Article

Generation of Data-Driven Expected Energy Models for Photovoltaic Systems

Michael W. Hopwood [1,2] and **Thushara Gunda [1,*]**

1 Sandia National Laboratories, Albuquerque, NM 87123, USA; mwhopwo@sandia.gov
2 Department of Statistics and Data Science, University of Central Florida, Orlando, FL 32826, USA
* Correspondence: tgunda@sandia.gov

Abstract: Although unique expected energy models can be generated for a given photovoltaic (PV) site, a standardized model is also needed to facilitate performance comparisons across fleets. Current standardized expected energy models for PV work well with sparse data, but they have demonstrated significant over-estimations, which impacts accurate diagnoses of field operations and maintenance issues. This research addresses this issue by using machine learning to develop a data-driven expected energy model that can more accurately generate inferences for energy production of PV systems. Irradiance and system capacity information was used from 172 sites across the United States to train a series of models using Lasso linear regression. The trained models generally perform better than the commonly used expected energy model from international standard (IEC 61724-1), with the two highest performing models ranging in model complexity from a third-order polynomial with 10 parameters ($R^2_{adj} = 0.994$) to a simpler, second-order polynomial with 4 parameters ($R^2_{adj} = 0.993$), the latter of which is subject to further evaluation. Subsequently, the trained models provide a more robust basis for identifying potential energy anomalies for operations and maintenance activities as well as informing planning-related financial assessments. We conclude with directions for future research, such as using splines to improve model continuity and better capture systems with low ($\leq$1000 kW DC) capacity.

Keywords: photovoltaic systems; expected energy models; fleet-scale; lasso regression; performance modeling; machine learning

1. Introduction

The increasing penetration of photovoltaic (PV) systems within the energy markets has established the need for evaluating and ensuring high system reliability. In particular, a large emphasis has been placed on monitoring algorithms that can contextualize observed energy generation at a site with information about how the system would have performed in a nominal state [1]. The latter are commonly estimated through expected energy models. Expected energy models are incorporated into many PV performance monitoring tasks, including anomaly detection [2–5], financial planning [6], fleet-level (site vs. site) comparisons [7], degradation analysis [7], and the evaluation of extreme weather effects [8]. The comparison of observed energy values to those derived from expected energy models serves as the basis for informing both tactical (i.e., short-term tasks such as field repair) and strategic (i.e., long-term activities such as site planning) operations and maintenance (O&M) activities.

Expected energy models can vary from asset-level to site-level estimates [9]. Asset-level models typically focus on using parameters provided by the manufacturer (e.g., maximum power) [9,10]. However, such approaches do not always work well for in-field performance since the parameters were developed in standardized test conditions and thus do not reflect operational conditions [11]. In response to these limitations, empirical methods that use field observations and regression methods have emerged to derive parameters across non-standardized test conditions (e.g., [12,13]). At the site-level, most expected

energy models leverage the correlation between power production and meteorological covariates [14,15]. For example, the standard expected energy model from the International Electrotechnical Commission (IEC) uses irradiance and site capacity information to develop an expected energy estimate [15]. Similarly, the PVUSA model trains a regression model for a given site by estimating power production using local irradiance, temperature, and wind speed conditions [16]. Industry research shows that most expected energy estimates tend to be overestimate production by a median of 3% but could be up to 20% [17]. Although the mismatch between observed and expected generation are well-recognized [18], limited attention has been given to date for improving the accuracy of expected energy models at the site-level, especially suited for fleet-level (i.e., site vs. site) comparisons.

This work aims to address this knowledge gap by generating a standardized, interpretable data-driven expected energy model that can be used for fleet-level comparisons. Although gradient-boosted and neural network-based methods have demonstrated significant successes for output performance [19–21], they often lack in model interpretability. In particular, models with high complexity can hide prediction biases or other vulnerabilities [22]. Thus for this work, we opted for more interpretable, regression-based models to increase the transparency of the implemented methods. In addition to identifying a more robust alternate for expected energy modeling, the associated publication of code used for training models (in the open source software pvOps) enables the extension of these methods to develop site-specific expected energy models for PV systems anywhere in the world or to other renewable energy systems. Such advancements in expected energy model estimates are needed to continue supporting better planning and field O&M activities, both of which ultimately influence the sustainability of PV sites. The following sections describes the data processing and model construction activities (Section 2), the performance of trained models (Section 3), and summarize primary findings (Section 4).

2. Methodology

The data-driven expected energy model training activities were supported by Sandia National Laboratories' PV Reliability, Operations, and Maintenance (PVROM) database [23]. Information about the PVROM database, as well as data processing, model training, and model evaluations, are described in the following Sections.

2.1. Data

The PVROM database contains 1.3 million data points of hourly production data across 176 sites in the United States [23], spanning multiple states (Figure 1) and generally ranging between 2017 and 2020. The database contains hourly measurements of expected energy in kiloWatt-hours (kWh), irradiance (Watts per square meter; $\frac{W}{m^2}$), ambient temperature, and module temperature; site-level direct current (DC) capacity is provided by the industry partners. The DC capacity (C_{DC}) for the sites within the database span from 37.8 kilowatts (kW) to 130,000 kW; a majority of the sites (140) are under 10,000 kW, with 67 of those sites under 1000 kW. A subset of the sites (100) contain industry-partner-provided expected energy estimates generated from proprietary models; these values serve as a basis for model validation activities (see Section 2.5).
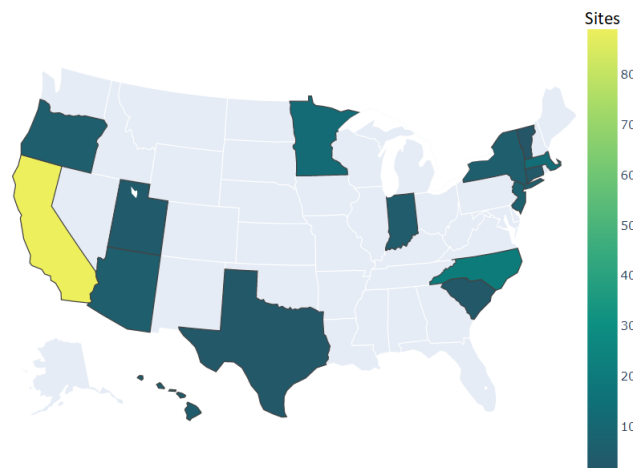
Figure 1. Geographical coverage of sites within the PVROM database. A majority of the sites are located in California.

2.2. Preprocessing

Data quality issues stemming from measurement errors and system anomalous conditions (reflecting local field failures, such as communication loss) could introduce signal variations in field data that would hinder model performance. Problematic data convolute the relationships between features, making it more difficult to measure the true parameter estimates; these potential irreducible errors are decreased through numerous data quality filters (Figure 2). Missing values (i.e., NaN or None values) were removed prior to applying data quality filters. An evaluation of these missing values revealed that a majority of them (~88%) occurred during nighttime hours (~7 p.m. to 8 a.m.), indicating that some sites captured night-time entries as null (Figure 2). After removing these missing values, ~900 K data points remained, which were then subject to a series of data quality filtering steps.
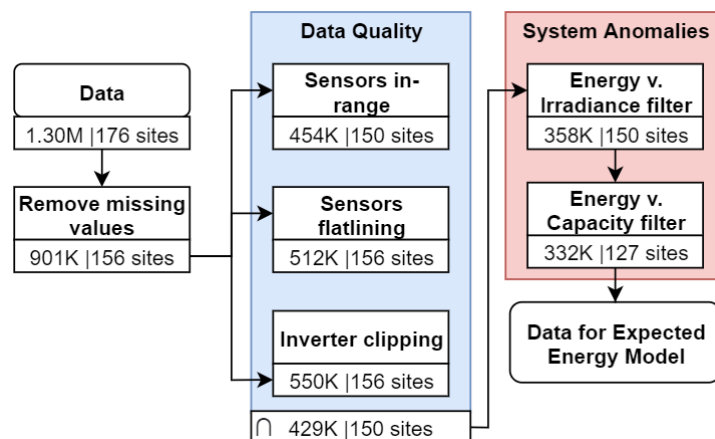
Figure 2. Data preprocessing activities included both data quality- and anomaly-related filters. Data quality filters were conducted independently; only data points that passed all quality-based filters were subject to the anomaly-based filters.

