UR Sustainability

Team 3:

Ian Costley, Marina Kupina,

Meghana Murthy, Ronald Michaels

DSC 383/483 – Final Project

University of Rochester

Fall 2020

Introduction

Background

With the rise of climate change at the forefront of concern in the 21st century**,** the rush for the innovation and uses of renewable energy sources has become ever more important. If this trend of increasing global temperatures isn't eliminated or reduced, our planet will be facing the costs of rising sea levels, droughts, heat waves, more intense hurricanes, and an arctic circle free from ice, amongst other results as well. There are many common characteristics of our lives that connect us as a society, however the main thread that intertwines all of us is that we live on the same planet. Thus, one small action over time can not only lead to harmful effects down the road for future generations, but also eradicate the livelihoods of people who rely on our now changing climates. Luckily, this problem is well known and there are many industries dedicated to reducing global carbon emissions. Furthermore, this goal doesn't rely on one method as there are multiple avenues for lessening our carbon footprint. While this project does focus on solar energy, it is important to also mention the other popular renewable energy methods of wind, hydro, tidal, geothermal, and biomass energy for reference [1]. Fig. 1 shown below from REN21 shows the estimated global breakdown in energy production in 2018, with solar energy coming in at only 2.4%. The piece of information that makes this image below so hopeful for the future of renewable energy is that according to the U.S. Department of Energy, the amount of sunlight received by the earth in just one hour is greater than the total energy used by the entire world for one year [2]. With these statements in mind, understanding the features that our climate has on solar production is paramount to unlocking such an abundant energy source.

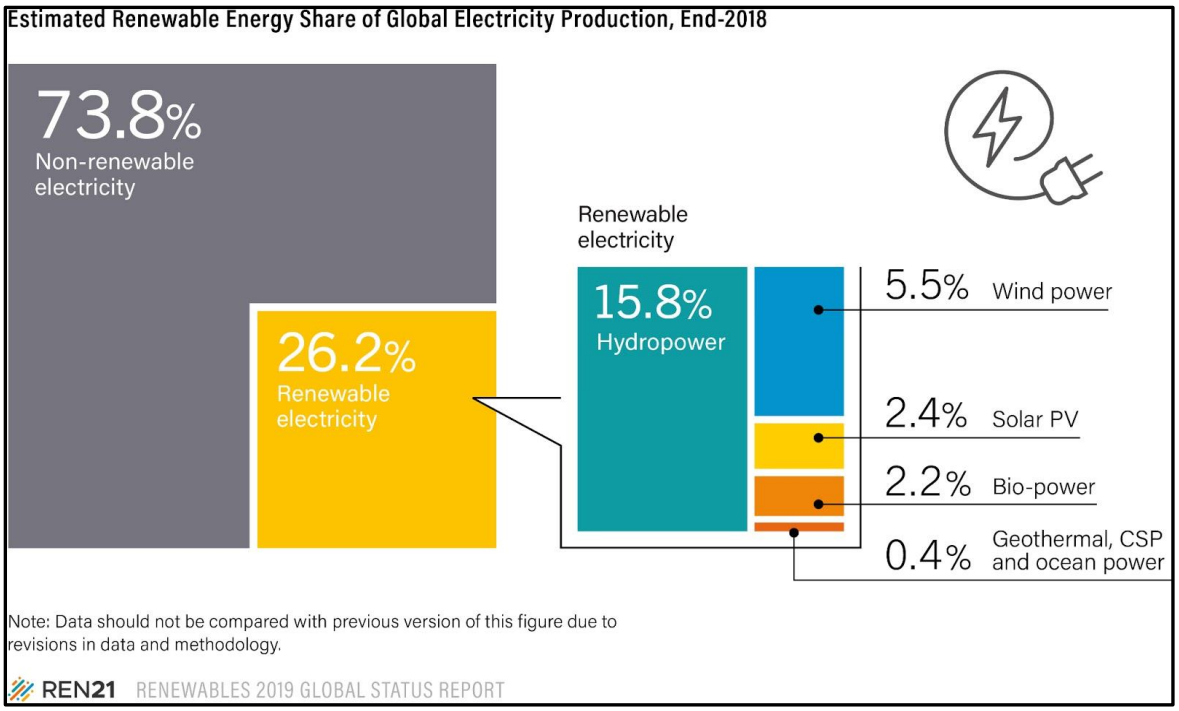

Figure 1: Breakdown of Energy Consumption [3]

Solar Panel System

The solar panel system at the University of Rochester sits on the roof of the Goergen Athletic Center (GAC) and contains 960 solar modules, each with a 350 Watt capacity, making the entire solar panel system capable of producing 336 kW at its capacity. These modules are oriented at a 10 degree vertical tilt with an azimuth alignment of 153 degrees. In terms of the flow of energy for this system, solar radiation is taken in by the panels, then that energy travels from the panels to eight different inverters where the solar energy is inverted from direct to alternate current. Afterwards, this current is then transferred to either TESLA battery packs for storage, for immediate use by the Goergen Athletic Center, or for redistribution amongst the University of Rochester power grid. At the moment, around 30 percent of the Goergen Athletic Center is powered by the solar panel system. Fig. 2 shown below displays a snapshot instance from the system in a one-line drawing.

