SMU Data Science Review

Volume 2 | Number 2 Article 2

2019

Forecasting Localized Weather-Based Photovoltaic Energy Production

Kevin Chang
Southern Methodist University, kchang@smu.edu

Afreen Siddiqui
Southern Methodist University, afreens@smu.edu

Robert Slater
Southern Methodist University, rslater@smu.edu

Forecasting Localized Weather-Based Photovoltaic Energy Production

Kevin Chang Afreen Siddiqui Dr. Robert Slater

Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
{kchang, afreets, rslater}@smu.edu

Abstract. Photovoltaic (PV) power system performance can vary from nominal specifications when put in application, making it difficult to accurately estimate real power generation at a localized level. As the usage and efficiency of PV systems has increased in recent years, the amount of power contributed to the national power grid from solar irradiation has also increased significantly. However, solar power installations are subject to variances in efficiency and output, driven by differences in system size, local weather, and atmospheric condition changes. With a significant install base in today's world, combined with extensive solar irradiance and meteorological data, the variables exist to explore the viability of a power generation forecasting model to predict PV power system performance across the United States at a localized level. Through this paper, we evaluate and analyze historical power generation from PV power systems and determine if power generation can be estimated across the United States with sufficient accuracy. Through the use of various models, a random forest regressor is found to provide the strongest model in respect to estimating PV power system energy generation.

1 Introduction

In recent years, the topic of global warming, greenhouse gases, damage to our atmosphere, and many other climate change focused topics have been discussed greatly. Growing concern has arisen from the effects of conventional power generation systems on earth's atmospheric conditions. Due to these issues, renewable energy has become the center of attention. In the United States, renewable energy has been supported by federal and state tax credits to support the adoption of renewable energy sources such as solar. According to the U.S. Energy Information Administration (EIA), in 2016 small-scale photovoltaic (PV) applications accounted for 37% of all PV generators in the US. Within the small-scale PV space, residential generators were responsible for over 52% of the total power generated, marking the first time residential sources provided the majority of small-scale power generation[1].

[1] U.S. Energy Information Administration. "More than half of small-scale photovoltaic generation comes from residential rooftops". June 1, 2017

Photovoltaic systems are semiconductor based systems that use hundreds of PV cells to form panels, and one or multiple panels will make up a system along with the inverter. When a system is installed, it is connected to the main power grid of a home in order to take power harvested from the sun's radiation energy and convert it into usable electrical power. The power created by the solar PV panels exits the system as raw direct current (DC) power. This power then passes through the inverter component of the solar PV system, and is converted to alternating current (AC) power for consumption by the homeowners and their appliances, lights, and other household items.

The energy received by the earth every day from the sun is multiples greater than the amount of energy consumed by the people each day. Although this means that solar radiation is widely abundant across the globe, energy quantity can be intermittent and energy strength can vary. This quantification of solar radiation received at the earth's surface changes daily and can vary greatly by location. In addition to the basic weather shifts due to time and location, anomalous shifts in weather conditions due to extreme events and natural disasters can add up to drastic changes in overall availability of solar radiance. Assuming a given PV system size, all of these changing factors on a daily basis can compound on one another and make it very difficult to understand how well a residential PV system will perform in generating power to specification prior to installation.

The basic issue to this problem is that solar PV system performance is not widely known and available to consumers. Also, the information around weather and solar radiation, which is not visible to the naked eye, is not a simple value to characterize and understand how much solar radiation is available for a certain home throughout the year. To begin to solve this problem, we need to find a source of historical performance for solar PV systems to know how different sized systems in various states have historically performed, but also take it down one level to zip code granularity to provide the most insights to consumers.

Looking at this problem, we are analyzing solar PV system historical performance data, weather data and solar insolation data, the amount of solar radiation received in a specified area, to build an analytical model that estimates realized power generation for a PV system of a specific size. To do this, we start with historical PV system performance data, contributed to The OpenPV Project, an initiative within the National Renewable Energy Laboratory (NREL) to create a comprehensive database of PV installations across the United States. In this data set, we are provided data points on solar PV installations such as the state and zip code of installation, install type, install date, system cost, system size, and the annualized energy production after installation of the system. In this analysis, we reduce the data to only look at the residential installations, along with their respective zip code, install date, system size, and the energy production values. After identifying the solar PV installations of interest, each sample needs to be matched with its respective weather and atmospheric data to provide explanatory variables in the same time frame as the measured installation performance.

For solar insolation data, NREL also provides a public data set called the National Solar Radiation Data Base (NSRDB), which is a collection of both meteorological and solar irradiance data for the United States. The NSRDB data provides multiple solar radiation variables, including three important values: global horizontal irradiation (GHI), diffuse horizontal irradiation (DHI), and direct normal irradiation (DNI). These attributes specifically correspond to the amount of radiation observed at a given location from different directions and through atmospheric disruptions. Global Horizontal Irradiance, the total radiation received from the sun by a surface horizontal to the ground, is a combination of DNI and DHI, making it one of the more valuable attributes of interest when discussing photovoltaic installations. The analysis will explore all values, but due to the mathematical relationship between GHI, DNI, and DHI, GHI is the value we focus on here.