Data were filtered to ensure they are within nominal sensor ranges, using thresholds following [24] and the IEC 61724-1 standard [15]. Specifically, we retained data that met the following criteria:

- $20 \frac{W}{m^2} \leq$ Irradiance $(I) \leq 1500 \frac{W}{m^2}$;
- Energy $(E) > 0$ kWh;
- Ambient temperature $(T_{amb}) \leq 50\,^{\circ}\text{C}$ and module temperature $(T_{mod}) \leq 90\,^{\circ}\text{C}$.

Wind speed was not consistently available from partners and thus was excluded from analysis. Although available temperature data were used in the preprocessing steps, they are not used as a predictor variable in the regression models, since they ate not included in current standard models [15].

Flatlining values—determined by periods where consecutive data changed by less than a threshold—were flagged for removal using the `pecos` package [25], which follows the IEC 61724-3 standard [26]. Specifically, four consecutive hours with either $\Delta E < 0.01\%$ of the site's capacity or $\Delta I < 10 \frac{W}{m^2}$ were filtered. Lastly, inverter clipping, which occurs when the DC energy surpasses an inverter's DC energy rating, was addressed by mathematically observing plateaus in the energy signal using the `pvanalytics` package [27]. Dropping energy measurements during inverter clipping, which manifest as a static value across high irradiance levels, would create a better linear fit. After data quality checks, 429 K data points across 150 sites remained (Figure 2).

Data points that passed all quality checks were also assessed for system-level anomalies. These anomalies likely reflect abnormal operating conditions (i.e., local failures) and thus require removal to ensure the trained baseline energy models reflect nominal system performance. Anomalous entries were detected using a comparison of observed energy to irradiance and site capacities (Figure 3). The comparison of observed energy and irradiance filter focuses on removing data where the E–I ratio (λ) is outside its nominal distribution by 3 standard deviations $\{\lambda : \lambda < \mu_\lambda - 3\sigma_\lambda \cup \lambda > \mu_\lambda + 3\sigma_\lambda\}$, where μ_λ and σ_λ are the mean and standard deviation of the E–I ratio, respectively [28]. This filter was implemented for each site separately to capture site-specific variations (including system capacity) and resulted in the removal of 70 K data points (Figure 2). The second system anomaly filter focused on removing sites with mismatches between observed energy and site capacity. Namely, if a site's maximum recorded energy was over $1.2 \times C_{DC}$ or under $0.7 \times C_{DC}$, then all data points for that site were excluded from subsequent analysis. This method filtered 23 sites; 50%+ of these sites were under 1000 kW, and only 1 was over 10,000 kW. Approximately 26 K data points were removed with this filter, resulting in a final dataset that contained 332 K data points across 127 sites for model training and testing activities (Figure 2). The age of the sites within the final dataset ranged from newly installed sites up to 10 years, with a majority being less than 5 years in age (Figure A2).
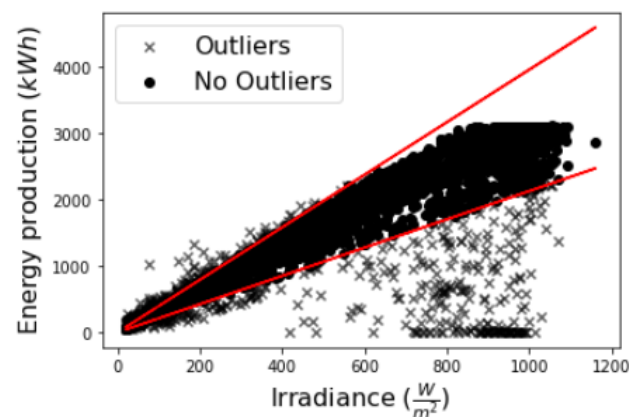
Figure 3. An example of anomaly-based filter (energy production vs. irradiance) for a particular site. Anomalous data points (visualized as Xs) are often lower than non-anomalous values within the distribution-derived bands (red lines).

2.3. Variable Standardization

The specific inputs used for model training mimic commonly available parameters used in current expected energy models (e.g., [15]), such as irradiance and site capacity. However, with covariates at different scales (e.g., $\{0\frac{W}{m^2} < I < 1.2 \times 10^3 \frac{W}{m^2}\}$ while $\{1 \times 10^2 \text{ kW} < C_{DC} < 1.3 \times 10^5 \text{ kW}\}$), variable standardization is required to reduce model sensitivity to parameter scales. In particular, without standardization, weights generated

for each parameter are more likely to reflect scalar nuances rather than the relative importance of the parameter to the outcome of interest. Variable standardization centers the data by subtracting data points in a feature from its associated mean value (μ) and then scales the data by dividing by the associated standard deviation (σ)—i.e., $Z = \frac{\mu - \bar{\mu}}{\sigma_X}$. The resulting standardized variables have a mean of zero and a standard deviation of one. This process makes parameters easier to rank in terms of influence; the variable with the larger parameter holds a more important effect on the output response. Thus, variable standardization also aids in the interpretability of the derived parameters, especially when variable interactions (e.g., $I \times C_{DC}$) are introduced. The mean and standard deviation parameters used to standardize irradiance, capacity, and energy values are captured in Table 1.

Table 1. Mean and standard deviation (StDev) parameters used to standardize variables prior to training regression models.

Parameter	>1000 kW Systems		<1000 kW Systems	
	Mean	**StDev**	**Mean**	**StDev**
Irradiance	571.459	324.199	413.533	286.110
Capacity	14,916.234	20,030.000	375.919	234.151
Energy	7449.152	12,054.525	119.008	119.829

2.4. Model Design and Training

Similar to other machine learning models, regression techniques leverage input data to learn relationships and use those relationships to predict unseen quantities. These relationships are generally contained in model parameters ($\hat{\beta}$), which map predictors, as summarized in a design matrix $\mathbf{X}$, to an output $\hat{Y} = \mathbf{X}\hat{\beta} + \epsilon$ with residual model error ϵ. Many different regression techniques exist; these techniques typically vary in the structure of the cost function, which quantifies the error between predicted and expected values. This cost function (C) is usually captured as a summation of loss functions (calculated on each data point) across the training set. The set $\hat{\beta}$, which renders the smallest cost, is defined as the learned parameters, mathematically notated as:

$$\hat{\beta} = \underset{\theta}{arg\min} C. \tag{1}$$

A popular regression model is the ordinary least squares (OLS), which defines its best model ($\hat{\beta}_{OLS} = \underset{\theta}{arg\min} SSE$) with an objective function equal to the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y - \sum_{j=0}^{p}\hat{\beta}_j x_{ij})^2 \tag{2}$$

where n is the number of samples, p is the number of predictors, and x_{ij} is the ith value for the jth explanatory variable. As shown in the equation, the SSE sums the squared difference between each sample (y) and its associated model estimate ($\hat{y}$). High emphasis is naturally placed on reducing high-error samples. Therefore, outliers can have a large effect on the learned parameters, so data preprocessing steps are required for robust model development. Additionally, OLS renders non-zero coefficients on all $\hat{\beta}$, which can create small, insubstantial parameters which are likely components of the training dataset and therefore contribute to model overfitting and thus should be removed from the model.

Alternate approaches to OLS include the Theil–Sen regressor [29], which is robust against outliers since it chooses the median of the slopes of all lines between pairs of points, as well as techniques such as Lasso regression [30] that explicitly address model overfitting by reducing model complexity (i.e., the number of parameters used). For this analysis, the latter was selected since Lasso regression models are able to incorporate both parameter regularization and residual sum of squares into the loss function. The cost function for

Lasso regression $\hat{\beta}_{lasso} = \underset{\theta}{arg\min}\left(SSE + \alpha \sum_{j=1}^{p}|\beta_j|\right)$ incorporates an L1 regularization term $\alpha \sum_{j=1}^{p}|\beta_j|$, which penalizes the magnitude of the β terms. This penalization tends to shrink coefficients to zero, rendering a more parsimonious model; we use an $\alpha = 0.003$ for defining the impact of the regularization on the regression kernel. Specifically, the penalization acts as a bias, which in turn can reduce overall error due to the bias–variance tradeoff [31].

Standardized variables are passed into Lasso regression to learn a linear model, which relates the input variables to energy. Multiple combinations of input variables were used to train the regression models (more details below). For all models, a randomized (80–20%) split is utilized to partition the preprocessed, standardized data into train and test partitions, respectively.