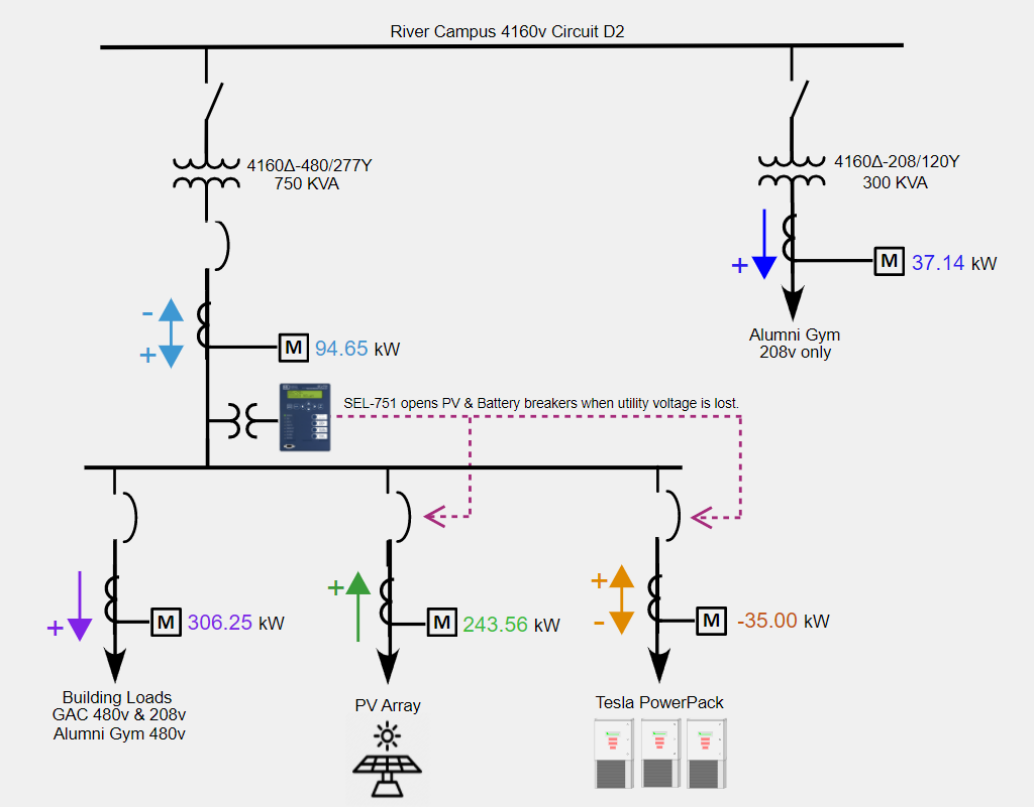

Figure 2: One-Line Drawing of UR Solar System [4]

Vision

The end goal of the solar panel system at the University of Rochester is to help the university become a place of net-zero emissions. In addition, benefits for this system extend far beyond just our carbon footprint. For example, this system is cost effective for the university as the average cost of electricity for the state of New York (14.34 cents/kWh) is greater than that of the United States (10.54 cents/kWh). [5] Furthermore, the cost of electricity continues to increase over time while innovation and economies of scale have led to a 70% drop in the cost of

Photovoltaic solar panel systems over the last nine years ($4.54/Watt in 2011 to now $1.34/Watt in 2020). [6] Education, of course, is another benefit as the data collected with this system can be used in classroom and lab settings, as well as a source for additional research and collaboration. Proof of net zero in upstate New York is widely desired and Rochester hopes to be a pioneer to help other schools in our area.

Purpose

The purpose of this project is to better understand why the University of Rochester solar panel system does not produce at its rated capacity. Specifically, to better understand panel production through the weather of Rochester, New York. Our sponsors, Prof. Doug Kelley from the UR Mechanical Engineering department, Prof. Karen Berger from the UR Earth & Environmental Sciences department, and Alex Pfadenhauer from the UR Utilities & Energy Management department, have graciously met and answered our questions on a regular basis to guide this project throughout the semester. By grasping an improved knowledge of the connection between Rochester's weather and panel output, we hope to provide new discoveries that will aid in not only solar production predictions, but financial expectations for solar energy as well.

Objective

Our objective is to provide newfound information to better help the university in understanding the relationships between solar panel output with seasonal and atmospheric variations.

Data Collection

The solar panel is located on the roof of the Goergen Athletic Center (GAC) at the University of Rochester, River Campus (Fig. 3). The panels are barely tilted and lay almost horizontal to the ground, at a tilt angle of 5-10 degrees. Exports of the solar panel output have been provided by the sponsor in an Excel Workbook. The exports include Real Power (3-phase average) – beginning Feb 2020 at a 1-minute resolution. The solar array consists of 8 inverters and the Real Energy per inverter beginning July 2020 is also provided at a 5-minute resolution. Lastly, solar irradiance in Watts per square meter is provided, beginning June 2020. The solar irradiance is calculated from a separate equipment located slightly further from the panels. Some external factors that might affect the data include the weather in Rochester, the shading of the buildings surrounding the GAC, or any electrical malfunction that doesn't allow the solar panel output to be recorded properly.

