CS 760: Machine Learning Fall 2020

Predicting the impact of climate change on solar PV generation

Authors: Samantha Hartke, Nomin Khishigsuren, and Michael Rebarchik

1 Abstract

To mitigate the effects of climate change, alternative renewable energy sources such as solar photovoltaics (PV) are being adopted to replace traditional electricity generation methods on a global scale. With a growing dependence on solar PV plants, the development of accurate solar energy prediction models will become essential for the reliable economic operation of the electric grid. Though there are many types of solar energy prediction, we focus on predicting the monthly efficiency of solar PV plants as a function of geographical and climatological variables. With the global climate predicted to change in response to increased levels of greenhouse gas emissions, we also seek to understand how climatological changes, such as temperature and cloud cover, will impact solar power. In this project, we first train linear and logistic regression models using climatological records along with historical data from 144 solar PV plants across the United States. We evaluate each model's ability to predict monthly solar PV efficiency. Then, using climatological data from regional climate models to represent future climate conditions, we predict the efficiency of solar PV plants in the future. This analysis allows us to understand the impact of geographical and climate features on monthly solar PV efficiency. Our findings provide insight into how a changing climate may affect the solar PV industry.

2 Introduction

With greenhouse gas emissions continuing to rise, the transition to more sustainable energy sources is essential for mitigating the consequences of climate change. Renewable energy technologies such as solar PV make use of renewable resources (direct solar radiation) to generate electricity. While alternative energy sources are promising for reducing the carbon footprint, they depend on the regional availability of natural resources, and thus are not suitable for all locations. For instance, solar PV generates the most energy in sunny, dry locations while wind turbines have higher energy generation rates in windy areas, such as coastal locations.

The integration of solar power into the electric grid continues to increase on a global scale. This increasing dependence on solar power has led to newfound challenges for the reliable and economic operation of the power grid. The high variability and uncertainty in solar irradiance leads to intermittency in the energy supplied by solar PV plants. To effectively match load demands and avoid the generation of excess energy, solar forecasting has become an essential component of solar power generation. The continued development of accurate solar forecasting is imperative for reliable electricity system operation. 'Solar forecasting' generally refers to predictions of solar irradiance or solar PV generation on the scale of hours and days, and is useful for operating solar PV systems.

For potential investors or communities, the longer term predictability of solar PV generation is also important. Before investing in a new solar PV array, a client may want to know how much energy they can expect to be generated on a monthly basis. Analytical models such as the System Advisor Model combine observed direct shortwave radiation, expected conversion efficiency of solar PV modules, and the number of PV modules to predict energy generation for a given solar PV array (Gilman, 2016). However, these models can require significant assumptions and oversimplification of module properties. Additionally, solar PV energy generation can be influenced by a number of factors related to geography and climate that analytical models may not take into account.

Climatological and geographical variables that influence solar PV generation include incoming shortwave radiation (also referred to as solar irradiance), cloud cover, temperature, wind speed, elevation, and latitude and longitude (which are sometimes used as a proxy for shortwave radiation). As climate change is predicted to become more prevalent in the next few decades. understanding the implications of these changes on solar PV production will be critical in the continued success and advancement of the technology. While there is uncertainty surrounding the impact of climate change on many sectors, the impact of climate change on solar PV generation is particularly relevant as many regions work to transition to a more carbon neutral electricity grid. The feedback loop between the adoption of solar PV and climate change impacts on solar PV is poorly understood. While regional climate models (RCMs) predict increasing temperature and solar radiation, they also predict increased cloud cover in some regions.

Since we want our prediction of solar PV generation to be applicable to current or prospective solar PV plants of any size, we have decided to predict monthly solar PV efficiency. We define efficiency as the monthly energy generated by a plant [in MWh] divided by the maximum possible energy generation of that plant [in MWh, calculated using nameplate capacity of the plant multiplied by hours in a given month]. Within the energy analysis field, this efficiency metric is known as the capacity factor of a plant and describes the proportion of energy generated by the plant relative to the theoretical generation of the plant if it was operating at nameplate capacity continuously. For many renewable energy technologies, intermittency of renewable energy sources (e.g. solar radiation and wind) does not allow for high capacity factors. In general, wind turbines have capacity factors between 20% and 40% and solar PV systems exhibit capacity factors as low as 10% (this would occur at high latitudes during winter when solar radiation is minimal) or as high as 30% (Boretti, 2020). Other energy generation technologies have higher capacity factors; nuclear power plants, for instance, regularly operate at capacity factors of 90%.