For the weather data, we are using data from NASA's Prediction of Worldwide Energy Resource (POWER) project, which specifically targets solar and meteorological data to support research on renewable energy. NASA's data set provides a number of attributes related to weather and solar insolation, including air temperature, surface temperature, humidity, wind speed, precipitation, and solar insolation clearness. The basic variables of weather exist here as well as the vision that the combination of variables here will be representative of non-quantifiable values such as cloud cover. The solar insolation clearness gives us the representative value for clearness for solar radiation to pass down to the earths surface, an alternative to true solar radiation values.

With the weather and solar radiation data from each zip code, the aggregated and transformed values were extracted from the raw data to be connected with each solar PV system installation. The solar PV install month is extracted and extended to the twelve months following installation in order to get the date range for joining on the weather and solar radiation data. Merging by zip code and install month, the data is transformed to create a single data point. With this merged data, a number of regression models were then tested in order to identify the best model for estimating annual power generation for a solar PV system. The four models tested here included linear regression, elastic net linear regression, random forest regression, and a dense neural network. These models showed varying results when trying to minimize mean absolute error when estimating power generated. After evaluating each model, the random forest regressor was found to have the most powerful model, based on minimizing the mean absolute error on the training and test data sets.

Solar PV system performance is difficult to estimate and understand for consumers prior to making a significant investment in solar PV technology. In an effort to alleviate and provide more information for buyers to make an informed decision, this model provides some value, but also has some limitations. Because the solar PV energy produced by a sample system is annualized, the ability for the model to extrapolate seasonal effects to power generation is lost. Also in relation to the annualized power output, the weather and solar variables needed to match the same timeline. After aggregating the weather and data over 12

months post-installation, we found that the wider variety of values is quickly lost due to the summation or averaging of all the values to create a singular dimension of both weather and solar radiation.

2 Photovoltaic Technology

Photovoltaic systems are packaged solutions that include photovoltaic modules and other electrical components in order to transform the sun's radiation into usable electric power for consumption. The photovoltaics in a grid-connected system deliver dc power to a power conditioning unit (PCU) that converts dc to ac and sends power to the building. If the PVs supply less than the immediate demand of the building, the PCU draws supplementary power from the utility grid, so demand is always satisfied. If, at any moment, the PVs supply more power than is needed, the excess is sent back onto the grid, potentially spinning the electric meter backwards [1].

The solar PV system is made up of a number of individual components that are necessary to convert the sun's radiation into electrical power for use. The main components of a system include the solar PV panels, a power conditioning unit or inverter, battery bank, a system controller, power balancing hardware, wiring, power protection and disconnection devices[2]. Out of all these components, the solar PV panels will vary the most, coming from different manufacturer, using different technologies, and being of different sizes.

Solar PV panels are made up typically one or more PV modules, which are subsequently made up of hundreds of PV cells. At the core of a solar PV system, PV cells are semiconductor with silicon layers underneath that are designed to build an electric field to absorb and extract power from the sun's radiation electrons. When sunlight hits the PV cell surface, a subsequent current flow is triggered, producing power at a specific design voltage and outputting direct current (DC) power. The typical PV cell produces approximately 0.5-0.6 volts DC with no load, but the current depends on the efficiency and size of the PV module [2]. The typical efficiency of a commercially available PV module for installation can vary between 5-15%, which means that less than 15% of sunlight ends up being converted into electrical power [3]. While there is still significant improvement opportunities to maximize PV module efficiency, this drastically impacts the power generation from a solar PV system. Overall, when multiple modules are connected together to form a panel and/or array, the efficiency does not change, but the increase in system size allows greater power to be generated. Significant efforts have been focused in recent decades to reduce the overall cost of a solar PV system. Between 2010 and 2018, NREL reports that the cost of solar panel modules has decreased by over 80% and total system hardware costs by over 70%, making the installation of solar panels less than $1 per watt[4].

[2] Florida Solar Energy Center. "How A PV System Works". 2014

[3] U.S. Energy Information Administration. "Photovoltaics and Electricity - Energy Explained, Your Guide to Understanding Energy". April 12, 2019

3 Solar Irradiance

With regard to photovoltaics, there are four main methods of measuring solar irradiance for understanding solar power generation. One of these methods is different from the others, as it pertains to the total solar irradiance (TSI) as measured at the earth's atmosphere. This measurement is taken perpendicular to the incoming sunlight angle, and when sampled over time, is used to calculate the solar constant, the measure of mean solar irradiance over an area unit. The solar constant is estimated to be 1.366 kilowatts per square meter (kW/m^2) [4]. Although it is named a solar "constant", this value does vary ever so slightly over time depending on a number of factors in space, but the variance is extremely small at less than 0.015% of variation experienced over the past 1000 years. The minimal variation is of importance because the solar constant is based on a mean calculation of daily measurements, which experiences small variations due to the 11-year solar cycle. The change in solar constant is triggered by periodic changes in solar irradiation and solar ejection activity.