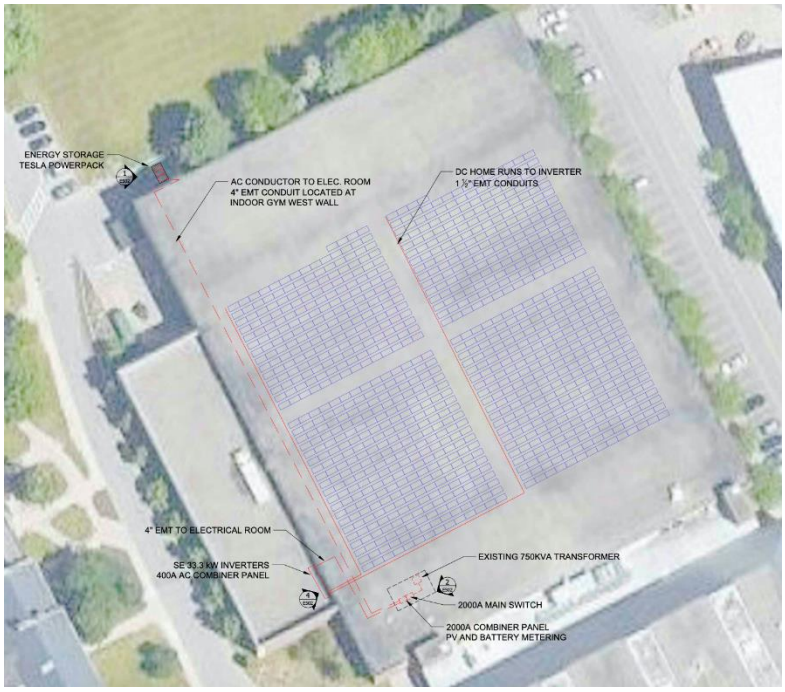

Figure 3: View of Solar Array from the roof of GAC

The weather data is collected from the National Oceanic and Atmospheric Administration [7] (NOAA). The weather data export contains features like temperature, relative humidity, solar elevation, solar duration, cloud cover et cetera. The attribute for cloud cover is measured on a numeric scale from 0 to 8 with 0 meaning clear skies and 8 being overcast. Rochester being the cloudy city that it is, predominantly has a high cloud cover number. Additionally, the modeling of the residuals (described later) requires the solar radiation at the top of the atmosphere. This feature isn't collected so much as it is calculated through a series of complex formulae and approximations.

Data Preparation

We prepare two separate datasets that each align with a separate modeling objective. The first dataset is comprised of daily aggregates and will be used to answer the main objective to describe the relationship between temperature, cloud cover, and solar radiation as predictors for solar panel power/output. An hourly dataset is comprised of hourly aggregates of the features describe prior. In addition to the hourly aggregates, feature engineering to the already present attributes is performed to aid in exploring features that might be important predictors of solar panel output, but were not considered in the main objective.

Daily

For the daily dataset, weather data is collected from NOAA and merged with the solar panel data. From there, we exclude any time values that are not within an 8AM to 4PM window. Finally, we average all attributes over each day and have a dataset consisting of ~250 rows/days.

Hourly

For the hourly dataset, weather data is also collected from NOAA and merged with the panel data. We then excluded any times values that occur before sunrise or after sunset for a given day. Next, average values for each attribute at an hourly level are calculated. Finally, we perform extensive feature engineering to increase the hourly aggregate dataset comprised of 19 features to 153 features. These further features include maximum, minimum, standard deviation, etc. for each attribute, at each hour. After consulting the relevant literature, we defined additional features that may have an impact on the target variable, such as sun declination angle and duration of sun exposure. In addition, June 21st is the day of the summer solstice when the tilt of a planet's axis in the northern hemisphere is most inclined toward the Sun, we calculated the deviation from this date. Since most predictive models take numeric values as input, we apply one-hot encoding for categorical variables that creates a binary column for each category and return a sparse matrix.

Exploratory Analysis

To begin answering our objective, we need to understand what proportion of solar panel efficiency, if any, can be explained by cloud cover and/or temperature. Our baseline models explore the relationship between these variables. Specifically, we will build a linear combination of the predictors that explain the most amount of variation with the solar panel output. An illustration of this linear combination is seen in Fig. 4.

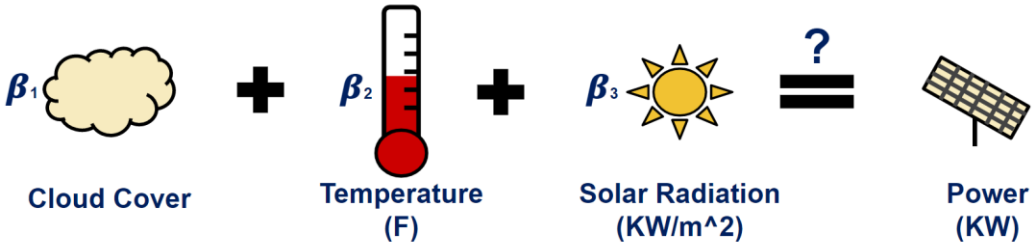

Figure 4: Baseline Regression Models

Before baseline modeling can occur data exploration is required to understand how these variables interact with each other, and the target variable. In Fig. 5, we show the correlations and daily averages for cloud cover, temperature, and solar radiation plotted against panel power output.

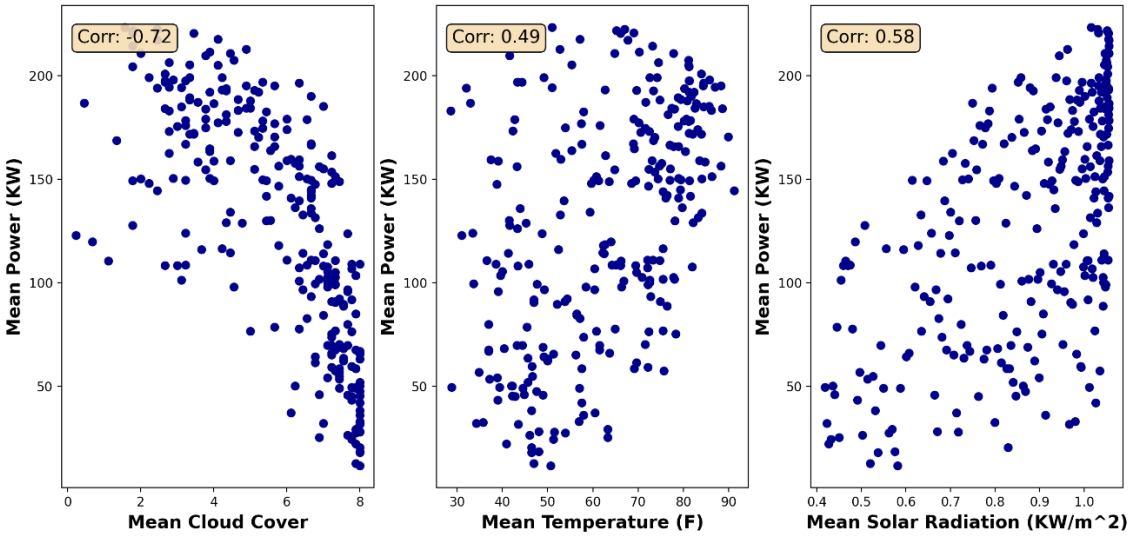

Figure 5: Cloud Cover, Temperature, and solar radiation versus panel output power.