In addition to individual parameter influences, interactions and temporal factors were incorporated as input features to capture nuances within the datasets. Interaction parameters, which allow the effect of one parameter on the response variable to be weighted by the value of another variable, are introduced by including terms which are the product of two or more predictor variables. For example, Figure 4 shows that the relationship between E and I does vary across C_{DC}. Thus, the inclusion of an I and C_{DC} interaction term may be helpful in predicting the generated energy. The suite of interaction combinations are instantiated using polynomial models up to the third order (i.e., degree $d = 3$). In a model with $d = 2$ and 2 covariates, the initiated regression model would take the following form:

$$y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_1 x_2 + \beta_4 x_1 + \beta_5 x_2. \tag{3}$$

Notice that a $d = 2$ also includes $d = 1$ parameters (i.e., $\beta_4 x_1$ and $\beta_5 x_2$). This remains true for all values of the polynomial power (e.g., for a model initiated with $d = 3$, terms from $d = 2$ and $d = 1$ are also included). Two interaction polynomial orders are tested: a second-order ($d = 2$) and a third-order ($d = 3$) (Table 2). The particular interaction noted above ($I \times C_{DC}$) is captured in multiple models, including an additive model with a single interaction term (Table 2).

In addition to interactions, temporal factors are used to capture a variable's changing effect on the energy generated over time. For instance, the correlation between I and E changes over the course of the year due to spectral irradiance effects [32,33]. Therefore, allowing the model to capture time-variant nuances may be important for capturing such nonlinearities. Three temporal based conditions were explored: seasonal (four per year), monthly, and hourly. A model with two predictor variables and monthly temporal-based variable conditions would be instantiated as:

$$\begin{aligned}
y &= a_{jan}1_{t \in jan}x_1 + a_{feb}1_{t \in feb}x_1 + \ldots + a_{dec}1_{t \in dec}x_1 \\
&\quad + b_{jan}1_{t \in jan}x_2 + b_{feb}1_{t \in feb}x_2 + \ldots + b_{dec}1_{t \in dec}x_2,
\end{aligned} \tag{4}$$

where the a and b parameters are coefficients describing the effect of parameter x_1 and x_2, respectively, when conditioned on a month of the year. For instance, a_{jan} describes the effect of x_1 on the y response variable during the month of January. The indicator function 1_t masks the predictor variable to ensure it is within its timeframe. With the various combinations of interactions and temporal conditions, a total of 13 regression kernels were evaluated (Table 2; see Appendix A for some of the mathematical formulations).

(**a**)

(**b**)
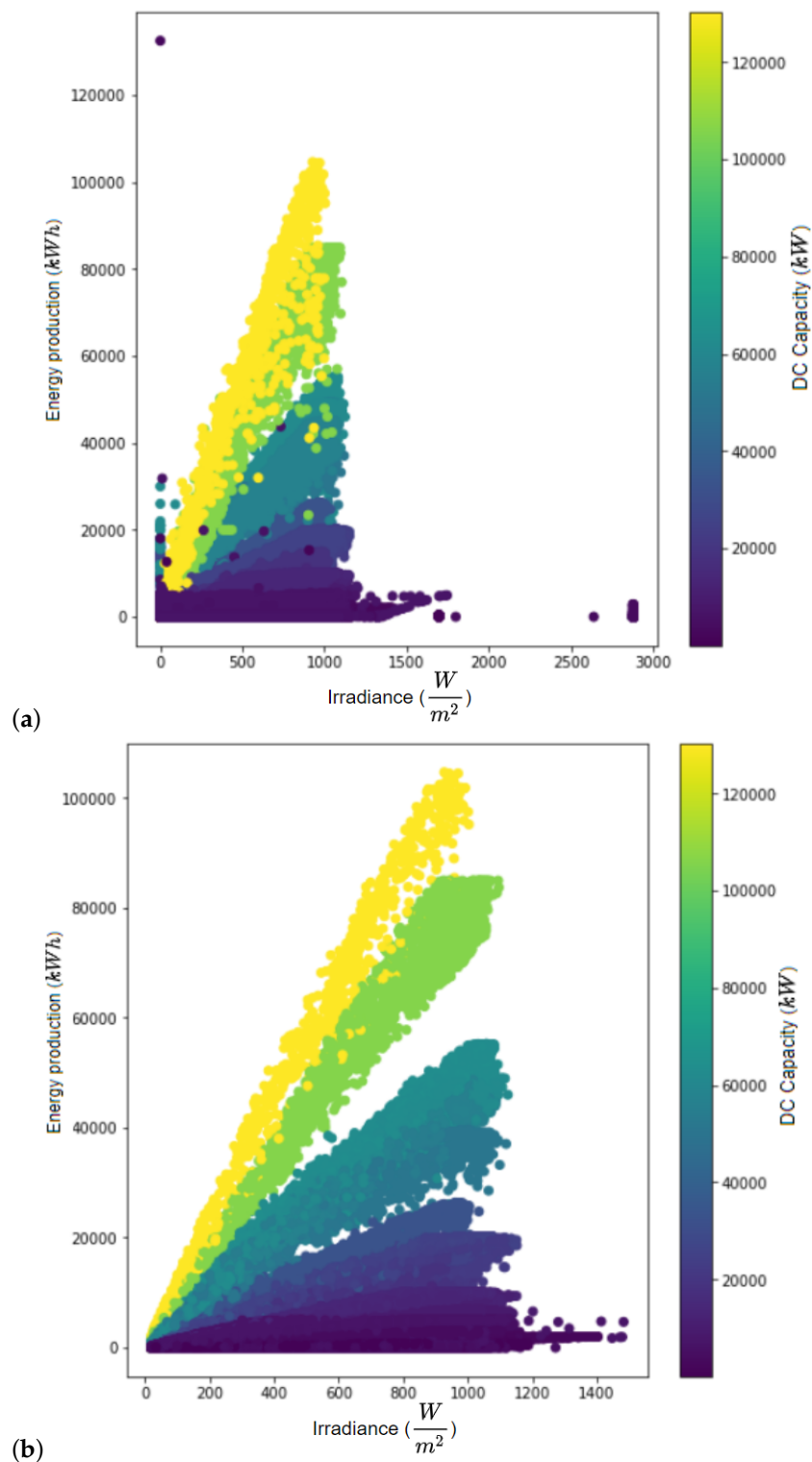
Figure 4. Correlation between energy production and irradiance for raw data (**a**) and preprocessed data (**b**) for different site capacities. Higher correlations in the preprocessed data indicate interaction between DC capacity and irradiance for energy production.

Table 2. Combinations of parameters used to initiate the 13 regression kernels evaluated in this study. The simple additive model attaches a parameter to each predictor variable to evaluate individual effects. * For hourly temporal conditions, only 15 h are used to reflect daytime hours.

Model Category	Time Variables (t)	Interaction Degree (d)	Number Parameters for M Inputs	Number Parameters for 2 Inputs		
Additive			$M + 1$	3		
Additive with interaction		2nd-order (d = 2)	$M + 2$	4		
Additive with Time-weighted	Season (t = 4) Month (t = 12) Hour (t = 15 *)		$tM + 1$	9 25 31		
Polynomial		2nd-order (d = 2) 3rd-order (d = 3)	$\sum_{i=1}^{d} C^R(M, i) + 1$	6 10		
Polynomial with Time-weighted	Season (t = 4) Month (t = 12) Hour (t = 15 *)	2nd-order (d = 2) 3rd-order (d = 3)	$t \sum_{i=1}^{d} C^R(M, i) + 1$	t = 4 t = 12 t = 15	d = 2 25 73 90	d = 3 41 121 151

2.5. Model Evaluation

Three metrics were used to evaluate the performance of the trained expected energy models: logarithmic root mean squared error (log $RMSE$), coefficient of determination (R^2), and percent error (δ). Both partner-provided expected energy values and those calculated by the leading standardized expected energy model (i.e., IEC 61724) were used as reference values for model evaluations.

The root mean squared error ($RMSE$) is a common goodness-of-fit statistic used for model evaluation. The $RMSE$ is expressed as:

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

where y_i and $\hat{y}_i$ are the measured and predicted values of the response variable, and n is the number of samples. $RMSE$ is in the same units as the response variable (i.e., kWh). Lower $RMSE$ values indicate a better, lower predicted error. Because the error can be quite large in magnitude (10^0 to 10^{10}), a logarithmic transform is applied to facilitate evaluations. Because the magnitude of the error is closely connected to a site's capacity, the log $RMSE$ cannot be used to compare model performance between sites unless the sites are similar in size.

The coefficient of determination (R^2), however, can be used to compare model performance across different site sizes. Specifically, R^2 is calculated as:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \tag{5}$$

where $\bar{y}$ is the average of the y values. R^2 denotes the proportion of variability in the response explained by the model with a value of 1, indicating a perfect fit. R^2 was used to compare trained model outputs with partner-generated expected energy values, whose underlying model structures were unknown.