The goal of this project is to develop a machine learning model to predict monthly PV efficiency of solar PV installations in the U.S. By analyzing the significance of various predictive features, we will better understand which variables are important for identifying new locations that may be promising for solar PV. Finally, we aim to understand how changes in regional climate may influence the production efficiency of current and prospective solar PV plants.

3 Related Work

Renewable energy forecasting has become increasingly important as renewable technologies generate greater and greater proportions of the electricity in many countries' electric grids. Forecasting wind and solar energy in the short term is important for letting grid operators know when to request greater generation from conventional power sources or for preparing the grid for a large influx of energy on particularly windy or sunny days. These requests to conventional power plants, referred to as re-dispatches, cost grid operators money because utility companies must be compensated for such last minute adjustments. In Germany, which aims to generate 80% of national electricity demand with renewables by 2050, considerable resources are being invested to create models that can reliably forecast renewable energy generation 48 hours in advance (Schiermeier, 2016). In the vein of long-term renewable energy forecasting, a number of studies have sought to understand or constrain the effects of climate change on PV (Panagea et al., 2014); to our understanding, none of these studies utilized machine learning techniques.

In recent years, forecasting solar irradiance using regression algorithms has become common practice (Lorenz, 2009 and Sharma, 2011). However, the high variability in climate data limits the accuracy of such models. Alternative methods such as artificial neural networks have shown improved performance (Prastawa, 2013). For example, Jawaid et al. performed a comparative analysis of various regression algorithms against artificial neural networks for predictive forecasting of solar irradiance (Jawaid, 2020). They find that the inclusion of key parameters such as azimuth and zenith parameters significantly improve model performance.

4 Dataset

Monthly energy generation data from solar photovoltaic plants around the U.S. was retrieved from the U.S. Energy Information Administration (EIA). Each record of monthly solar PV generation is considered one sample data point. Our final dataset includes generation data from 144 plants across 19 states (Figure 1). The EIA dataset for each solar PV plant includes plant nameplate capacity (e.g. 5 MW), latitude, longitude, and start date of operation.

The nameplate capacity is the intended-full load capacity of a plant, and in our dataset it ranges from less than 1 MW to 52 MW. Our dataset is skewed towards solar PV plants with nameplate capacities less than 10 MW. One deficiency of our dataset is the lack of solar PV plants in the midwest and northern U.S.; having sample data from these geographical regions could help our machine learning models be more robust for predicting solar PV efficiency throughout the country.

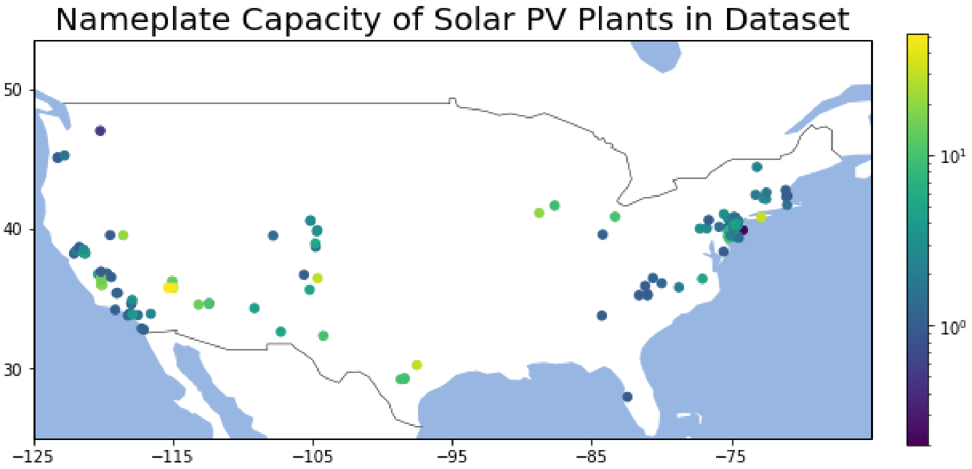

Figure 1: Regional map of PV plants used in this study obtained from the EIA repository. The color indicates the nameplate capacity [MW] of each plant. Note the log scale of the color bar.