In the plots above, each blue dot represents daily average values between 8AM-4PM. It is clear that both average daily temperature, cloud cover and solar radiation are linearly correlated with average daily power. Now, we examine the correlation between each predictor in Fig. 6.

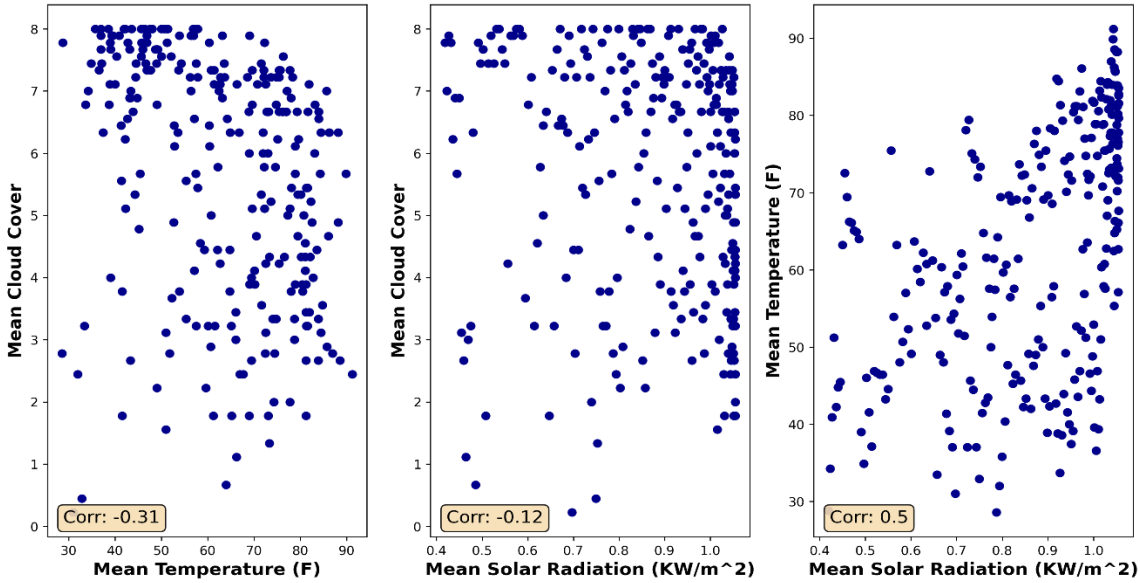

Figure 6: Cross correlation between predictor variables.

Notice, there is only a slight linear correlation between cloud cover vs. temperature and cloud cover vs. solar radiation. Although, there is a correlation between temperature and solar radiation. This correlation is hypothesized to be attributed to the already seasonal dependence of temperature. For example, on days of the year when there is a higher average solar radiation, it is expected that there is a higher average daily temperature.

Methods

Predictive Models

First, we employ multiple linear regression (MLR) as a baseline model for our analysis. The data used for both the baseline analysis and more advanced predictive modeling will be the daily data set. More specifically, we will only consider cloud cover, temperature, and solar radiation (used as a proxy for explaining the effects of solar seasonal variation) as predictors. Our main goal with MLR is not predictive performance, but rather the coefficients and confidence intervals around the coefficients for the MLR model. Since the predictors used to fit the model have been standardized, the magnitude of the coefficients will convey the relative importance of each predictor to the MLR model. Furthermore, as a result of the predictor's standardization in the data processing set, if zero lies within the confidence interval for any of the coefficients, then this predictor is not statistically significant and can possibly be removed from further analysis.

For the advanced predictive model, we will use Gaussian Processes (GP). The main advantage of GPs over MLR is the increased flexibility and ability to model non-linearities in the data. Furthermore, GPs are highly interpretable and make the same basic assumption as MLR: the data are normally distributed. GPs work by creating a conditional distribution of the target variable, given the predictors. Since normality is conserved under conditioning, the resulting distribution is also Gaussian. Thus our resulting model only needs a mean and covariance function (parameters of Gaussian distribution). As stated prior, our data are standardized and have a mean of zero, which in turns allows us to set the mean function of the GP to zero. Now, we effectively only have to fit one parameter for the model: the covariance function.

The covariance function for GPs can range from periodic to linear correlation functions. They are highly customizable and are one of the main advantages for using GPs. For our modeling procedure, we used the radial basis function (RBF) to describe the covariance between points in the data. The main advantage to using RBF as the covariance function is maintaining generality in the model. Using RBF, we do not assume any predefined shape of the data. That is, we do not assume periodicity, linearity, or even non-linearity, as the RBF covariance function approximates to MLR if the data follows a linear trend.

Models for Further Feature Understanding

In the models for "*Predictive Modeling*", it is detailed how we will explore the effects of cloud cover, temperature, and solar radiation on solar panel power. Therefore, it seems feasible to explore other variables and how they impact solar panel power. In order to analyze more variables and their impact on total panel efficiency, we decided to create a series of gradient boosted models, which are known for their ability to recognize feature significance.

The Gradient Boosting Model is an extremely popular machine learning algorithm that has proven itself in many fields and is one of the leading state-of-the-art predictive analytics methods.

The model allows you to detect and adjust for nonlinearity, while ignoring missing values. The algorithm sequentially builds an ensemble of shallow decision trees, where each tree improves the performance of previous trees. Even though each decision tree is a fairly weak predictive model, a combination of decision trees can provide high performance. One of the main benefits of using ensembles of decision trees is that they can automatically provide estimates of feature importance from a trained predictive model. Feature importance provides a score that indicates how valuable each predictor variable was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.