Generally, however, R^2 is not well-suited for comparing models across varying numbers of parameters. Thus, when comparing the 13 trained models to one another, we utilize an adjusted R^2_{adj} metric, which checks whether the added parameters contribute to the explanation of the predictor variable and penalizes models with unnecessary complexity [34]. Low-effect parameters (i.e., $\beta \approx 0$) reduce the model's overall fit score. The adjusted R^2_{adj} is calculated as follows:

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1} \right), \tag{6}$$

where n is the number of samples, and p is the number of predictors.

Finally, δ was used to capture the directionality of error (i.e., overprediction vs. underprediction):

$$\delta = 100 \times \frac{\hat{y} - y}{y}. \tag{7}$$

The log $RMSE$ and R^2 were implemented to evaluate model performance at both site and fleet (i.e., across multiple sites) levels, while δ was only implemented at the fleet level; all metrics were reported on the test dataset. T-tests were used to evaluate significance in performance variations between the trained and reference values.

3. Results and Discussion

Data processing activities generally increased the correlations between the predictor variables (i.e., irradiance and capacity) and the response variable (i.e., energy) (Table A1). The processed data were inputted into a total of 13 trained models—ranging in model complexity (pre-lasso) from 3 parameters for the 'simple additive' model pre-lasso to 151 parameters for the 'third-order-hour' (see Table 3). Generally, the number of parameters were lower for all models post-lasso fit, except for the 'simple additive' and 'additive interaction' models, likely indicating the already sparse construction of these models.

Table 3. This table describes the parameterization and performance of all of the models evaluated in the results of this paper. Wins are summarized by showing the percentage of sites where a given model was the top performer according to the associated goodness-of-fit metric (Adj. R^2 or log $RMSE$); the IEC model was used as the reference value for log $RMSE$ calculations. The third-order interactions model and basic model perform consistently well, a conclusion also found on the heatmaps (Figure 5). Additionally, because lasso regression was leveraged, the models decrease in size after training the model.
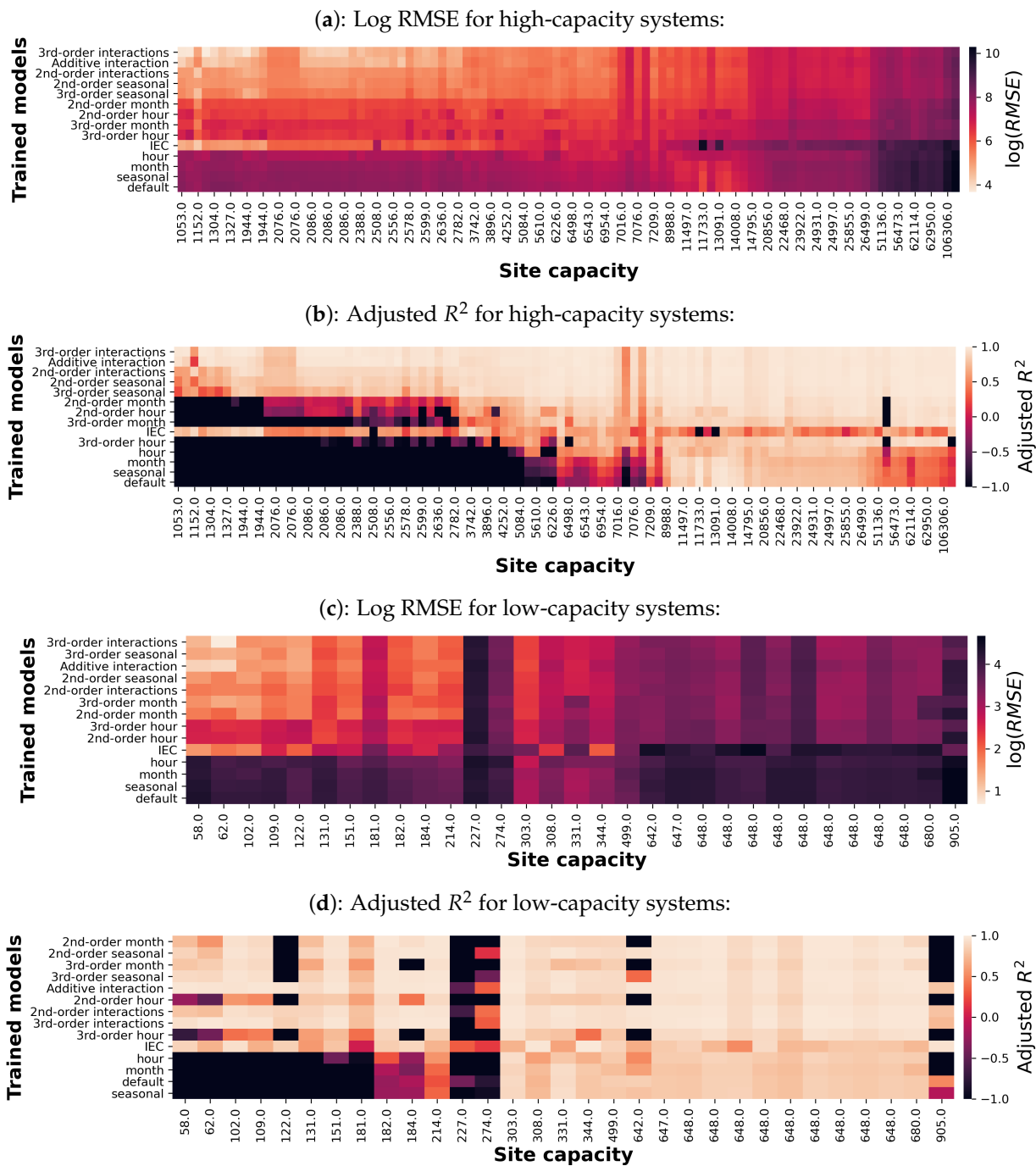
Models	Time Variable	Interaction Degree	Number Parameters Pre-Lasso	Number Parameters Post-Lasso	>1000 kW Systems		<1000 kW Systems	
					Wins: Adj. R^2	Wins: logRMSE	Wins: Adj. R^2	Wins: logRMSE
Third-order interactions		3	10	9	47.9	49.0	38.7	38.7
Additive interaction		1	4	4	22.9	17.7	32.3	29.0
Second-order interactions		2	6	5	15.6	10.4	6.5	0.0
Second-order seasonal	season	2	25	18	9.4	15.6	3.2	3.2
Third-order seasonal	season	3	41	27	1.0	4.2	3.2	3.2
Second-order month	month	2	73	43	0.0	0.0	3.2	3.2
Second-order hour	hour	2	90	44	0.0	0.0	0.0	0.0
Third-order month	month	3	121	68	0.0	0.0	0.0	3.2
Third-order hour	hour	3	151	81	1.0	1.0	0.0	6.5
IEC					1.0	1.0	12.9	12.9
hour	hour	1	31	26	0.0	0.0	0.0	0.0
month	month	1	25	25	0.0	0.0	0.0	0.0
seasonal	season	1	9	9	1.0	1.0	0.0	0.0
simple additive		1	3	3	0.0	0.0	0.0	0.0

Initially, the various models were trained using data across all system sizes. However, this approach demonstrated systemic underperformance for low-capacity systems (<1000 kW DC capacity). Specifically, the best trained models (i.e., 'third-order interactions' and 'additive interaction') outperformed the IEC model in terms of log *RMSE* when tested on every single system above 1300 kW DC capacity; however, 12 of 34 systems below a 1300 kW DC capacity underperformed relative to the IEC model. This result likely reflects the varying relationships between the site DC capacity and the energy generated; systems of higher capacity tend to receive a higher maximum energy generated per DC capacity (Figure A1). To better deal with this varying linearity, two separate sets of models were trained: one for models under 1000 kW DC capacity and another over.