The other three types of solar irradiance measurement pertain specifically to measurements taken on the earth's surface. These measurement types consist of direct normal irradiance (DNI), diffuse horizontal irradiance (DHI), and global horizontal irradiance (GHI). Starting with direct normal irradiance, this is the measurement of solar radiation at a specific location on the earth's surface that is perpendicular to the sun [5]. This is the direct calculation of the sun's irradiance effect on the upper atmosphere of the earth minus the effects of passing through the atmosphere and other environmental losses from meteorological effects. DNI is also referred to as beam radiation since it is the direct irradiation measurement taken as if the sun is beaming directly through the earth's atmosphere and arriving at a perpendicular surface. The second type, diffuse horizontal irradiance, is the measure of radiation that arrives on the earth's surface from light scattered after passing through the earths atmosphere [5]. This measurement is focused on identifying solar radiation that is diffracted from the direct solar beam through clouds and the atmospheric sky, which is what illuminates the sky. And the last measurement of solar irradiation is the global horizontal irradiance, which is a collective measure of total irradiation taken by combining the diffuse horizontal irradiance and the direct normal irradiance after factoring in the zenith angle of the sun[5]. This total irradiance measurement is most valuable for understanding PV power systems capabilities because it encompasses the measurements of direct energy from the sun as well as the atmospheric energy that has been diffused and spread out, though still usable for harvesting energy.

When evaluating PV power systems performance, the optimal solar value to be used is the global horizontal irradiance. However, because GHI data from the NSRDB is based on a limited number of locations [6], roughly 15,000, with respect

[4] Pietro P. Altermatt. "Altermatt Lecture: The Solar Spectrum - 2.1: Measurement of the solar constant". 2019

[5] Vaisala Energy. Solar Online Tools FAQ :: Support :: 3TIER. 2019

[6] National Centers for Environmental Information. "Solar Radiation". 2019

to individual solar panel system installations, the ability to leverage hyper-local meteorological weather data to further tune the solar irradiance for a specific location will provide the individual focus necessary to ensure the analysis is done at a granularity of most value. This means that for this analysis we require zip code level weather data specific to where the solar PV system is installed.

The ability to forecast, assess and map solar PV system outputs at all levels has consistently been the subject of interest for both academic and commercial purposes. There are a number of different approaches for solar irradiance forecasting across various time scales. For example, short-term forecasting for utility scale systems are often based on satellite imaging data and traditional numerical weather predictions[2]. There are companies that provide solar energy generation results, meteorological data and weather forecasts that can be useful, however, this is too expensive for the small PV systems that are used for residential purposes and can only be useful at a commercial level. The size of solar panel installed is directly proportional to the energy produced by the PV plant. As the size increases, the overall solar output increases, but the cost of installation and the type of install also needs to be considered. The meteorological conditions, the cloud coverage and the movement of clouds, all have an impact on the output generation of the PV plants. Various data science teams who have attempted to tackle this problem have used a variety of statistical models to predict solar radiation values using methods like an auto-regressive integrated moving average (ARIMA) model which leverages the power of historical time series data but the drawback with time series data is it cannot leverage the non-linear features like meteorological and cloud moments efficiently and therefore different Machine learning models like artificial neural networks (ANN) and support vector machine (SVM) models have been used to forecast the global and horizon solar irradiance to estimate power generation for solar PV systems[3].

4 Solar Radiation Data

In the analysis provided herein, the data used is based on a number of sources focused on providing data supporting solar power research. The first data set is from NREL's National Solar Radiation Database (NSRDB). In this data set, NREL compiles a collection of solar and meteorological data from the United States. The data granularity is based on either half-hour or hourly intervals, providing excellent resolution for conducting research and analysis studies. Variables in the NSRDB include common measurements for solar irradiation along with extra meteorological attributes which are not used. Unfortunately, the data which NSRDB provides is sourced from solar measurement stations that are spaced widely around the country. While this represents accurate measurements for specific locations, the accuracy between stations begins to drop off as data points between measurement stations is inferred. Although this data set does not provide measurements respective of each individual solar panel installation location, solar irradiation is not expected to vary significantly over small distances. Solar radiation will impact a region with very similar energy, whereas

Year	Month	Day	Hour	Minute	GHI	DHI	DNI
2015	1	1	8	30	100	29	600
2015	1	1	9	30	240	46	796
2015	1	1	10	30	349	57	875
2015	1	1	11	30	410	62	911

Table 1. Example solar insolation data from the National Solar Radiation Database

local weather attributes such as cloud cover, temperature, humidity, rainfall, and other meteorological attributes will affect overall visibility and strength of solar irradiation that reaches the solar PV panel surface. The combination of these meteorological conditions with specific solar radiation parameters like GHI, DHI, DNI, and solar zenith angle will provide the base input solar energy data of the data model.