The goal of this feature importance analysis is to implement two gradient boosting models using the preprocessed hourly dataset, compare their accuracy, and determine their feature importance that we will use for further exploration. The hourly dataset is used due to the gradient boosting model's need for a large amount of data. Thus, the hourly dataset clearly has more data points and is therefore used for this analysis. For the boosting models, XGBoost and LightGBM are chosen, which are two popular gradient boosting algorithms. Both algorithms are based on a similar idea, however, they use different methods to filter data instances to find the split value. LightGBM grows trees vertically (leaf-wise tree growth) while XGBoost grows trees horizontally (level-wise tree growth). Thus, the leaf-wise algorithm can reduce more loss than the level-wise algorithm. [8]

Results

Predictive Models

From the baseline analysis using multiple linear regression, we have the following coefficients for the models detailed in Table 1. Note, the coefficients are accompanied by their upper and lower bounds for a 95% confidence interval. Finally, as a reminder, all data fed into the model has been standardized. For simplicity, we refer to "T" as temperature, "CC" as cloud cover, "SR" as solar radiation, and "P" as solar panel power when describing the models.

Models	Temperature Coef.	Cloud Cover Coef.	Solar Radiation Coef.
$P = \beta_1 \cdot T$	0.49 ± 0.11	NA	NA
$P = \beta_2 \cdot CC$	NA	-0.72 ± 0.08	NA
$P = \beta_3 \cdot SR$	NA	NA	0.58 ± 0.10
$P = \beta_1 \cdot T + \beta_2 \cdot CC + \beta_3 \cdot SR$	0.046 ± 0.071	-0.64 ± 0.06	0.48 ± 0.07

Table 1: Coefficient values and confidence intervals for the baseline regression models.

The first observation that is apparent is the change in magnitude of the temperature coefficient when compared to the simple linear regression model and the final multiple linear regression model. Notice, the temperature coefficient is no longer significant at the 5% significance level once the other attributes have been added. This further confirms our hypothesis stated in the "Exploratory Analysis" section that temperature acts as a proxy for seasonal solar effects when solar radiation is not present. Additionally, notice the decrease in magnitude for cloud cover and solar radiation from the simple linear regression model to MLR model. This suggests that both cloud cover and solar radiation are required for modeling solar panel output, but there is small correlation effect between the two predictors. As is expected, since cloud cover is slightly seasonal.

Now, we fit a more sophisticated GP model to the data above and report the 5-fold root-mean squared error (RMSE). The term "5-fold" is a shorthand for 5-fold cross validation, where in which the data is split into 5 equal length partitions. We iterate through the partitions leaving one partition to be testing and the other 4 to be training. The RMSE metric is calculated for iteration and averaged after the final iteration. The equation for RMSE is given in equation 1.

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

Equation 1: Root-mean squared error

where $\hat{y}$ is the predicted output.

From the equation above, we calculate the RMSE for GPs fit on different combinations of the predictors, and compare it to the 5-fold RMSE for the MLR baseline model. The 5-fold RMSE for the compared models is given in Table 2.

Model Type	Predictors	5-Fold RMSE
MLR	T, CC, SR	27.46 KW
GP	T	48.75 KW
GP	CC	34.24 KW
GP	SR	46.42 KW
GP	CC, SR	24.07 KW
GP	T, CC, SR	23.43 KW

Table 2: 5-Fold RMSE values for different models.

Cleary, the best model above is the Gaussian Process with all predictors. Note, the model with just solar radiation and cloud cover as predictors is slightly worse than the full model. Hence, the following analysis will assume that the full model is indeed the best model and that temperature should be included in the final model. Now, we perform a sensitivity analysis using the full GP model. For the sensitivity analysis, we will choose a predictor to view the entire spectrum of, and vary the values of the other predictors and observe the changes in the chosen predictor. For example, if we choose cloud cover, we want to observe the models output over the entire spectrum of cloud cover values where solar radiation and temperature are kept static. For this, we plot the spectrum of cloud cover under two scenarios. One, temperature is set at its mean value (0 since standardized predictors), and solar radiation is varied at different constant values. We then repeat the process where solar radiation is set at its mean and temperature is allowed to vary. The following example is shown in Fig. 7.