For each record of monthly generated energy at a solar PV plant, we obtained monthly temperature data, including average temperature, minimum temperature, and maximum temperature at the corresponding latitude, longitude, month, and year from Meteostat's Python library. Monthly climatological data, including cloud cover in the lower, middle, and upper atmosphere, incoming shortwave radiation, and outgoing longwave radiation, was obtained from EUMETSAT's Satellite Application Facility on Climate Monitoring reanalysis product. Our final dataset consists of 13,977 samples of monthly data. Table 1 shows 5 such samples. Note that the efficiency variable was not retrieved from a database, but was calculated using nameplate capacity and generated energy (see Approach section).

Table 1: Dataset sample. Plant code is used to identify the solar PV plant from which each data sample was obtained. Efficiency is the predicted variable, y, and all other features are predictors.

Plant Code	57310	53710	53710	56900	56900
Efficiency (%)	0.06	0.19	0.19	0.20	0.20
Age (months)	10	32	78	92	116
Latitude (°)	33.79	33.79	33.79	39.36	39.36
Longitude (°)	-118.24	-118.24	-118.24	-74.44	-74.44
Avg. Temp. (°C)	14.9	22.4	22.7	21.8	20.9
Min. Temp. (°C)	6.7	13.9	17.2	10.0	9.4
Max. Temp. (°C)	30.0	36.7	31.7	35.0	30.6
Avg. Windspeed (mph)	5.6	7.7	9.1	12.1	12.2
Cloud Fraction - Low (%)	17	11	11	8	11
Cloud Fraction - Mid (%)	9	3	8	4	10
Cloud Fraction - High (%)	18	4	5	35	26
Incoming Shortwave Radiation (W/m^2)	133	259	315	193	187
Outgoing Longwave Radiation (W/m^2)	374	453	472	439	428

We used the CMIP6 regional climate model (RCM) to obtain the temperature, radiation, and cloud cover conditions that could be expected under climate change (World Climate Research Programme, n.d.). Although uncertainty in climate model predictions is a nontrivial factor of RCM data, we feel that the atmospheric conditions under each RCM scenario are a possible representation of the future and are important to evaluate when making decisions about energy infrastructure. We refer to the monthly climatological data obtained from the RCM as the climate change dataset. Similarly to the feature vectors used to evaluate students' chances of surviving the titanic in assignments this semester, the climate change dataset is not used to train or evaluate our machine learning models. Rather, we use the machine learning model to make predictions about the monthly solar PV efficiency under future climate change conditions. We obtained climate change data representative of the year 2050 near Palm Springs, CA and Chicago, IL, where plants 57743 and 57191 from our dataset are located, respectively.

5 Approach

Using the acquired datasets, linear and logistic regressions were performed to identify significant climatological features that influence solar PV efficiency and develop a predictive model for solar PV performance. To calculate monthly power efficiency, the nameplate capacity was converted from an annual to a monthly basis. Based on monthly power generation, the monthly efficiency was calculated using Eq. 1.

$$Efficiency = \frac{Power\ Generation\ [MWh]}{(Nameplate\ Capacity\ [MW])\ *\ (Days\ in\ Month)\ *\ (24\ Hours)} \tag{1}$$

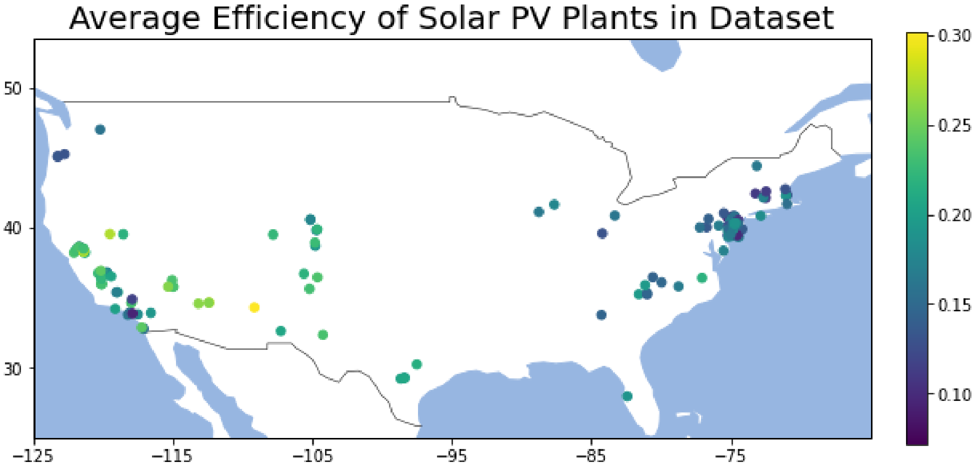

Figure 2: Average efficiency of all solar PV plants in our dataset. While the maximum efficiency is 1.0, most solar PV installations attain efficiencies closer to 0.2 due to the intermittent availability of shortwave radiation. Although this figure displays the average monthly efficiency of each plant, monthly efficiency fluctuates on an annual cycle; efficiency is lowest during the winter months and highest during the summer months.