With the solar radiation data, because we are estimating annualized power generation, the solar radiation data needed to be transformed into an equivalent time scale as the training data. The raw data format can be seen in Table 1, showing that for a given zip code, the solar irradiance data is sampled at an hourly basis on every half-hour. Since solar radiation is cumulative from hour to hour and day to day, the data transformation is a summation of all hours and days for a given month. This provides the total monthly solar radiation available at a specific zip code. The monthly solar radiation is accumulated based on the twelve month window post-installation of the solar PV system starting from the installation month.

After aggregation, we compared the solar radiation measured in GHI, DHI, and DNI in Figure 1 to see how the three measurements are related or differ from one another. We found that the amplitude difference between GHI and DHI is significant, where DHI is a much flatter trend from month to month than GHI, providing less variance, whereas DNI is significantly different in trend than the horizontal irradiance measurements. Comparing two states, Oregon and Wisconsin, in Figure 2, we see that seasonal transition months like March and October are similar, but in various summer and winter months, Oregon encounters significantly more solar radiation than Wisconsin, potentially providing more success with solar PV systems.

5 Meteorological Data

Adding to the solar irradiation data available from the NSRDB, localized weather data is needed in order to analyze the solar radiation effects on a specific location. In order to add local adjustments to the solar radiation data, detailed meteorological data is provided by NASA's POWER (Prediction of Worldwide Energy Resource) project. NASA provides an enhanced renewable energy data set augmented by new satellite systems. One of the key communities NASA

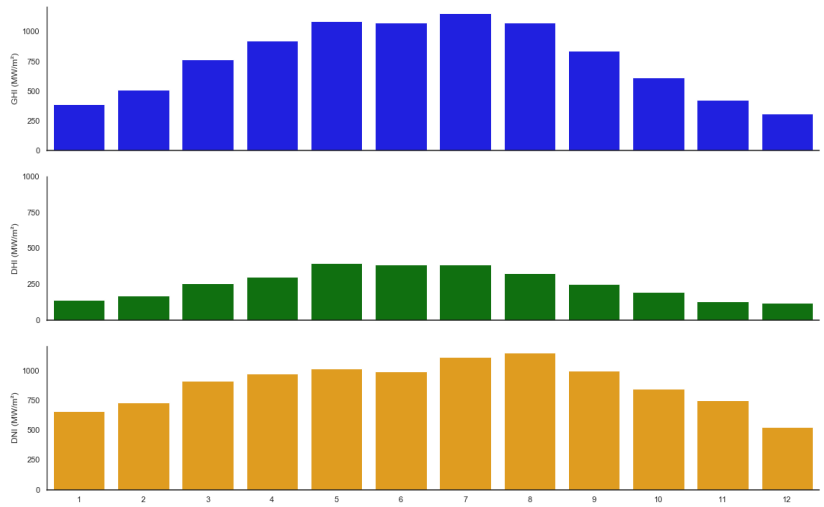

Fig. 1. Monthly solar irradiance in 2015 from the National Solar Radiation Database

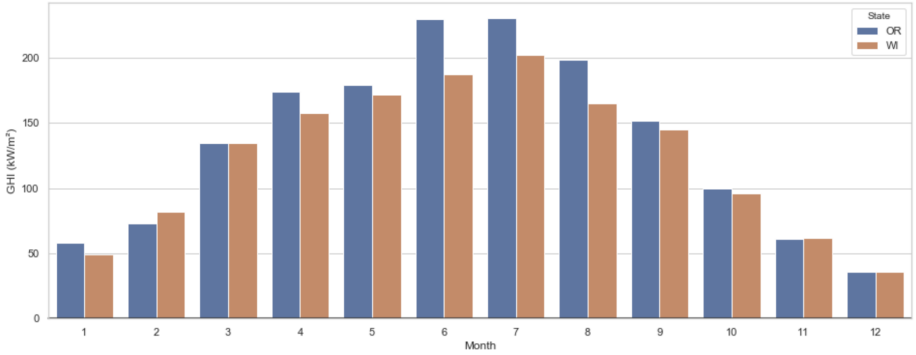

Fig. 2. Monthly Solar Irradiance in 2015 in Oregon and Wisconsin from the National Solar Radiation Database

DATE	KT	PRECTOT	PS	QV2M	T2M	TS	WS2M
201501	0.42	91.88	98.63	0.001950	-7.95	-8.26	0.47
201502	0.45	53.92	98.51	0.001455	-12.58	-13.00	0.61
201503	0.46	56.74	98.65	0.002713	-3.10	-3.33	0.43
201504	0.45	69.97	98.40	0.004868	6.37	6.28	0.14

