number of decision variables for each tree. Using the method described in Section 2.5, the variable importance rankings—across all locations–are presented in Table 5. In the modeling process, multiple values for the number of decision trees were explored; the default value was 50 trees. For comparison, the rankings for 500 trees were also presented—and the rank order did not change. The three most important variables were ambient temperature, humidity, and cloud ceiling.

Table 5. Variable importance rankings using DRF.

Variable	Scaled Performance for 50 Trees	Scaled Performance for 500 Trees
Ambient temp	100%	100%
Humidity	55%	46%
Cloud ceiling	52%	42%
Month	50%	36%
Pressure	26%	24%
Time	25%	22%
Latitude	25%	21%
Wind speed	19%	17%
Visibility	4%	3%

The data from each location was then modeled separately using the DRF algorithm. Table 6 presents the results and shows there is location-dependent variation between ambient temperature, cloud ceiling, and humidity as the main drivers of model performance. This was expected as the locations of the test sites vary across eight climate regions where solar energy potential is affected by accompanying variations of temperature, humidity, and cloud cover [48]. Ambient temperatures and humidity were the top two primary influencers of solar power output in nine of the 12 locations. The results in Table 6 were relatively consistent across locations, which indicated that the independent variables provided sufficient information to be applied across a range of geographical locations.

Table 6. Accuracy metrics and variable importance rankings by location.

Location	R^2	MAE (W)	RMSE (W)	First Variable	Value	Second Variable	Value
Camp Murray	0.962	0.876	1.339	Ambient Temp	37%	Humidity	26%
Grissom	0.948	0.957	1.534	Ambient Temp	34%	Humidity	23%
JDMT	0.929	1.461	1.999	Humidity	27%	Ambient Temp	24%
Travis	0.968	0.779	1.193	Ambient Temp	29%	Humidity	21%
Hill Weber	0.955	0.988	1.445	Ambient Temp	27%	Humidity	24%
Kahului	0.908	1.699	2.187	Humidity	25%	Ambient Temp	23%
Malmstrom	0.951	1.023	1.564	Ambient Temp	32%	Humidity	23%
Offutt	0.937	1.456	2.038	Humidity	33%	Ambient Temp	22%
USAFA	0.924	1.160	1.609	Ambient Temp	21%	Cloud Ceiling	16%
MNANG	0.955	1.069	1.643	Ambient Temp	34%	Cloud Ceiling	17%
Peterson	0.947	1.050	1.561	Ambient Temp	30%	Humidity	17%
March	0.936	0.919	1.296	Month	23%	Ambient Temp	23%
All Locations	0.939	1.187	1.754	Ambient Temp	32%	Humidity	15%

An important variable for predicting power would seemingly be latitude; however, it was ranked seventh in both the 50-tree and 500-tree models. The relative unimportance of latitude might be due to the limited latitude range included in the model. Latitudes in the northern hemisphere range from 0–66 degrees; however, the latitude range for the dataset was only 21–48 degrees. As shown in Table 5, the DRF algorithm best predicted the Travis data, whose location is 38.16 degrees latitude and 121.56 degrees longitude within the hot-dry climate region. Camp Murray in Washington had the second-best model performance; this site is located at 47.11 degrees latitude and 122.57 degrees longitude in the mixed-humid climate region. Between these two sites, the higher percentage of humidity and ambient temperature influence in Camp Murray was likely due to larger seasonal variations in these variables. In contrast, the model performance for the Kahului, Hawaii site, was the poorest. The seasonal weather variation there was substantially different from the remainder of the sites. A final observation from Table 6 was the difference in model performance between USAFA and Peterson. While the sites are only 20 miles apart, they have significantly different geographical characteristics as USAFA is nestled on the Rocky Mountain foothills, and Peterson is located on the plains. In such a scenario, predicting output in the absence of irradiation data may be beneficial as irradiation may not vary significantly between locations.