Across both high-capacity and low-capacity systems, models with the $I \times C_{DC}$ interaction term perform better than those without the interaction term (i.e., 'hour', 'month', 'seasonal', and 'simple additive' models). For example, two of the top-performing models (across both high-capacity and low-capacity systems) are the 'simple additive' and the 'third-order interactions' models, both of which contain this interaction term (Table 3 and Figure 5). The 'additive interaction' trained (AIT) model has four parameters:

$$
\hat{e} = \begin{cases} 0.07 + 0.69i + 0.65c + 0.42ic & C_{DC} < 1000 \text{ kW} \\ -0.06 + 0.29i + 0.76c + 0.40ic & C_{DC} \geq 1000 \text{ kW} \end{cases} \tag{8}
$$

where *i* and *c* define the standardized irradiance and capacity variables, respectively, as defined in Table 1. The 'third-order interactions' model, on the other hand, contains these four terms as well as higher-order interactions (e.g., irradiance^2 × capacity, capacity^3). Although the variables within both of these two models are similar to the IEC standard, the inclusion of the interaction term, which highlights that the linear relationship between *I* and *E* is moderated by C_{DC} (Figure 4), likely explains the superior performance of these models relative to that standard. The heatmaps of the log *RMSE* values highlight the evaluation metric's dependence on site capacity (Figure 5a,c), while the adjusted R^2 heatmaps show consistent performance across site capacity (Figure 5b,d). The vertical concentration of dark bars likely reflects data quality issues not addressed by data preprocessing steps. A comparison of the associated partner generated expected energy estimates to those predicted from models also demonstrates that the AIT-derived estimates have lower average percent errors than the other models and the IEC (Table A2). Further evaluation of 2 years of records at a single site demonstrates that the AIT-derived estimates have a lower standard deviation and do not overestimate as much as the IEC-derived estimates (Figure A3). Given its parsimonious nature, the AIT model is subjected to further evaluation for both high- and low-capacity systems.
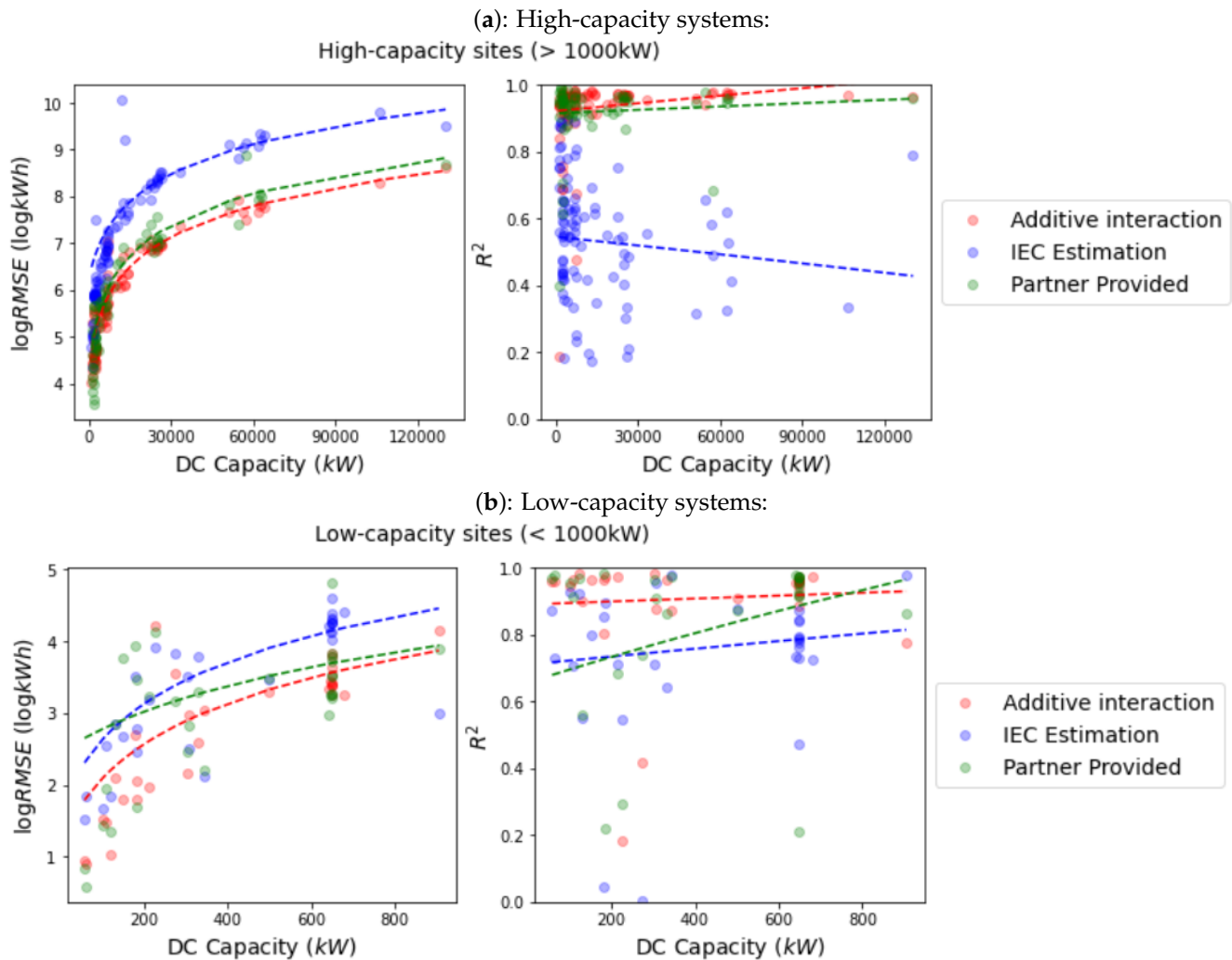
(a): Log RMSE for high-capacity systems:

(b): Adjusted R^2 for high-capacity systems:

(c): Log RMSE for low-capacity systems:

(d): Adjusted R^2 for low-capacity systems:

Figure 5. High-capacity (**a**,**b**) and low-capacity (**c**,**d**) site-level model evaluations with test data. Models with lighter colors (i.e., low values for log *RMSE* and values closer to 1 for adjusted R^2) indicate better performance.
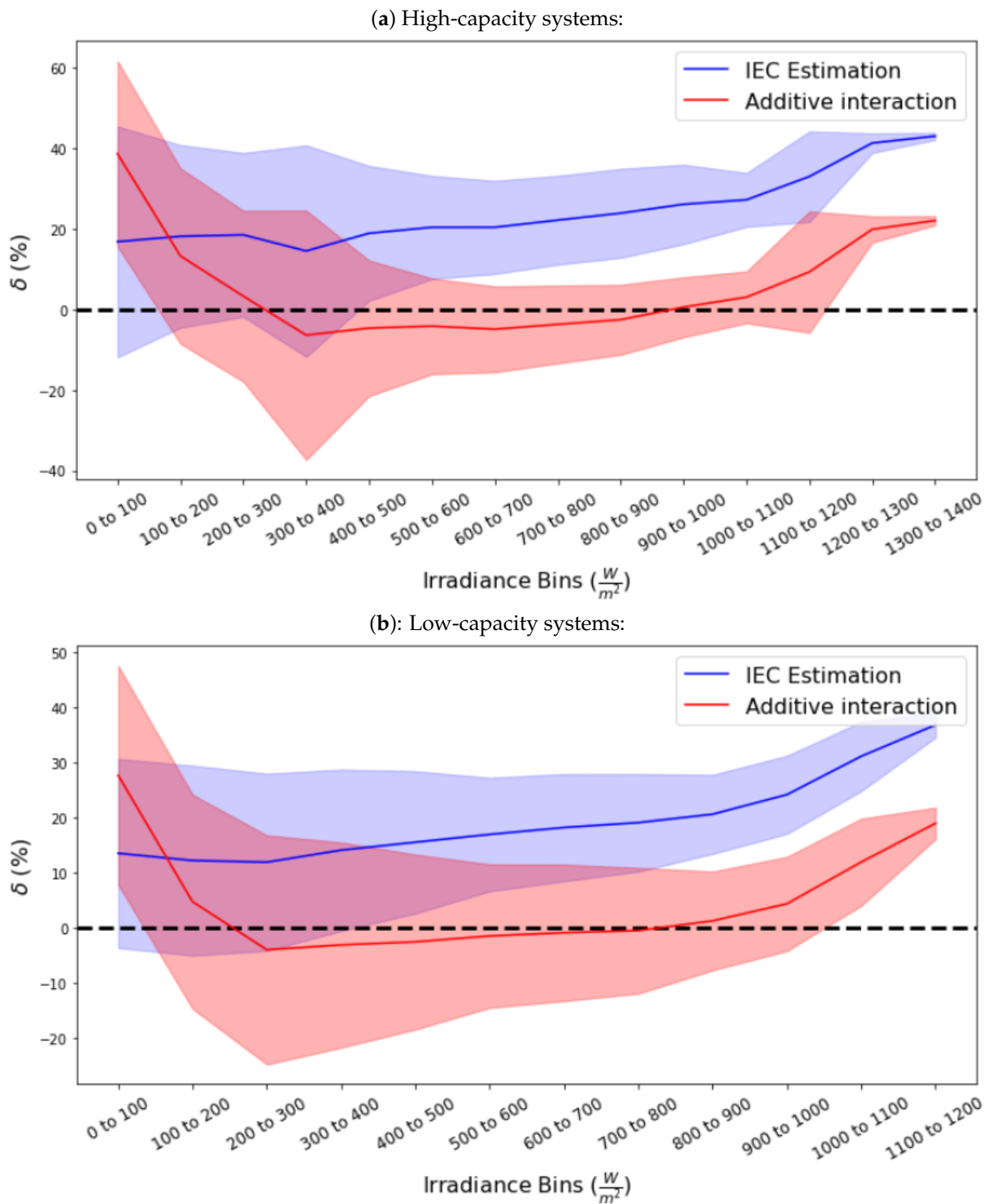
3.1. High-Capacity Systems