Table 2. Example weather attribute data from NASA POWER Project

targets with this data set is providing parameters specifically tailored to helping design renewable energy systems. Through the NASA POWER API, data is available for specific latitude and longitude coordinates over daily and interannual aggregations. From this data, meteorological attributes are gathered that have distinct factors related to solar PV system installations. Beyond just traditional meteorological attributes such as temperature, wind speed, and humidity, these variables also have counterparts relative to altitude. Each meteorological attribute also includes a measurement at 2 meters and/or 5 meters above surface level that relates more directly to the application of solar PV systems. When solar panels are installed on the roof of a home, the temperature, humidity, and other weather variables begin to shift a bit due to the altitude change. NASA's POWER database provides an altitude-adjusted value to provide the most relevant measurements for planning renewable energy systems. These adjusted values provide a more accurate measurement as to what the solar panels physically will experience further away from ground level as compared to measurements made at ground level as temperature and humidity levels can vary at different levels of altitude.

The weather data provides some visibility to understand similarities and differences between various locations across the nation. In Table 2, a sample output of the information provided by the NASA POWER data set can be seen, providing values such as temperature, precipitation, solar irradiance, and wind speed. Although geographically cities and locales can be far away from each other, some areas still encounter similar weather trends. For example, when comparing Oregon and Wisconsin, although both are on the northern half of the United States, their weather can differ quite drastically. Wisconsin is known to have some cold winters with heavy snowfall, as Figure 3 shows in January and February where the average temperature is significantly lower in Wisconsin than in Oregon. Although precipitation is expected in the winter time due to snowfall, Figure 4 shows us different, with Wisconsin having significant rainfall throughout the summer as well, compared to Oregon. Also when it comes down to just the average wind speed between the two states, Wisconsin has a significantly higher average wind speed than Oregon as Figure 5 shows, an attribute that could affect solar energy production due to disruptions from cloud cover and dust.

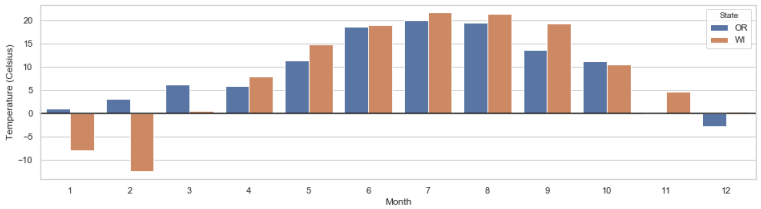

Fig. 3. Monthly Temperature in 2015 in Oregon and Wisconsin from the NASA POWER Database

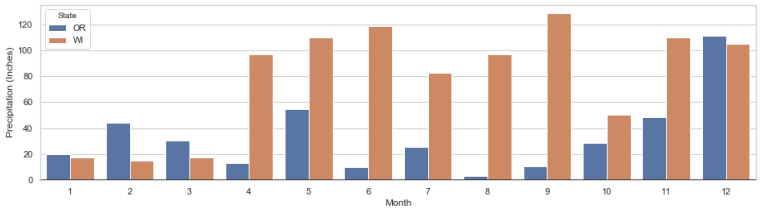

Fig. 4. Monthly Precipitation in 2015 in Oregon and Wisconsin from the NASA POWER Database

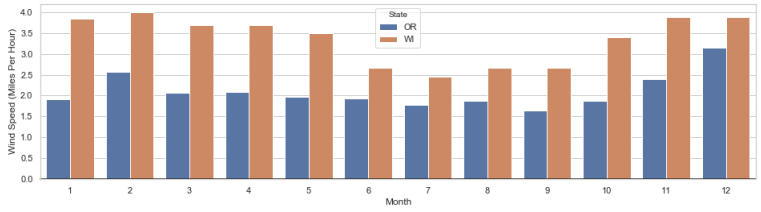

Fig. 5. Wind Speed in 2015 in Oregon and Wisconsin from the NASA POWER Database

6 Historical PV System Performance Data

The next data source provides actual solar panel system data by individual installation. This data is sourced from the National Renewable Energy Laboratory (NREL), from their OpenPV Project. The OpenPV Project is a NREL effort supported by the government, solar industry, and the public, to compile a comprehensive public data set of photovoltaic installations across the United States. The data here is community and publicly sourced based on voluntarily provided information from disparate sources. Due to the nature and method of data collection used for this data set, there are potential risks to data quality, primarily that the data provided is only as accurate as the contributors of the data. Fortunately for the industry of photovoltaic installations, installers and vendors are diligent in helping provide to data sources like this as well as more reputable sources, such as governmental entities and large utilities. This support provides a significant amount of data through well trusted means and methods, providing much more stability and accuracy for predictive modeling.

state	date_installed	size_kw	zipcode	install_type	city	annual_PV_prod	tech_1	tilt1	cost
TX	12/03/2015	13.550	78266	residential	San Antonio	20178.126730	Mono	N/A	42779.60
TX	06/26/2015	5.490	78414	residential	Corpus Christi	7484.855032	Poly	N/A	20659.99
OR	12/18/2015	3.180	97734	residential	Culver	4369.586682	Poly	23.0	13801.20
OR	10/01/2015	13.550	97701	residential	Bend	18898.993170	Mono	14.0	41997.60