4. Discussion

To the best knowledge of the authors, this work provided the first study to predict the power output of geographically distributed horizontal polycrystalline solar panels in the absence of irradiation or previous power output data. Although it can be challenging to make exact comparisons with previous research due to the range of potential differences, it is still insightful to see how these results compare to other studies. There has been modeling done for a range of algorithms and datasets to forecast solar PV energy output using solar irradiance. Ahmad et al. (2018) predicted hourly energy output and reported training set R^2 values of 0.9105, 0.9272, and 0.9367 for support vector machine, extremely randomized trees, and random forest models, respectively [49]. Ramsami and Oree (2015) used single-stage and stepwise regression and neural network models with correlation coefficients ranging from 0.914 to 0.937 [50]. Additionally, Pedro and Coimbra (2012) used previous power output data in time series, neural network, and nearest neighbor models to forecast one-hour ahead energy output with R^2 values ranging from 0.91 to 0.96 for the full validation set [51].

We also presented our results in the context of the three quantitative studies summarized in Table 1. Table 7 displays the present results next to the most applicable subset of results from the three other studies. It is important to emphasize that the purpose of these comparisons was to understand the context of forecasting solar power output in the absence of irradiation data. The results presented were chosen to make the comparisons as close as possible, i.e., most similar algorithms and type of solar panel, but there were still differences in the tuning parameters, the definition of power for the dependent variable, the available independent variables, and the time period of the data.

Table 7. Comparison with recent studies.

Measure	Present Work	Busquet et al. [35]	Kayri et al. [12]	Lahouar et al. [13]
Model	DRF	Linear regression	Random forest	Random forest
Dependent variable	Power	Daily energy	Power	Power
R^2	0.939	0.87 *	0.986	N/A
MAE (W)	1.176	N/A	2.376	30144 **
RMSE (W)	1.754	N/A	N/A	44343 **
Importance—1st variable	Ambient temp	High/low irradiation	Global radiation	Solar irradiance **
Importance—2nd variable	Humidity	Ambient temp	Solar elevation angle	Humidity **
Importance—3rd variable	Cloud ceiling	Wind speed	Temperature	Temperature **

* based on S3 solar panel; ** based on the January timeframe.

In general, the results of this study indicated that solar power prediction might be suitable in the absence of irradiation data as the quantitative performance measures were not out of a family with the other studies. One notable difference was this study included nine independent variables, whereas the three comparison studies in Table 7 used six. These additional parameters might have sufficiently compensated for the lack of irradiation data, which was consistently shown to be the most important variable in the other studies. Lahouar et al. (2017) conducted an additional analysis excluding irradiation data, with a resulting MAE = 44,271 W and RMSE = 59,391 W [13]; these measures were significantly higher than the DRF results in this study. These differences might be due to the larger power systems, short timeframe (a single week in January), the exclusion of other independent variables, or a smaller data set.

5. Conclusions

In summary, using only weather, time, and geographic variables, 14 months of data from 12 northern-hemisphere locations were modeled using a variety of machine learning techniques. These data contributed to an $R^2 = 0.94$ model accuracy using the distributed random forest algorithm on the full dataset within the H2O.ai platform. This work indicated that advances in machine learning could potentially facilitate accurate prediction of horizontal photovoltaic panels without irradiation data; this type of prediction was beneficial as irradiation data could be time-consuming to measure or contain significant uncertainty if modeled. Additionally, we identified the three most important weather variables for power prediction in the absence of irradiation data as ambient temperature, humidity, and cloud ceiling.

This type of analysis could be practically useful in supporting feasibility and cost-effectiveness decisions for the use of solar power for geographically distributed entities—especially ones of an agile or expeditionary nature. For example, consider the power requirements of a small, quick response team—such as those responding to a humanitarian crisis or natural disaster—that operate over a diverse range of locations. These teams often work in austere environments without a local, reliable power source. The ability to determine the feasibility and scale of a distributed PV power in support of these teams without requiring the time to gather or model irradiation data could be valuable. This benefit could also extend to the growth of distributed residential PV in rural areas.

In addition to the applicability of this forecasting, the scalability of this particular study had both advantages and disadvantages. On the advantage side: (1) the study was conducted over a relatively diverse set of locations (as noted in Section 3); (2) the data was collected in a controlled manner—e.g., there was specific installation and operation guidance provided to each site; (3) by each measure of accuracy, there were multiple machine algorithms that gave a similar performance, indicating some degree of robustness to the choice of algorithm. Conversely, the collection of the weather data was not pre-planned as part of gathering the solar panel data, nor were the comparison studies identified prior to executing the machine learning algorithms.