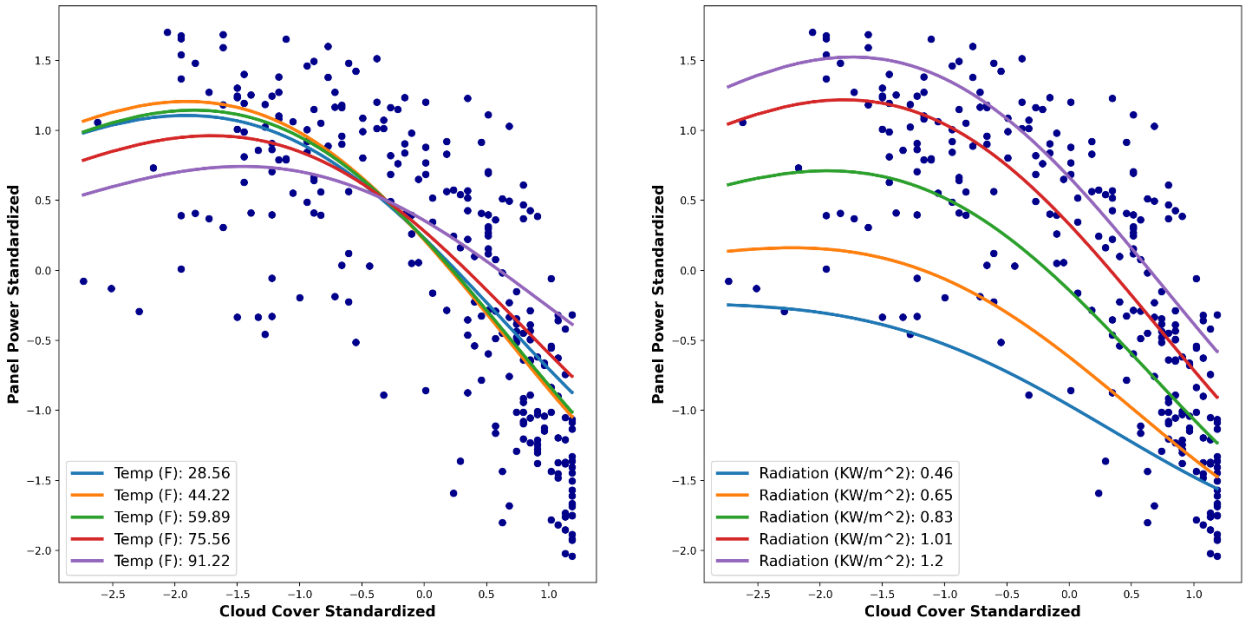

Figure 7: Sensitivity analysis for cloud cover.

Notice, for the left graph is Fig. 7, we observe the entire spectrum output for cloud cover under different constant temperatures where solar radiation is set to its mean. From the graph, we notice that temperature does not have much of an effect on solar panel output, unless we are at the tail ends of the spectrum for cloud cover. In which case, a higher temperature appears to raise the solar panel output for high values of cloud cover and lower the panel output for low values of

cloud cover. In the graph on the right of Fig. 7, it is apparent that solar panel output is very sensitive to different values of solar radiation. The sensitivity analysis for temperature and solar radiation are seen in Fig. 8 and 9 respectively.

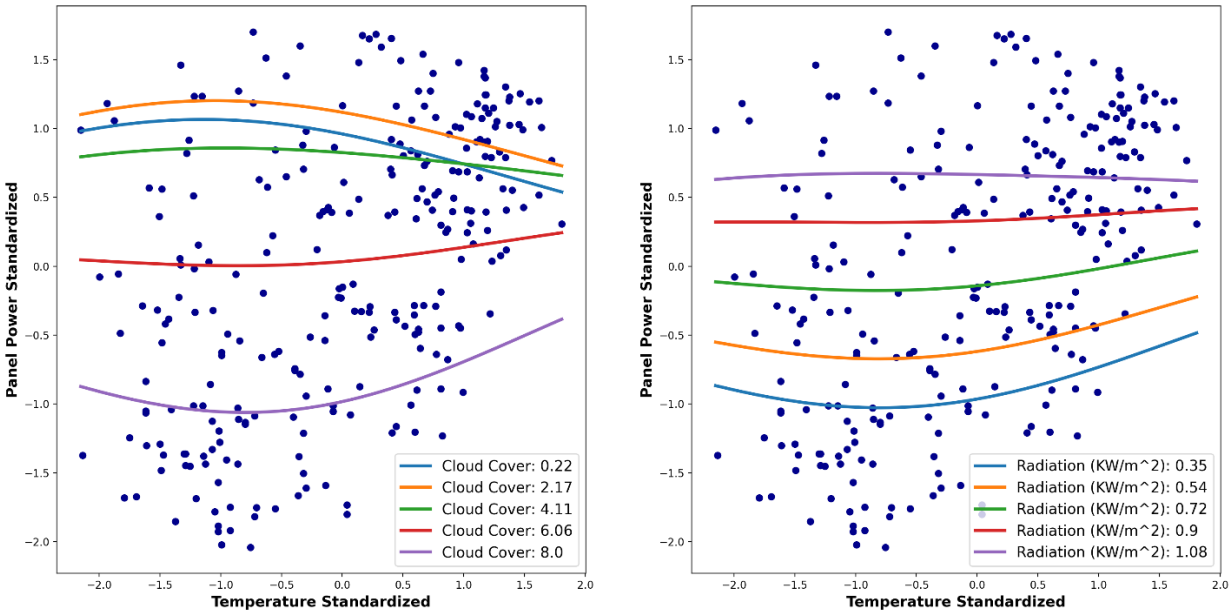

Figure 8: Sensitivity analysis for temperature.