Table 3. Example solar PV system installation data from The OpenPV Project

Within the data shown in Table 3, we have important location identifier attributes such as installation city, state and zip code. For each installation record, we also have specific installation information, such as the system install date, system size (in kW), install type-residential, commercial, or utility, technology-the PV technology installed, the system tilt-the position angle of the PV panel surface, and the cost of the system. Finally, this data set includes the value used as the target variable in this analysis, the solar PV system's annualized energy generation. Unfortunately, the energy production numbers are a singular value, eliminating the ability for a higher resolution model to be conducted. Without at minimum a monthly distribution of energy production, the accuracy of the model could be limited as seasonal changes in different parts of the nation will not be apparent when matched to weather data.

In Figure 6, the annual photovoltaic energy production for all residential installations in the top 10 states in the United States are shown. The results are ordered descending based on total PV production. Based on the distribution, we find that the majority of PV solar installations are located in California, followed by Massachusetts and New Jersey. The data includes sufficient samples from a number of states, but is limited to 14 states with sufficient samples sizes to be considered for inclusion in the analysis.

As we take an alternative look at the Open PV Project data in Figure 7, we see a significantly increasing trend of solar PV installations starting in the

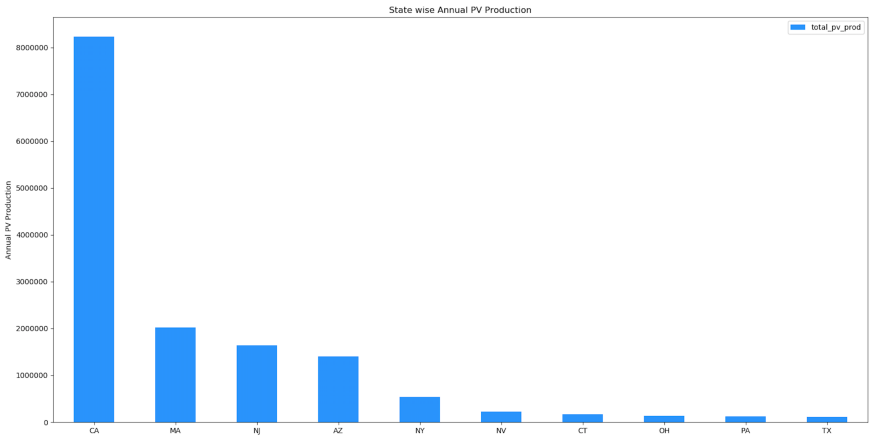

Fig. 6. Annual Solar PV Production By State in the U.S. Residential Sector From 2000 to 2019 based on submitted data to The OpenPV Project

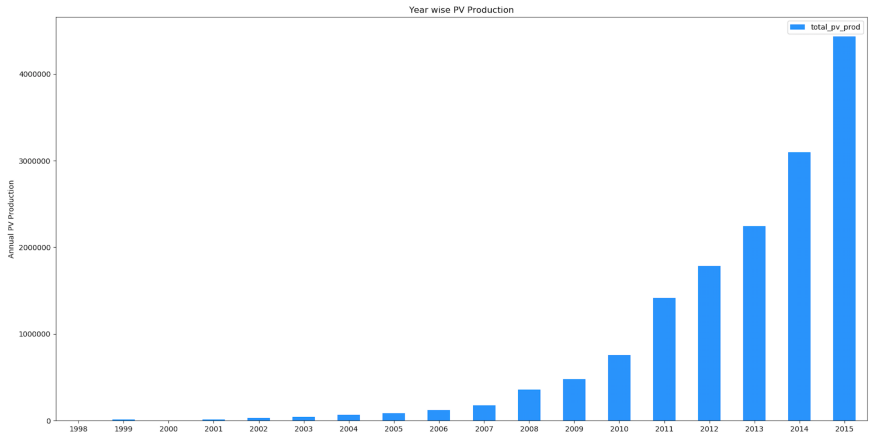

Fig. 7. Annual Solar PV Production in the U.S. Residential Sector From 2000 to 2019 based on submitted data to The OpenPV Project

early 2000s up to 2015. Compared on industry trends, the OpenPV Project data set contains a sample quantity gap after year 2015, which does not accurately represent solar PV installations from 2016 and onwards, so the data sample is limited only to year 2015. The data from the OpenPV Project includes the independent or response variable, the annual energy produced by the PV system.

7 Model Approach

In building a model for predicting annualized power generation from a solar PV system, since the predictors and the response variable are continuous variables, regression models are the best fit model type for this case. Four regression models were fitted in order to determine best fit with a focus towards minimizing mean absolute error (MAE) and root mean squared error (RMSE), to identify the most accurate predictive model. The four models that were fitted included linear regression, elastic net regularization linear regression, random forest regressor, and a dense neural network using Keras.