Linear regression was performed using linear models available from the sklearn package through Python. These models include Ordinary Least Squares (OLS), Ridge, Lasso and Elastic Net models. Ridge model uses L2-norm regularization of the coefficients, while Lasso uses L1-norm regularization instead. On the other hand, Elastic Net model uses both L1 and L2-norms for regularizing the coefficients. The coefficient of determination, R^2, better known as the R^2 score, was calculated for each model using Eq. 2. Subsequently, 10-fold cross validation was carried out on the estimator using the built-in cross validation helper function . The accuracy of the models from 10-fold cross validation was similarly quantified using the R^2 metric.

If $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value for total n samples, the estimated

$\mathbf{R^2}$ is defined as:

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

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$.

The comparison of the resulting values is shown in Table 2.

Table 2: R^2 scores of four linear models and their 10-fold cross validation

Linear Model	**R^2 score**	**10-fold CV R^2 score**
OLS	0.510	0.457
Ridge	0.509	0.458
Lasso	0.505	0.468
Elastic Net	0.450	0.426

Based on this assessment, Ordinary Least Squares (OLS) method was chosen as the linear regression model for our analysis. The OLS linear regression model fits a linear model by minimizing the Residual Sum of Squares (RSS) between the true and predicted targets y. The coefficient vector $\hat{\theta}$ is then defined as shown in Eq.3.

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \tag{3}$$

where $\boldsymbol{X}$ is the training dataset that comprises of feature vectors.

A logistic regression was performed using the sklearn package available through python. Our predicted variable (solar efficiency) was transformed into a binary variable based on a minimum solar efficiency (i.e. Efficiency < 20%, y=0; else y=1). A benchmark of 20% efficiency was used as this was the average annual efficiency of most plants (Figure 2). L2-norm penalization was used alongside the saga algorithm to compute the maximum likelihood. Maximum likelihood asymptotics were used to evaluate feature significance by computing p-values.

In order to calibrate and evaluate our regression models, a 10-fold cross validation was performed for each model. The data for all 144 plants was split into training and validation sets based on plant identity. All monthly data from each plant was assigned to the corresponding training or validation set. In this way we avoid having data from a single plant be present in both training and validation data. Using the same methodologies described previously for linear and logistic regression, fits were determined for each training set. Using the corresponding test sets, the accuracy of each model was computed.

6 Results

The significant features under the linear regression model we used were year, month, latitude, longitude, average temperature, maximum temperature, wind speed, medium and high atmosphere cloud cover, incoming/outgoing radiation and age of the power plant were all found to be significant. Using 10-fold cross validation with plant identity splitting, the linear regression model had an accuracy of 43%. Significant features along with their corresponding p-values are provided in Table 3.

To determine how the linear regression model performed in predicting the efficiency of the solar power plants, the dataset was split into training and testing sets and the root-mean-square error (RMSE) was calculated between the true and predicted efficiencies. The resulting error was +/- 0.07 in terms of the efficiency.

Similarly, logistic regression was used to model our dataset. Significant features identified by both the linear and logistic models along with their corresponding p-values are provided in Table 3. Similar to our findings using linear regression, the month, latitude, longitude, mid-atmosphere cloud cover, incoming/outgoing radiation, and age were

Table 3: Significant features identified by linear and logistic regressions along with their corresponding p-values. Blank entries indicate that a feature was not identified as significant.

Feature	**p-value (Linear)**	**p-value (Logistic)**
Nameplate Capacity	-	0.00
Year	0.00	-
Month	7.00×10^{-4}	0.00
Latitude	0.00	8.00×10^{-3}
Longitude	8.16×10^{-6}	0.00
Avg. Temp.	0.00	-
Max. Temp.	1.29×10^{-6}	-
Wind Direction	-	3.00×10^{-3}
Wind Speed	$3.63\ \times10^{-5}$	-
Cloud Fraction - Low	-	0.00
Cloud Fraction - Medium	0.00	3.00×10^{-3}
Cloud Fraction - High	1.90×10^{-2}	-
Incoming Shortwave Radiation	1.50×10^{-3}	0.00
Outgoing Longwave Radiation	0.00	3.80×10^{-2}
Age	5.00×10^{-3}	03.00×10^{-2}

all found to be significant. Using 10-fold cross validation, our logistic regression was identified to have an accuracy of 79%.