Therefore, future research could extend the benefit of the efficacy of this type of forecasting. An experiment could be conducted whereby a distributed solar PV system is sized based on a nominal requirement and the forecasted power output using this model; then, measure how well the system

met the power requirements. Additionally, the collection of weather data could be automated or linked directly with the location of the PV system (as opposed to the local weather station). Finally, further comparisons of these results with other models could be studied.

References

1.　International Energy Agency. Renewables 2019. Available online: https://www.iea.org/reports/renewables-2019 (accessed on 16 April 2020).

2.　Lorenz, E.; Kühnert, J.; Heinemann, D. Overview of Irradiance and Photovoltaic Power Prediction. In *Weather Matters for Energy*; Troccoli, A., Dubus, L., Haupt, S., Eds.; Springer: New York, NY, USA, 2014; pp. 429–454.

3.　Raza, M.Q.; Nadarajah, M.; Ekanayake, C. On recent advances in PV output power forecast. *Sol. Energy* **2016**, *136*, 125–144. [CrossRef]

4.　Yang, D.; Sharma, V.; Ye, Z.; Lim, L.I.; Zhao, L.; Aryaputera, A.W. Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. *Energy* **2015**, *81*, 111–119. [CrossRef]

5.　Gueymard, C.A. Prediction and validation of cloudless shortwave solar spectra incident on horizontal, tilted, or tracking surfaces. *Sol. Energy* **2008**, *82*, 260–271. [CrossRef]

6.　Lorenz, E.; Scheidsteger, T.; Hurka, J.; Heinemann, D.; Kurz, C. Regional PV power prediction for improved grid integration. *Prog. Photovol.* **2010**, *19*, 757–771. [CrossRef]

7.　Qing, X.; Niu, Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy* **2018**, *148*, 461–468. [CrossRef]

8.　Chakraborty, P.; Marwah, M.; Arlitt, M.; Ramakrishnan, N. Fine-Grained Photovoltaic Output Prediction Using a Bayesian Ensemble. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence AAAI, Toronto, ON, Canada, 22–26 July 2012; pp. 274–280. [CrossRef]

9.　Yaniktepe, B.; Genc, Y.A. Establishing new model for predicting the global solar radiation on horizontal surface. *Int. J. Hydrogen Energy* **2015**, *40*, 15278–15283. [CrossRef]

10.　Su, Y.; Chan, L.; Shu, L.; Tsui, K. Real-time prediction models for output power and efficiency of grid-connected solar photovoltaic systems. *Appl. Energy* **2012**, *93*, 319–326. [CrossRef]

11.　Ma, T.; Yang, H.; Lu, L. Solar photovoltaic system modeling and performance prediction. *Renew. Sustain. Energy Rev.* **2014**, *36*, 304–315. [CrossRef]

12.　Kayri, M.; Kayri, I.; Gencoglu, M.T. The Performance Comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using Photovoltaic and Atmospheric Data. In Proceedings of the 2017 14th International Conference on Engineering of Modern Electric Systems (EMES 2017), Oradea, Romania, 1–2 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4. [CrossRef]

13.　Lahouar, A.; Mejri, A.; Slama, J.B.H. Importance based selection method for day-ahead photovoltaic power forecast using random forests. In Proceedings of the 2017 International Conference on Green Energy Conversion Systems (GECS), Hammamet, Tunisia, 23–25 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 554–560. [CrossRef]

14.　Wilcox, S. *National Solar Radiation Database 1991–2010 Update: User's Manual*; National Renewable Energy Laboratory: Golden, CO, USA, 2012.

15.　Letendre, S.; Makhyoun, M.; Taylor, M. *Predicting Solar Power Production: Irradiance Forecasting Models, Applications and Future Prospects*; Solar Electric Power Association: Washington, DC, USA, 2014. Available online: https://forecasting.energy.arizona.edu/media/papers/sepa2014.pdf (accessed on 16 April 2020).