7.1 Linear Regression

Starting with the simplest form of regression, a linear regression model is used to identify relationships between continuous variables. In a linear regression model, there are two types, simple and multiple, which is based on the quantity of independent variables provided to train the model. The model looks for a statistical relationship between one or multiple independent variables and their explanatory power to estimate the response variable. We fitted the model to determine if there is a linear relationship between the independent variables, a combination of weather, solar irradiance, and system size, and the dependent variable, the real power generation of the system as a whole.

7.2 Elastic Net Regularization

Following the multiple linear regression model, we extend this model to implement elastic net regularization to look for improved model fit and resulting metrics. The implementation of elastic net regularization is used to combine both lasso (L1) and ridge (L2) regularization methods into a single model to solve for each individual method's limitations. Since regularization is a defined methodology to prevent overfitting the regression model to training data, allowing for the two different penalties to work together to build an optimial linear regression model. With lasso regularization, the L1 penalty is equal to the magnitude of the coefficients to limit their size. With ridge regularization, the L2 penalty is equal to the square of the magnitude, reducing the size of all coefficients. The addition of regularization helps to manage the fact that there is a significantly higher number of solar PV installations in the training data with a lower system size and resulting power generation than larger systems.

7.3 Random Forest Regressor

The third model switches into a different form of regression model, a random forest regressor, which is a form of regression model focused on low bias but high variance. While linear regression models are the opposite, with high bias and low variance, the random forest regressor focuses on providing adequate results with variability based on simple decisions. Since a random forest model is based on multiple decision trees to make their decisions, although the resulting values are continuous in nature as opposed to categorical, the continuous values have limitations. The resulting split of a tree provides two values as opposed to a calculated value, which can limit the resulting output accuracy, but in this case, the specific accuracy does not need to be 100% accurate, but within a reasonable range of true power generation. The random forest regressor is good for reducing bias towards smaller or larger solar PV systems, but allowing variability to exist in the resulting estimations.

7.4 Dense Neural Network

The last model implemented was a dense neural network using Keras and Tensorflow. A dense neural network is characteristic of having fully connected layers where all of the inputs and outputs are connected to all neurons of a layer. In building this neural network, we leveraged Keras' high level API's to build the model on a Tensorflow backend. In our implementation we used the Keras regressor from the python scikit learn package with a mean squared error (MSE) loss function and three activation functions, relu, sigmoid, and linear.

8 Results

From our model outputs we found that if evaluating R^2 scores, the models performed quite well. Unfortunately, in this scenario, we believe the R^2 value is a false sense of accuracy. While the R^2 value provides a method of comparing how accurate we are in estimating the amount of solar power generated, the estimation could be off by 5% for a 20kW system and that would mean our estimation is off by 1kW, a significant amount of power to be generated for consumption that would greatly impact the value gained from a solar PV system. For the most accurate representation of accuracy, the mean absolute error (MAE) is used because it tells us exactly how far from actual measurements our estimations were. In contrast to the MAE, the mean absolute percent error (MAPE) provides us a representative accuracy for different PV system sizes. These values help to calculate how much of a power generation gap would be missed by our estimation model when compared to real world output. Table 4 shows the four models and the estimation performance for each.

In Table 5, we find that the random forest regressor identifies as the model with the lowest MAE. With this model, we find that the average error of the samples is approximately 424 watts, but this is an average across various system

	Linear Regression	Elastic Net	Random Forest	Dense Neural Net
R^2	0.963	0.961	0.978	0.000
MAE	0.593	0.597	0.424	3.2636
MAPE	7.60%	7.65%	5.03%	44.9602
RMSE	0.840	0.853	0.648	19.5165

Table 4. Estimation Model Performance Metrics

size_kw	mean absolute error	mean absolute percent error
1	0.094	9.42%
2	0.187	9.36%
3	0.246	8.19%
4	0.319	7.98%
5	0.378	7.56%
6	0.437	7.29%
7	0.508	7.26%
8	0.601	7.51%
9	0.620	6.89%
10	0.690	6.90%
11	0.724	6.58%
12	0.828	6.90%
13	0.847	6.51%
14	0.907	6.48%
15	0.847	5.65%
16	0.955	5.97%
17	0.994	5.85%
18	1.139	6.32%
19	0.973	5.12%
20	0.726	3.63%

Table 5. Random Forest Model MAE and MAPE By System Size

sizes. The assumption is that with different system sizes, the mean absolute error will vary, but the mean absolute percent error will be more consistent. The dense neural network significantly under-performed and obviously does not fit this use case well, resulting in zero accuracy power and a MAE of 3.2636 kW, larger than the majority of solar PV system sizes. The linear regression and elastic net models performed very similarly, so the penalty functions did not help to significantly improve the estimation accuracy.

If we look at the mean absolute error across various system sizes for the random forest model in Table 5, it is obvious that as the system size grows, the error grows. From the MAE by system size, we are able to calculate the mean absolute percent error (MAPE), which shows how large the error is relative to the size of the solar PV system. From this we find that as the system size grows, the percent error actually decreases, meaning that we are more accurate on larger systems than we are on smaller systems.