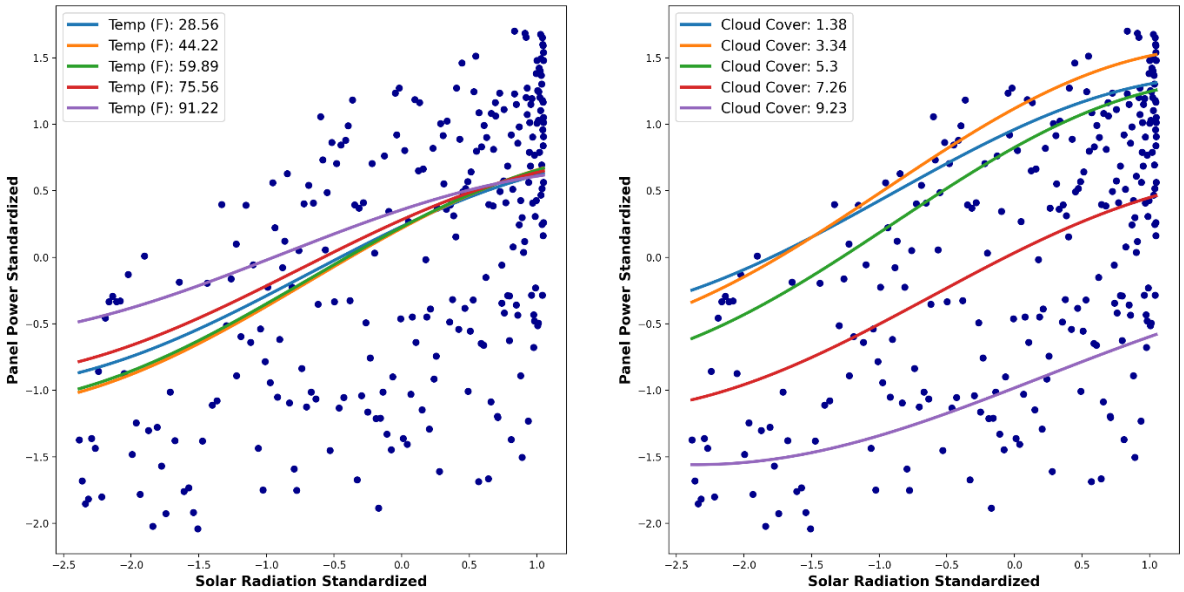

Figure 9: Sensitivity analysis for solar radiation.

Note, from all three sensitivity analysis graphs, it is clear that temperature is the least important predictor of the three predictors. Solar panel output is highly variant to changes in solar radiation and cloud cover, but not temperature. Temperature becomes more of a helpful attribute at the extremes of the solar radiation or cloud cover spectrums. Furthermore, observing the sensitivity lines in Fig. 8 shows they are almost linear with slope 0, suggesting that temperature

provides little use to the model except at extremes, which we observe both in the 0.35 KW/m^2 sensitivity line on the right graph of Fig. 8 and varied solar radiation sensitivity curves on the left graph of Fig. 9.

Gradient Boosted Models

Both models are trained and tested on the same hourly dataset created in "Data Preparation" section. The model takes hourly observations as input, predicting the total power for a given hour. Since gradient boosting models have many hyperparameters, such as maximum tree depth and learning rate, parameter tuning is an important step that improves overall model performance. To select hyperparameters, Gridsearch is used, which trains the model on various sets of parameters and selects the best configuration. In this part, R^2 and RMSE are used as metrics for evaluating and comparing models. The equation for R^2 is given in equation 2.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i \left(y_i - \bar{y}_i\right)^2}$$

Equation 2: R^2

where $\hat{y}$ is the predicted output and $\bar{y}$ is the average value for the output.

As a result of training the model, the performance of both models is comparable. The feature importance histograms (Fig. 10, 11) show that the 12 most important features are similar for both algorithms.

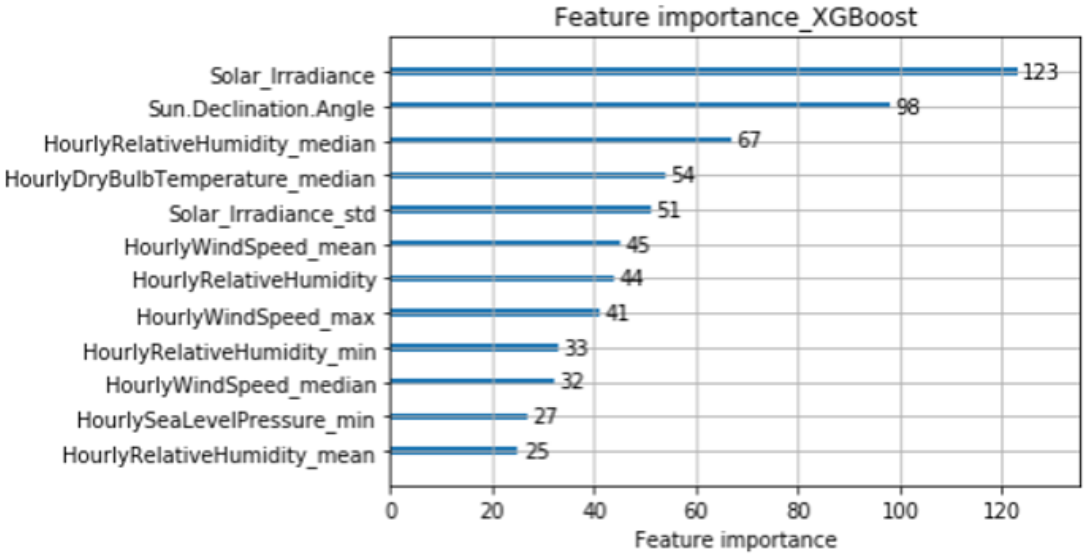

Figure 10: Feature importance for XGBoost ($R^2 = 0.817$).

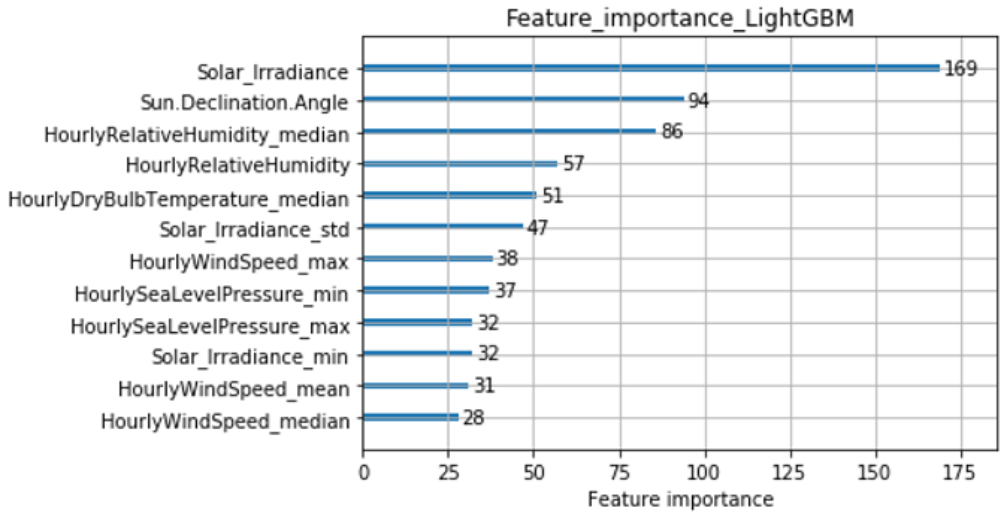

Figure 11: Feature importance for LightGB ($R^2 = 0.831$).