For high-capacity systems, a significant (almost uniform) difference is found in both the log *RMSE* and R^2 values between the AIT and the IEC reference models (Figure 6a,b). Across the sites, the AIT model improves the goodness of fit by 0.42 in R^2 (IEC: 0.501; AIT: 0.93) and 1.16 in log *RMSE* (IEC: 6.99; AIT: 5.83). Generally, there are very few systems for which the IEC model performs better than the trained models (Table 3). The percent error (δ), on average, of the AIT model (3.65) is significantly lower than the IEC model (20.86) for high-capacity systems. An evaluation of percent error shows that the AIT generally

performs well (i.e., $\delta \approx 0$ and thinner standard deviation bars) for most irradiance levels, except at the two extremes (i.e., <200 and $>1100 \frac{W}{m^2}$) (Figure 7).

The difference in performance relative to the IEC model is especially pronounced for larger system sizes (Figure 6a,b). Although the AIT model appears to show a small improvement over the partner-provided values (Figure 6a,b), a *t*-test concluded that the distributions of both the log *RMSE* (*p*-value: 0.78) and R^2 (*p*-value: 0.48) are not significantly different.

(**a**): High-capacity systems:

(**b**): Low-capacity systems:

Figure 6. Model evaluation using log *RMSE* and R^2 metrics for high-capacity systems (**a**) and low-capacity systems (**b**). Data points reflect site-level summaries of associated test data while dotted lines reflect best line fits to support visual pattern identification. The R^2 metric was used for this analysis (vs. adjusted R^2) since partner-provided model architectures are unknown. The 'additive interaction' regression model (in red) is comparable to the partner-provided proprietary values (green) and consistently performs better than the IEC standard (blue), especially at higher capacity values.

(**a**) High-capacity systems:

(**b**): Low-capacity systems:

Figure 7. Percent error as a function of irradiance shows that the 'additive interaction' model outperforms the IEC standard across both high-capacity (**a**) and low-capacity (**b**) systems. Lines indicate mean values, while shaded region captures one standard deviation. The 'additive interaction' model performs best ($\delta \approx 0$) at 500–1100 $\frac{\text{W}}{\text{m}^2}$ and at 200–1000 $\frac{\text{W}}{\text{m}^2}$ for high-capacity and low-capacity systems, respectively.

3.2. Low-Capacity Systems

The model performance of the AIT for low-capacity systems was generally comparable to that of high-capacity systems, although the improvements were not as high. Across all low-capacity sites, the AIT model's goodness of fit improved by 0.165 in R^2 (IEC: 0.74; AIT:

0.90) and 0.61 in log $RMSE$ (IEC: 3.35; AIT: 2.75) (Figure 6a,b). Out of the 31 low-capacity sites (comprising 50K hours), the IEC-based model outperformed the trained models in 4 systems (Table 3). In some of the low-capacity systems, the measured energy is much higher than expected (Figure 5). The tendency of the IEC model to overpredict likely describes why this model performs better for some of the lower-capacity systems.

The δ, on average, is 4.42 and 15.37 for the AIT model and IEC model, respectively. Similar to the high-capacity system, the percent error values are greater at the extremes (Figure 7b). However, the standard error is generally higher in the low-capacity systems, as evidenced by wider standard deviation bars across the irradiance levels (Figure 7b).

3.3. Limitations and Future Work

The methodological approach of this analysis was strongly guided by available data. However, future work could extend these methods to consider: (1) energy generation at finer resolutions, (2) additional co-variates, and (3) alternate model formulations. For example, scaling could be used to consider alternate frequencies (beyond the hourly intervals considered in this study) post-evaluation. The methods used in this analysis explicitly omitted variables not included in current standard models (e.g., [15]). However, future assessments could more explicitly incorporate co-variates such as temperature, wind speed, and even age of the site. The latter would especially enable active consideration of degradation influence, which can influence long-term energy generation of PV sites [35]. Additional co-variates (such as type of inverters and modules) could also be included in subsequent iterations to capture more subtle impacts associated with differing site designs. Finally, future work could consider alternate model formulations (e.g., splines) to improve model continuity and better capture energy generation for smaller system sizes.

4. Conclusions

This work demonstrates the opportunities for leveraging data-driven, machine learning methods to generate more robust expected energy models. Generally, when compared to partner-provided values, the trained regression models outperform the IEC standard, especially in high-capacity systems. Detailed evaluation of the parsimonious AIT or 'additive interaction' model, in particular, demonstrated significant potential for use as a standardized, fleet-level expected energy model. The specific code used to train the regression models as well as the AIT model have been integrated with pvOps, an open source Python package which supports the evaluation of field data by PV researchers and operators; pvOps can be accessed at https://github.com/sandialabs/pvOps, accessed on 22 December 2021. Although this work presents findings specific to PV systems, the general methodologies can be applied to any domain that uses expected energy models to support site planning and O&M activities. Ongoing evaluations and improvements of these standardized expected energy models will continue to increase the accuracy and precision of site-level PV performance evaluations, which is critical to supporting reliability and economic assessments of PV operations and maintenance.

Author Contributions: Conceptualization, T.G.; methodology, M.W.H. and T.G.; software, M.W.H.; validation, M.W.H. and T.G.; data curation, M.W.H.; writing—original draft preparation, M.W.H.; writing—review and editing, T.G.; visualization, M.W.H.; supervision, T.G.; project administration, T.G.; funding acquisition, T.G. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Abbreviations

The following abbreviations are used in this manuscript:

E	Energy (kWh)
$\hat{E}$	Expected energy (kWh)
I	irradiance ($\frac{W}{m^2}$)
C_{DC}	DC capacity (kW)
T_{amb}	ambient temperature (°C)
T_{mod}	module temperature (°C)
e	standardized E (see Table 1)
i	standardized I (see Table 1)
c	standardized C (see Table 1)
$RMSE$	Root mean squared error
n	number of samples
p	number of predictors

Appendix A

Appendix A.1. Tables & Figures

Table A1. Correlations with energy generation and regression model parameters pre- and post- data processing. Correlations with energy production are generally comparable with irradiance across raw and filtered data, while correlations for site capacity are significantly higher for the filtered data than the raw data.

	Data Subsets		
Parameters	**Raw**	**Post-Data Quality Filters**	**Post-System Anomaly Filters**
Irradiance (I)	0.46	0.41	0.40
Capacity (C_{DC})	0.55	0.85	0.89

Table A2. Summary of average percent errors for each of the trained models, using the partner-generated values as the reference value.