16.　Cameron, C.P.; Boyson, W.E.; Riley, D.M. Comparison of PV system performance-model predictions with measured PV system performance. In Proceedings of the 2008 33rd IEEE Photovoltaic Specialists Conference (PVSC), San Diego, CA, USA, 11–16 May 2018; IEEE: Piscataway, NJ, USA, 2008; pp. 2099–2104. [CrossRef]

17. Lave, M.; Kleissl, J. Optimum Fixed Orientations and Benefits of Tracking for Capturing Solar Radiation in The Continental United States. *Renew. Energy* **2011**, *36*, 1145–1152. [CrossRef]

18. Kelly, N.A.; Gibson, T.L. Improved photovoltaic energy output for cloudy conditions with a solar tracking system. *Sol. Energy* **2009**, *83*, 2092–2102. [CrossRef]

19. Nelson, A.; Kelly, N.A.; Gibson, T.L. Increasing the solar photovoltaic energy capture on sunny and cloudy days. *Sol. Energy* **2011**, *85*, 111–125. [CrossRef]

20. Antonanzas, J.; Urraca, R.; Martinez-de-Pison, F.J.; Antonanzas, F. Optimal solar tracking strategy to increase irradiance in the plane of array under cloudy conditions: A study across Europe. *Sol. Energy* **2018**, *2018 163*, 122–130. [CrossRef]

21. Faine, P.; Kurtz, S.R.; Riordan, C.; Olson, J.M. The influence of spectral solar irradiance variations on the performance of selected single-junction and multijunction solar cells. *Sol. Cells* **1991**, *31*, 259–278. [CrossRef]

22. Baklouti, I.; Driss, Z.; Abid, M.S. Estimation of solar radiation on horizontal and inclined surfaces in Sfax, Tunisia. In Proceedings of the 2012 1st International Conference on Renewable Energies and Vehicular Technology (REVET 2012), Nabeul, Tunisia, 26–28 March 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 131–140. [CrossRef]

23. Breyer, C. Economics of Hybrid Photovoltaic Power Plants. In Proceedings of the 27th European Photovoltaic Sol. Energy Conference and Exhibition (27th EU PVSEC), Frankfurt, Germany, 24–28 September 2012; EU PVSEC: Lisboa, Portugal, 2012; pp. 4582–4593. [CrossRef]

24. Wei, C. Predictions of Surface Solar Radiation on Tilted Solar Panels using Machine Learning Models: A Case Study of Tainan City, Taiwan. *Energies* **2017**, *10*, 1660. [CrossRef]

25. George, A.; Anto, R. Analytical and experimental analysis of optimal tilt angle of solar photovoltaic systems. In Proceedings of the from 2012 International Conference on Green Technologies (ICGT), Trivandrum, Kerala, India, 18–20 December 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 234–239. [CrossRef]

26. Bakirci, K. General models for optimum tilt angles of solar panels: Turkey case study. *Renew. Sustain. Energy Rev.* **2012**, *16*, 6149–6159. [CrossRef]

27. Mekhilef, S.; Saidur, R.; Kamalisarvestani, M. Effect of dust, humidity and air velocity on efficiency of photovoltaic cells. *Renew. Sustain. Energy Rev.* **2012**, *16*, 2920–2925. [CrossRef]

28. Hosseini, S.A.; Kermani, A.M.; Arabhosseini, A. Experimental study of the dew formation effect on the performance of photovoltaic modules. *Renew. Energy* **2019**, *130*, 352–359. [CrossRef]

29. Ayvazoğluyüksel, Ö.; Filik, U.B. Estimation methods of global solar radiation, cell temperature and solar power forecasting: A review and case study in Eskişehir. *Renew. Sustain. Energy Rev.* **2018**, *91*, 639–653. [CrossRef]

30. Aldali, Y.; Celik, A.N.; Munee, T. Modelling and Experimental Verification of Solar Radiation on a Sloped Surface, Photovoltaic Cell Temperature, and Photovoltaic efficiency. *J. Energy Eng.* **2012**, *139*, 8–11. [CrossRef]

31. Skoplaki, E.; Palyvos, J.A. On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. *Sol. Energy* **2009**, *83*, 614–624. [CrossRef]