Through linear regression, most variables were determined to be significant whereas a more limited subset was obtained through logistic regression. The increased accuracy of the logistic regression is believed to be primarily due to the transformation of the predicted variable (solar efficiency) into a binary variable. Additional testing was performed to determine the importance of the binary transformation. Regardless of the efficiency cutoff chosen, an accuracy of approximately 80% was achieved in all cases. In agreement with previous studies, we find that regression algorithms fail to accurately model solar power predictions due to high variability in climatological data.

The significant features shared between the two methods are consistent with *a priori* expectations. For example, the month, latitude, cloud cover, and incoming/outgoing radiation should all directly related to the incoming solar flux. Longitude is believed to be significant primarily due to geographical features (i.e. proximity to coast). This feature is expected to become less significant if a more comprehensive dataset was included. Surprisingly, the nameplate capacity was determined to be significant by the logistic regression model. This may support our assessment that the dataset is skewed towards plants of lower capacity. It may also indicate that plants with greater nameplate capacities benefit from some sort of economies of scale effect.

We made a more in depth analysis of OLS regression results at plants located in Palm Springs, California (Plant Code: 57743) and Chicago (Plant Code: 57191) in order to understand the potential effects of climate change on solar PV efficiency. We note that no data from these plants was used in training the linear regression model. First, we found that linear regression predictions of efficiency in 2016 (and other years) roughly follow the pattern of observed efficiency throughout parts of the year, but can make erroneous predictions (Figures 3, 4). The climatological pattern of PV generation shows that efficiency increases for both plants throughout the spring, peaks in summer, and decreases throughout fall. Under climate change conditions (represented by expected conditions in 2050), our model predicts a overall decrease in solar PV efficiency at both plants (Figures 3, 4). Interestingly, the decrease in efficiency in Palm Springs is greater than the decrease in efficiency predicted in Chicago.

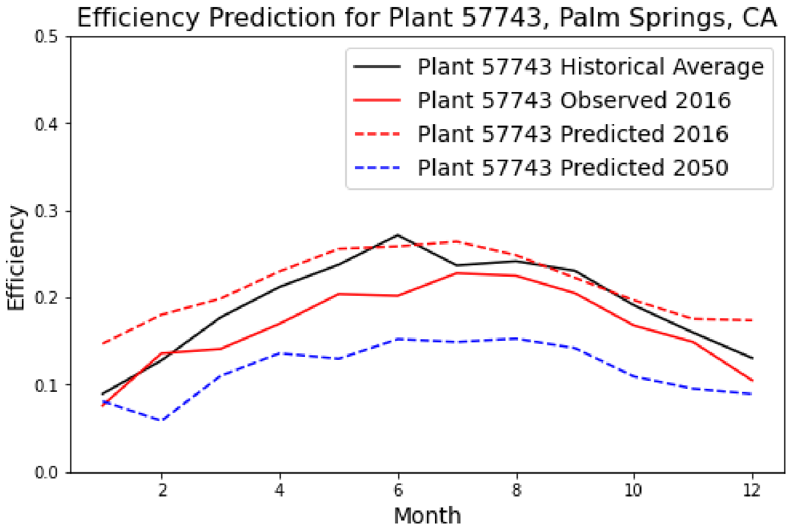

Figure 3: Predicted efficiency of plant 57743 in Palm Spring, CA using a linear regression model. Note that predicted solar PV efficiency in 2016 generally follows the pattern of observed efficiency in the same year. Also note that predicted efficiency in 2050 is roughly 50% less than the historical average.