In addition to the UR solar data, an experiment is conducted on external data from the Rochester Institute of Technology (RIT). Since these solar panels are also located in Rochester, the dataset contains the same weather observations. The main difference between the UR and RIT datasets is the characteristics of solar panels and installation specifics. As a result of model training, the performance of the model trained on UR data is higher than for the RIT dataset, which can be caused by the peculiarities of the solar panels. Table 3 below demonstrates the output for both models. It can be seen that humidity is one of the most common characteristics encountered in terms of the feature importance. Thus, we take into account these results for the further exploration.

	UR Data	RIT Data
R^2	0.817	0.754
Feature Importance	Solar Irradiance Sun Declination Angle **Hourly Relative Humidity - median** Hourly Dry Bulb Temperature - median Solar Irradiance - std Hourly Wind Speed - mean **Hourly Relative Humidity** Hourly Wind Speed - max **Hourly Relative Humidity - min** Hourly Wind Speed - median	Sun Duration Per Day Sun Declination Angle **Hourly Relative Humidity - max** Hourly Dry Bulb Temperature - mean **Hourly Relative Humidity** Hourly Dry Bulb Temperature Cloud Cover overcast **Hourly Relative Humidity - min** **Hourly Relative Humidity - median** Hourly Wind Speed – mean

Table 3: Results comparing UR and RIT model output.

As we noticed from the provided results of feature importance, humidity aggregations are considered as the most important feature in implemented predictive boosted models. Hence, we decided to make additional research on this feature. It can be seen from the plots in Fig. 12 that average daily humidity has a negative linear correlation with average daily power. Recall that each blue dot represents daily average values between the time frame of 8 AM - 4 PM. Moreover, there

is a strong positive correlation between the average daily humidity and cloud cover. Also, we explored the influence of wind speed on the average daily power. However, after our research, we were unable to confirm the existing relationship between these variables.

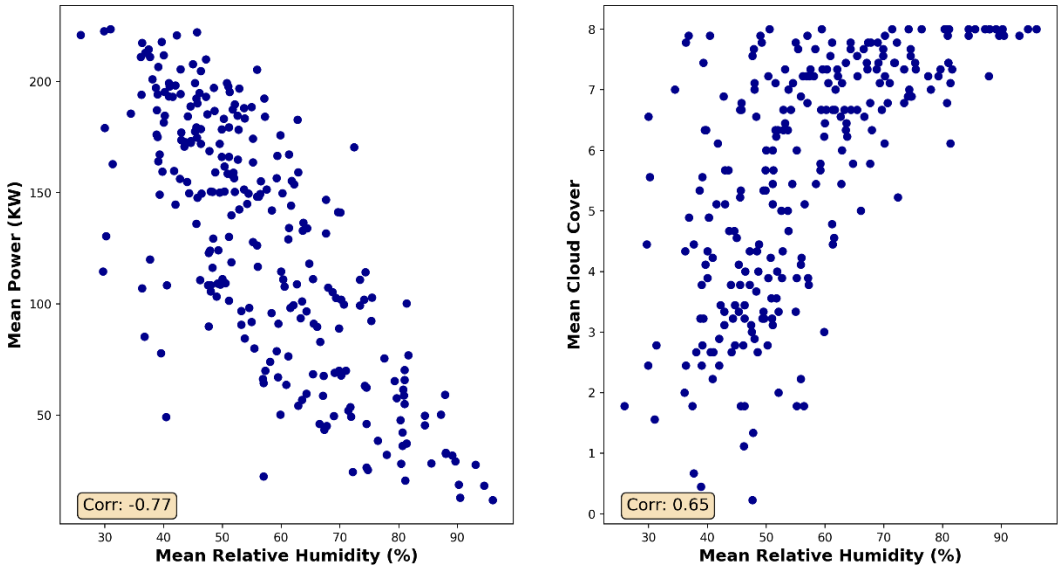

Figure 12: Exploratory graphs for humidity.

Conclusion and Next Steps

Main Objective

Investigation into our main objective yields 3 main conclusions in regards to the effect of temperature, cloud cover, and solar radiation on the effect of solar panel output. One, temperature on its own serves mostly as a proxy for solar seasonal variation, and contributes little to modeling in the presence of solar radiation as an additional predictor. Second, cloud cover is the most important attribute when modeling solar panel output, followed by solar radiation. Finally, all three predictors are necessary to build a reliable model for solar panel output, as temperature aids in prediction at the extremes of cloud cover and solar radiation.

Features outside of Main Objective

From the analysis provided by the gradient boosted models, we have that humidity is also an important attribute and should be considered when modeling solar panel output. Although, the details surrounding this attribute and its effect on solar output have not been solidified and is left as a next step if this project be continued upon in the future. For example, we observed a relatively high correlation between cloud cover and humidity. Thus, to fully understand the effect humidity has on solar panel output, variation in cloud cover should be removed from humidity. That is, the effect of humidity on solar panel output should be observed in a sub-analysis where cloud cover is kept constant.

References

[1] https://www.edfenergy.com/for-home/energywise/renewable-energy-sources

[2] https://www.osti.gov/biblio/899136

[3] https://www.ren21.net/reports/global-status-report/

[4] https://www.facilities.rochester.edu

[5] https://www.eia.gov/electricity/state/

[6] https://www.seia.org/solar-industry-research-data

[7] https://www.noaa.gov/

[8] https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db