32. Mellit, A.; Saglam, S.; Kalogirou, S.A. Artificial neural network-based model for estimating the produced power of a photovoltaic module. *Renew. Energy* **2013**, *60*, 71–78. [CrossRef]

33. Zhou, W.; Yang, H.; Fang, Z. A novel model for photovoltaic array performance prediction. *Appl. Energy* **2007**, *84*, 1187–1198. [CrossRef]

34. Hammad, B.; Al-Abed, M.; Al-Ghandoor, A.; Al-Sardeah, A.; Al-Bashir, A. Modeling and analysis of dust and temperature effects on photovoltaic systems' performance and optimal cleaning frequency: Jordan case study. *Renew. Sustain. Energy Rev.* **2017**, *82*, 2218–2234. [CrossRef]

35. Busquet, S.; Kobayashi, J. In Proceedings of the Modelling daily PV performance as a function of irradiation, ambient temperature, soiling, wind speed, and aging—Applied to PV modules operating in Maui. In Proceedings of the 2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC), Waikoloa, HI, USA, 10–15 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3401–3406. [CrossRef]

36. Lu, H.; Zhao, W. Effects of particle sizes and tilt angles on dust deposition characteristics of a ground-mounted solar photovoltaic system. *Appl. Energy* **2018**, *220*, 514–526. [CrossRef]

37. International Civil Aviation Organization. Meteorological Service for International Air Navigation. In *International Standards and Recommended Practices: Annex 3 to the Convention on International Civil Aviation*, 16th ed.; ICAO: Montreal, QC, Canada, 2007.

38. UCAR Center for Science Education. *The Highs and Lows of Air Pressure*. Available online: https://scied.ucar.edu/shortcontent/highs-and-lows-air-pressure (accessed on 16 April 2020).

39. Energy Informative. *Which Sol. Panel Type Is Best? Mono- vs. Polycrystalline vs. Thin Film*. Available online: https://energyinformative.org/best-solar-panel-monocrystalline-polycrystalline-thin-film/ (accessed on 16 April 2020).

40. Hines, P.A.; Wagner, T.J.; Koschnick, C.M.; Schuldt, S.J. Analyzing the Efficiency of Horizontal Photovoltaic Cells in Various Climate Regions. *J. Energy Nat. Resour.* **2019**, *8*, 77–86. [CrossRef]

41. Williams, J.; Wagner, T. *Northern Hemisphere Horizontal Photovoltaic Power Output Data for 12 Sites*; Mendeley Data; Mendeley Ltd.: London, UK, 2019. [CrossRef]

42. National Oceanic and Atmospheric Administration. "National Center for Environmental Information," 2019. Available online: https://www.ncdc.noaa.gov/cdo-web/ (accessed on 24 April 2020).

43. H2O.ai. Available online: https://www.h2o.ai/ (accessed on 14 May 2019).

44. Cook, D. *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2016.

45. H2O.ai. *H2O 3.24.0.3 Documentation*. Available online: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/flow.html (accessed on 20 May 2019).

46. H2O.ai. *H2O 3.30.0.1 Documentation*. Available online: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html (accessed on 25 April 2020).

47. Hastie, T.; Tibshirani, R.; Friedman, J. *The Element of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Spring Science and Business Media: New York, NY, USA, 2009.

48. Baechler, M.C.; Williamson, J.L.; Gilbride, T.L.; Cole, P.C.; Hefty, M.G.; Love, P.M. *Building America Best Practices Series: Volume 7.1: Guide to Determining Climate Regions by County*; Pacific Northwest National Lab.(PNNL): Richland, WA, USA, 2010. [CrossRef]

49. Ahmad, M.W.; Mourshed, M.; Rezgui, Y. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy* **2018**, *164*, 465–474. [CrossRef]

50. Ramsami, P.; Oree, V. A hybrid method for forecasting the energy output of photovoltaic systems. *Energy Convers. Manag.* **2015**, *95*, 406–413. [CrossRef]

51. Pedro, H.T.C.; Coimbra, C.F.M. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Sol. Energy* **2012**, *86*, 2017–2028. [CrossRef]