	Average Percent Error	
	>1000 kW Systems	**≤1000 kW Systems**
Third-order interactions	0.03	0.23
Second-order interactions	0.03	0.21
Third-order seasonal	0.05	0.20
Second-order seasonal	0.04	0.16

Table A2. *Cont.*

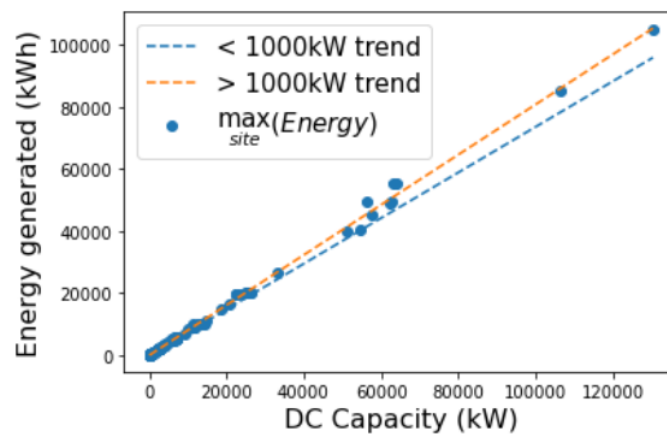|	Average Percent Error	
	>1000 kW Systems	**≤1000 kW Systems**
Third-order month	0.12	0.27
Second-order month	0.08	0.27
Third-order hour	0.13	0.15
Second-order hour	0.10	0.05
Additive interaction	0.03	0.00
simple additive	0.36	1.44
IEC	0.39	1.75
hour	0.28	1.01
month	0.32	1.30
seasonal	0.32	1.55

Figure A1. Relationship between DC capacity (kW) and hourly generated energy (kWh). The blue dots show a site's DC capacity versus its maximum recorded energy generated in a single hour. Although the trends are largely linear, the slopes differ for sites smaller than 1000 kW (blue dashed lines) and sites larger than 1000 kW (orange dashed lines). Since slight deviations in slope can render large prediction error, we train two separate model based on site size.
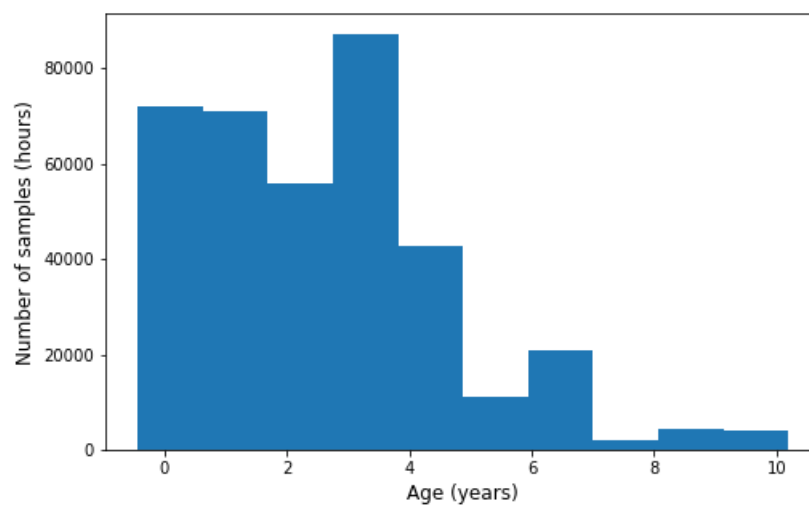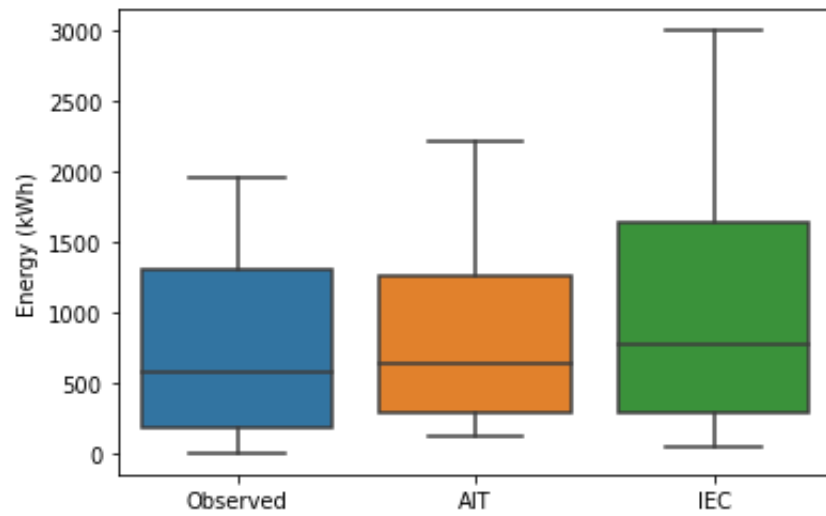

Figure A2. Age of sites within the final dataset. A majority are less than five years of age.

Figure A3. A 2-year comparison of observed energy with expected energy estimates from the trained additive interaction trained (AIT) model and the IEC standard at a single site. The IEC-derived estimates overestimate much more so than AIT-derived estimates. Slightly higher values of AIT-derived estimates relative to observed values likely indicate local failures at the site.

Appendix A.2. Top-Performing Trained Models

Mathematical equations associated with the trained regression models. In general, these model formulations contain more parameters and do not perform as well as the additive interaction model. High-capacity refers to systems with greater than or equal to 1000 kW in C_{DC} while low-capacity refers to systems smaller than 1000 kW in C_{DC}.

High-capacity second-order seasonal:

$$\begin{aligned}
e = \ & 0.309iI_{\text{winter}} + 0.292iI_{\text{spring}} + 0.281iI_{\text{summer}} + 0.287iI_{\text{fall}} \\
& + 0.762cI_{\text{winter}} + 0.734cI_{\text{spring}} + 0.724cI_{\text{summer}} + 0.730cI_{\text{fall}} \\
& - 0.003i^2I_{\text{spring}} - 0.009i^2I_{\text{summer}} - 0.007i^2I_{\text{fall}} + 0.430icI_{\text{winter}} \\
& + 0.397icI_{\text{spring}} + 0.380icI_{\text{summer}} + 0.384icI_{\text{fall}} + 0.008c^2I_{\text{winter}} \\
& + 0.008c^2I_{\text{spring}} - 0.003c^2I_{\text{fall}} - 0.054
\end{aligned}$$

High-capacity third-order interactions:

$$e = 0.293i + 0.769c - 0.019i^2 + 0.394ic + 0.021i^2c + 0.004ic^2 + 0.001c^3 - 0.039$$

High-capacity third-order seasonal:

$$\begin{aligned}
e = \ & 0.265iI_{\text{winter}} + 0.270iI_{\text{spring}} + 0.252iI_{\text{summer}} + 0.243iI_{\text{fall}} \\
& + 0.749cI_{\text{winter}} + 0.740cI_{\text{spring}} + 0.722cI_{\text{summer}} + 0.730cI_{\text{fall}} \\
& - 0.006i^2I_{\text{summer}} - 0.388icI_{\text{winter}} + 0.378icI_{\text{spring}} + 0.365icI_{\text{summer}} \\
& + 0.346icI_{\text{fall}} + 0.019c^2I_{\text{winter}} + 0.005c^2I_{\text{spring}} - 0.002c^2I_{\text{fall}} \\
& + 0.016i^3I_{\text{winter}} + 0.008i^3I_{\text{spring}} + 0.014i^3I_{\text{summer}} + 0.017i^3I_{\text{fall}} \\
& - 0.013ic^2I_{\text{winter}} + 0.008ic^2I_{\text{spring}} + 0.007ic^2I_{\text{summer}} + 0.019ic^2I_{\text{fall}} \\
& - 0.002c^3I_{\text{winter}} - 0.054
\end{aligned}$$

High-capacity second-order interactions:

$$e = 0.295i + 0.745c - 0.018i^2 + 0.404ic + 0.004c^2 - 0.042$$

Low-capacity second-order seasonal:

$$
\begin{aligned}
e = {} & 0.711iI_{\text{winter}} + 0.686iI_{\text{spring}} + 0.680iI_{\text{summer}} + 0.714iI_{\text{fall}} \\
& + 0.617cI_{\text{winter}} + 0.576cI_{\text{spring}} + 0.605cI_{\text{summer}} + 0.636cI_{\text{fall}} \\
& - 0.046i^2 I_{\text{winter}} - 0.028i^2 I_{\text{spring}} - 0.046i^2 I_{\text{summer}} - 0.038i^2 I_{\text{fall}} \\
& + 0.401icI_{\text{winter}} + 0.381icI_{\text{spring}} + 0.382icI_{\text{summer}} + 0.427icI_{\text{fall}} \\
& + 0.001c^2 I_{\text{winter}} - 0.006c^2 I_{\text{summer}} - 0.104
\end{aligned}
$$

Low-capacity third-order interactions:

$$e = 0.746i + 0.642c - 0.043i^2 + 0.426ic - 0.019i^3 - 0.034i^2 c - 0.017c^3 + 0.115$$

Low-capacity third-order seasonal:

$$
\begin{aligned}
e = {} & 0.757iI_{\text{winter}} + 0.684iI_{\text{spring}} + 0.667iI_{\text{summer}} + 0.712iI_{\text{fall}} \\
& + 0.577cI_{\text{winter}} + 0.486cI_{\text{spring}} + 0.567cI_{\text{summer}} + 0.619cI_{\text{fall}} \\
& - 0.023i^2 I_{\text{winter}} - 0.027i^2 I_{\text{spring}} - 0.050i^2 I_{\text{summer}} - 0.033i^2 I_{\text{fall}} \\
& + 0.413icI_{\text{winter}} + 0.382icI_{\text{spring}} + 0.381icI_{\text{summer}} + 0.428icI_{\text{fall}} \\
& - 0.028i^3 I_{\text{winter}} - 0.007i^3 I_{\text{summer}} - 0.022i^2 cI_{\text{winter}} - 0.004i^2 cI_{\text{fall}} \\
& - 0.047c^3 I_{\text{winter}} + 0.068c^3 I_{\text{spring}} + 0.027c^3 I_{\text{summer}} + 0.016c^3 I_{\text{fall}} + 0.099
\end{aligned}
$$