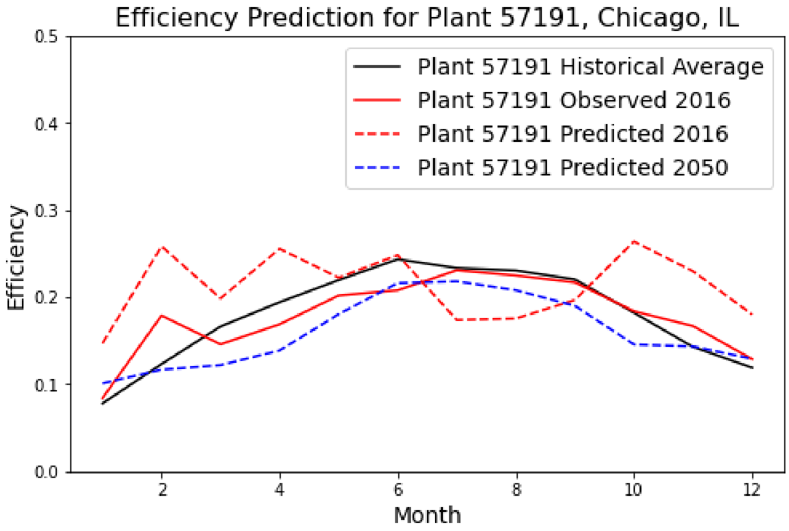

Figure 4: Predicted efficiency of plant 57191 in Chicago, IL using a linear regression model. Note that, although the predicted efficiency in 2050 is lower than the historical average, there is not a large decrease in expected efficiency.

7 Conclusions and Future Work

In this work, linear and logistic regression models for solar PV efficiency prediction were trained and evaluated using data from 144 solar PV plants across the United States. Using a more comprehensive dataset with more samples across the Midwest, Northern U.S., and Southeast U.S. would likely improve the accuracy of our models. Additionally, using higher resolution data at a daily or hourly scale is believed to provide more insight into the significance of climatological effects such as cloud cover and wind speed on solar efficiency. Unfortunately, this type of information is not readily available for all solar plants, and retrieving the data can be an intensive task. Additionally, climate change models are generally more accurate at coarser resolutions. Researchers pursuing similar work would benefit from a more consolidated database. The number of databases that we had to utilize, and the varying access formats of each database, make it difficult to describe the data retrieval process in such a way that other students or researchers could replicate it.

We found that linear regressions are able to roughly predict the annual pattern of efficiency of solar PV plants, although they cannot replicate month-to-month historical observations. Although we felt we had included a comprehensive list of possible predictors of solar PV efficiency, there appears to be some significant predictors missing. By applying a linear regression to geographical data and climatological data representative of the year 2050, we found that solar PV efficiency may decrease due to climate change.

8 References

Boretti, A., 2020. High-frequency standard deviation of the capacity factor of renewable energy facilities: Part 1—Solar photovoltaic. Energy Storage, 2(1), p.e101. https://onlinelibrary.wiley.com/doi/pdf/10.1002/est2.101

Gilman, P., DiOrio, N.A., Freeman, J.M., Janzou, S., Dobos, A. and Ryberg, D., 2018. SAM Photovoltaic Model Technical Reference 2016 Update (No. NREL/TP-6A20-67399). National Renewable Energy Lab.(NREL), Golden, CO (United States). https://www.nrel.gov/docs/fy15osti/64102.pdf

Panagea, I. S., Tsanis, I. K., Koutroulis, A. G., and Grillakis, M. G. (2014). Climate change impact on photovoltaic energy output: The case of Greece. Advances in Meteorology, 2014. https://doi.org/10.1155/2014/264506

Schiermeier, Q., 2016. And now for the energy forecast: Germany works to predict wind and solar power generation. Nature, 535(7611), pp.212-214. https://doi.org/10.1038/535212a

E. Lorenz, J. Hurka, D. Heinemann and H. Beyer, ”Irradiance forecasting for the power prediction of grid-connected photovoltaic systems”, Selected Topics in Applied Earth Observations and Remote Sensing IEEE Journal of, vol. 2, no. 1, pp. 2-10, March 2009.

N. Sharma, P. Sharma, D. E. Irwin and P. J. Shenoy, ”Predicting solar generation from weather forecasts using machine learning”, SmartGridComm, pp. 528-533, 2011.

A. Prastawa and R. Dalimi, ”New approach on renewable energy solar power prediction in indonesia based on artificial neural network technique: Southern region of sulawesi island study case”, QiR (Quality in Research) 2013 International Conference on, pp. 166-169, June 2013.

F. Jawaid and K. NazirJunejo, ”Predicting daily mean solar power using machine learning regression techniques,” 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, 2016, pp. 355-360, doi: 10.1109/INTECH.2016.7845051.

Stanley, S. (2020), Evaluating cloud cover predictions in climate models, Eos, 101, https://doi.org/10.1029/2020EO141682. Published on 23 March 2020.

World Climate Research Programme. CMIP6. Variables: Temperature, Windspeed, Direct shortwave radiation. Accessed Dec. 9, 2020. https://esgf-node.llnl.gov/search/cmip6/