In a detailed analysis of forecasting energy production of a solar PV system at the University of Belgrade, faculty members Marko Ikić, Jovan Mikulović, and Željko Đurišić conducted a study comparing two types of models to estimate PV energy production for a given system. Two models, including a "clear day" model and a "realistic" model were developed to estimate the power output on the roof of the engineering building. In their analysis, it was found that using purely a clearness index of the location was not enough to provide a valuable estimation, while the realistic model which leveraged 10 minute interval horizontal irradiance measurements and ambient temperature provided a model that resulted in an error of -0.47%, providing a very accurate estimation.

9 Ethics

The ethical issues and impact presented by the data and the application of statistical models will be discussed in terms of the ethical requirement to reduce the impact of any form of bias towards class, wealth, or means. Three firm issues that drive an impact to ethics when evaluating this analysis of solar power systems focus on home-ownership rates, upfront costs, and finally government tax credit value. In order to determine if the features which have been used have a bias towards discrimination, we must evaluate these variables and determine if they can be eliminated from the data without significantly impacting the power of the model when predicting power generation.

9.1 Homeownership Rates

The first issue that is brought to light in the context of this analysis is the most encompassing, the bias towards citizens of means who are able to afford solar panel systems for their home because they own the home as opposed to rent. According to the Federal Reserve Bank of St. Louis, at the beginning of 2015, home-ownership rate was at a historical high of 69.1% and has dropped to a

recent rate of 64.2% [7]. With these numbers in today's world where less of the younger generation is focused on home ownership, the bias towards those who own homes is shifting and may not represent the same bias towards financially challenged individuals and families that it may have a decade ago.

9.2 PV System Costs

A secondary issue stems from the fact that a solar power system is a luxury item, that even for many home-owners is an unnecessary cost and upkeep financially to support. Because nationwide power is a basic utility of life and even in some states deregulated for fair competition, Identifying areas that are focused on the basis of existing PV system installations tend to bias towards neighborhoods and areas where there are more financially healthy individuals and families with disposable income to afford a PV system. Based on the January 2019 average cost per watt for a solar power system of $3.05 and the average size of an installation of 6kW, the average estimated cost for a solar PV system installation equates to $18,300 [8]. This is the real system cost taken into account as the impact of tax credits are not realized immediately but only after receiving a tax return the following year. Therefore, the true out-of-pocket cost for a solar panel installation is not an insignificant amount and can easily be considered a major purchase for a typical homeowner.

9.3 Federal Tax Credits

Finally, continuing on the topic of government tax credits, there are specific considerations that one needs to think through before taking these tax credits into consideration as a positive sign to move forward with the installation of a PV system. Government tax credits are significant in order to try and drive adoption of solar. What most persons don't realize about tax credits is that in order to get the credits, federal taxes must be paid in order for these significant credits to apply. In 2018, the estimated percentage of households who will have no tax liability is roughly 45% [9]. This means that for most, the inclusion of the flaunted tax credit will have no impact to the installation of their PV system and is often incorrectly advertised to all households despite the fact that almost half of the country would see zero dollars of the credit when tax season comes around. This can have a significant impact to the choice of installing a solar PV system as well as the time to gain a return on investment (ROI) as the cost will significantly change and the annual savings will take longer to cover the overall cost.

[7] Federal Reserve Bank of St. Louis - FRED Economic Data. "Homeownership Rate for the United States". June 12, 2019

[8] EnergySage. "How much do solar panels cost in the U.S. in 2019?". June 12, 2019

[9] CNN Money. "45% of tax filers will owe nothing in federal income taxes this year". June 12, 2019

10 Conclusions

When considering the power generation capability of a solar PV system, a number of different attributes can contribute to its performance capabilities. From this analysis, we find the estimation of power generation for a solar PV system can be estimated on an annual basis with relatively good accuracy and small errors. On this annualized scale, we do lose a significant amount of resolution to seasonality and other variances over time that can shift and change when evaluating small local areas such as zip codes. Unfortunately, without higher granularity of data in the power generation data set, this model is limited in its predictive power. Although supporting data for weather and solar irradiance are available in very fine granularity, we require a significant amount of historical PV system power generation performance broken down in monthly intervals to improve the model performance.

References

1. *Renewable and Efficient Electric Power Systems*. A John Wiley and Sons, Inc., Publication, 2004.
2. Mikael Karpe Conde Emil Isaksson. Solar power forecasting with machine learning techniques, 2018.
3. Han Seung Jang, Kuk Yeol Bae, Hongshik Park, and Dan Sung. Solar power prediction based on satellite images and support vector machine. *IEEE Transactions on Sustainable Energy*, 7:1255–1263, 07 2016.
4. David Feldman Ran Fu and Robert Margolis. U.s. solar photovoltaic system cost benchmark: Q1 2018.