Low-capacity second-order interactions:

$$e = 0.719i + 0.630c - 0.063i^2 + 0.411ic - 0.124$$

References

1. Golnas, A. PV system reliability: An operator's perspective. In Proceedings of the 2012 IEEE 38th Photovoltaic Specialists Conference (PVSC) PART 2, Austin, TX, USA, 3–8 June 2012; pp. 1–6.
2. Feng, M.; Bashir, N.; Shenoy, P.; Irwin, D.; Kosanovic, D. SunDown: Model-driven Per-Panel Solar Anomaly Detection for Residential Arrays. In Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies, Guayaquil, Ecuador, 15–17 June 2020; pp. 291–295.
3. Dyamond, W.; Rix, A. Detecting Anomalous Events for a Grid Connected PV Power Plant Using Sensor Data. In Proceedings of the 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA), Bloemfontein, South Africa, 29–31 January 2019; pp. 287–292.
4. Harrou, F.; Sun, Y.; Taghezouit, B.; Saidi, A.; Hamlati, M.E. Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches. *Renew. Energy* **2018**, *116*, 22–37. [CrossRef]
5. Platon, R.; Martel, J.; Woodruff, N.; Chau, T.Y. Online fault detection in PV systems. *IEEE Trans. Sustain. Energy* **2015**, *6*, 1200–1207. [CrossRef]
6. Blair, N.; Dobos, A.P.; Freeman, J.; Neises, T.; Wagner, M.; Ferguson, T.; Gilman, P.; Janzou, S. *System Advisor Model, Sam 2014.1. 14: General Description*; Technical Report; National Renewable Energy Lab.: Golden, CO, USA, 2014.
7. Deline, C.; Anderson, K.; Jordan, D.; Walker, A.; Desai, J.; Perry, K.; Muller, M.; Marion, B.; White, R. *PV Fleet Performance Data Initiative: Performance Index-Based Analysis*; Technical Report; National Renewable Energy Lab.: Golden, CO, USA, 2021.
8. Gilletly, S.; Jackson, N.; Staid, A. Quantifying Wildfire-Induced Impacts to Photovoltaic Energy Production in the western United States. In Proceedings of the 2021 IEEE 48th Photovoltaic Specialists Conference (PVSC), Online, 20–25 June 2021.
9. Klise, G.T.; Stein, J.S. *Models Used to Assess the Performance of Photovoltaic Systems*; Sandia National Laboratories: Albuquerque, NM, USA, 2009.
10. De Soto, W.; Klein, S.A.; Beckman, W.A. Improvement and validation of a model for photovoltaic array performance. *Sol. Energy* **2006**, *80*, 78–88. [CrossRef]

11. Arab, A.H.; Taghezouit, B.; Abdeladim, K.; Semaoui, S.; Razagui, A.; Gherbi, A.; Boulahchiche, S.; Mahammed, I.H. Maximum power output performance modeling of solar photovoltaic modules. *Energy Rep.* **2020**, *6*, 680–686. [CrossRef]

12. King, D.L.; Kratochvil, J.A.; Boyson, W.E. *Photovoltaic Array Performance Model*; Sandia National Laboratories: Albuquerque, NM, USA, 2004.

13. Boyson, W.E.; Galbraith, G.M.; King, D.L.; Gonzalez, S. *Performance Model for Grid-Connected Photovoltaic Inverters*; Technical Report; Sandia National Laboratories: Albuquerque, NM, USA, 2007.

14. Dobos, A.P. *PVWatts Version 5 Manual*; Technical Report; National Renewable Energy Lab.: Golden, CO, USA, 2014.

15. *ISO/IEC TR 29110-1:2016*; *Photovoltaic System Performance—Part 1: Monitoring*. International Electrotechnical Commission: Geneva, CH, USA, 2017.

16. Myers, D. *Evaluation of the Performance of the Pvusa Rating Methodology Applied to Dual Junction pv Technology: Preprint (Revised)*; Technical Report; National Renewable Energy Lab.: Golden, CO, USA, 2009.

17. DNV GL. *Narrowing the Performance Gap: Reconciling Predicted and Actual Energy Production*; Technical Report; DNV GL: Oakland, CA, USA, 2019.

18. Hanawalt, S. *The Challenge of Perfect Operating Data*; Technical Report; Power Factors: San Francisco, CA, USA, 2020.

19. Livera, A.; Theristis, M.; Makrides, G.; Sutterlueti, J.; Ransome, S.; Georghiou, G.E. Performance Analysis of Mechanistic and Machine Learning models for Photovoltaic energy yield prediction. In Proceedings of the 36th European Photovoltaic Solar Energy Conference and Exhibition, Marseille, France, 9–13 September 2019; pp. 9–13.

20. Ascencio-Vásquez, J.; Bevc, J.; Reba, K.; Brecl, K.; Jankovec, M.; Topič, M. Advanced PV Performance Modelling Based on Different Levels of Irradiance Data Accuracy. *Energies* **2020**, *13*, 2166. [CrossRef]

21. Wang, S.; Zhang, Y.; Zhang, C.; Yang, M. Improved artificial neural network method for predicting photovoltaic output performance. *Glob. Energy Interconnect.* **2020**, *3*, 553–561. [CrossRef]

22. Bertsimas, D.; Delarue, A.; Jaillet, P.; Martin, S. The price of interpretability. *arXiv* **2019**, arXiv:1907.03419.

23. Klise, G.T. *PV Reliability Operations Maintenance (PVROM) Database: Data Collection & Analysis Insights*; Technical Report; Sandia National Lab.: Albuquerque, NM, USA, 2015.

24. Klise, K.A.; Stein, J.S.; Cunningham, J. Application of IEC 61724 Standards to Analyze PV System Performance in Different Climates. In Proceedings of the 2017 IEEE 44th Photovoltaic Specialist Conference (PVSC), Washington, DC, USA, 25–30 June 2017; pp. 3161–3166.

25. Klise, K.A. *Performance Monitoring Using Pecos*; Technical Report; Sandia National Lab.: Albuquerque, NM, USA, 2016.

26. *IEC TS 61724-3:2016*; Photovoltaic System Performance—Part 3: Energy Evaluation Method. International Electrotechnical Commission: Geneva, CH, USA, 2016.

27. Vining, W.; Hansen, C. PVAnalytics. 2021. Available online: https://github.com/pvlib/pvanalytics (accessed on 22 December 2021).

28. Malik, S.; Dassler, D.; Fröbel, J.; Schneider, J.; Ebert, M. Outdoor data evaluation of half-/full-cell modules with regard to measurement uncertainties and the application of statistical methods. In Proceedings of the 29th European Photovoltaic Solar Energy Conference and Exhibition, Amsterdam, The Netherlands, 22–26 September 2014; pp. 3269–3273.

29. Theil, H. A rank-invariant method of linear and polynomial regression analysis. *Indag. Math.* **1950**, *12*, 173.

30. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B* **2011**, *73*, 273–282. [CrossRef]

31. Geman, S.; Bienenstock, E.; Doursat, R. Neural networks and the bias/variance dilemma. *Neural Comput.* **1992**, *4*, 1–58. [CrossRef]

32. Dirnberger, D.; Blackburn, G.; Müller, B.; Reise, C. On the impact of solar spectral irradiance on the yield of different PV technologies. *Sol. Energy Mater. Sol. Cells* **2015**, *132*, 431–442. [CrossRef]

33. Eke, R.; Betts, T.R. Spectral irradiance effects on the outdoor performance of photovoltaic modules. *Renew. Sustain. Energy Rev.* **2017**, *69*, 429–434. [CrossRef]

34. Arias, A.L. R-Squared, Adjusted R-Squared, the F Test, and Multicollinearity. In *Understanding Regression Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020; pp. 185–200.

35. Jordan, D.C.; Deline, C.; Kurtz, S.R.; Kimball, G.M.; Anderson, M. Robust PV degradation methodology and application. *IEEE J. Photovoltaics* **2017**, *8*, 525–531. [CrossRef]