Springboard Capstone Project 1 – Final Report
Predicting Short Term Solar Energy Production

Connor McAnuff

September 25, 2019

1. Overview

1.1 Problem Statement

1.1.1 Solar energy generation

Global solar energy generation capacity has been increasing exponentially since 2006 and is forecast to continue along the same trend through 2021 [1]. A form of solar energy generation is using photovoltaic (PV) panels, which produce electricity through interaction with photons from the sun [2].

1.1.2 Predicting solar energy generation

The solar energy availability at a given location can be described by the level of solar irradiance (generally stated in W/m^2) and can be measured using a pyranometer [3]. Solar irradiance can be summed daily to get a measure of J/m^2 and be used as a measure of the energy available to enter a solar PV system. The system configuration and losses through the PV system are well known, thus if the solar irradiance is known, solar energy production can be accurately predicted. Therefore, the short term solar energy production of a PV array can be predicted by proxy through the prediction of daily incoming solar energy (the sum of the solar irradiance for the day). Solar irradiance at the Earth's surface is in part determined by weather conditions, as cloud coverage and precipitation obscure the suns' rays from reaching the Earth's surface through reflection and refraction [4].

1.2 Value to client

Solar energy generation presents a unique challenge for electric utility companies as the energy generation varies in part depending on weather conditions. Utilities companies must be able to accurately forecast electricity production to prevent energy shortages and surpluses. Energy shortages can result in costly emergency purchases from neighbouring utility companies or blackouts, while energy surpluses can result in wasted energy as electricity cannot currently be feasibly stored in large amounts. Forecasting solar energy generation is a key component in several grid-balancing decisions such as reserve activation, short-term power trading (with other utility companies), peak load matching, and congestion management [4].

Accurately predicting short term solar energy production across many "solar farms" would provide value to utility companies by providing information for better grid-balancing decisions, resulting in a more efficient operations and reduced costs.

2. Data Wrangling

2.1 Data overview

The raw data has been provided in the following format (raw data source):

- station_info.csv:
  - Array of station ID, latitude, longitude, and elevation (98 rows x 4 columns).
- train.csv:
  - Array of dates and the recorded daily available solar energy measured using a pyranometer at each of the 98 Mesonet Solar Farms from 1994-01-01 to 2007-12-31 (5113 rows x 98 columns).
- Weather Variable Forecasts:
  - 15 NETCDF4 files (one file for each weather variable) listing the variable forecast value for each of the 11 predictive models, 5 forecast hours, 9 latitudes, and 16 longitudes for each of the 5113 forecast days (dimensions 11, 5, 9, 16, 5113).

2.2 Importing

Stations_info.csv and train.csv were imported directly into Pandas DataFrames named stations and energy respectively. The 15 weather variable forecast files were located using `glob` and the data were imported into a list of data using `xarray`. Next, the list of data was converted into a list of DataFrames.

2.3 Cleaning and Organization

2.3.1 Missing values

There are no null values in the energy, station, and weather variable data. Null value checks were performed using `isnull()`.

2.3.2 Outliers

The client stated that the pyranometers occasionally ceased functioning correctly. The client filled in the missing values with fictional values. Using `value_counts()`, it was determined that these fictional values end in non-zero numbers whereas the remaining (non-fictional) values end with zero. Figure 1 shows the 1999 solar energy availability for ACME and CLAY stations. From May-July, CLAY station shows a constant energy value of 12320768 J/m^2.

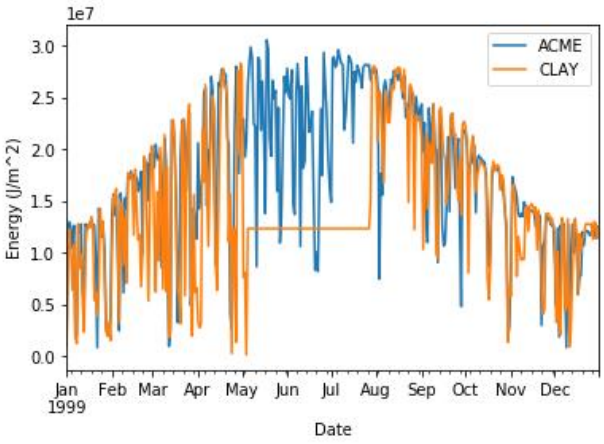

Figure 1: Daily available solar energy at two stations for the year 1999. CLAY has fictional data for May-July.

The fictional values were removed from the dataset after data formatting. They comprise <1% of the raw data.

2.3.3 Formatting

The goal of data formatting was to create a list of observations that include the date, station, available solar energy (target variable), and the machine learning features for that day and station. The process utilizes a nested for loop, iterating through each station, and for each station iterating through each weather variable. The steps are as follows:

1) Merge the list of energy data for a specific station with the list of stations.
2) Determine the closest weather forecast gridpoint (using longitude and latitude) to the station.
3) Get the weather variable forecasts for all dates, forecast hours, gridpoints, and 11 predictive models.
4) Reduce the data by using only the weather variable forecasts from the latitude and longitude of the closest gridpoint to the station.
5) Take the median value of the 11 different predictive models as a single weather variable forecast.
6) Pivot the forecast hour to be 5 different columns for each of the 5 forecast hours (and therefore 5 different features).
7) Merge the weather variable predictions/forecasts for each date with the total energy availability and add to the final DataFrame.

3. Exploratory Data Analysis

3.1 Data storytelling

3.1.1 Investigation

The following questions were asked of the data:

- What are the distributions of Mesonet station location (latitude/longitude) and elevation?
- How do the energy measurements vary over time (month-to-month, year-to-year) for a single station and for all stations combined? For all stations combined, how variable is the data for a given month/year?
- What is the distribution of total energy by station location?
- What are the differences in the energy data of stations in the east of Oklahoma vs the west of Oklahoma?
- How do the total/average of weather forecast variables vary in space and time and how do they relate to the energy?

3.1.2 Mesonet stations

The stations are spread evenly across Oklahoma (Figure 2). The state spans approximately 8.5° of latitude and 3° of longitude. The difference in elevation from the lowest (110 m) to highest (1322 m) station is 1212 m. The elevation of the stations steadily increases from east to west. This trend is indicative of Oklahoma's variable geography. The eastern side of Oklahoma is lower and contains extensive forested areas, central Oklahoma contains a transition from forest to prairies, and in the north-west there are flat grasslands.

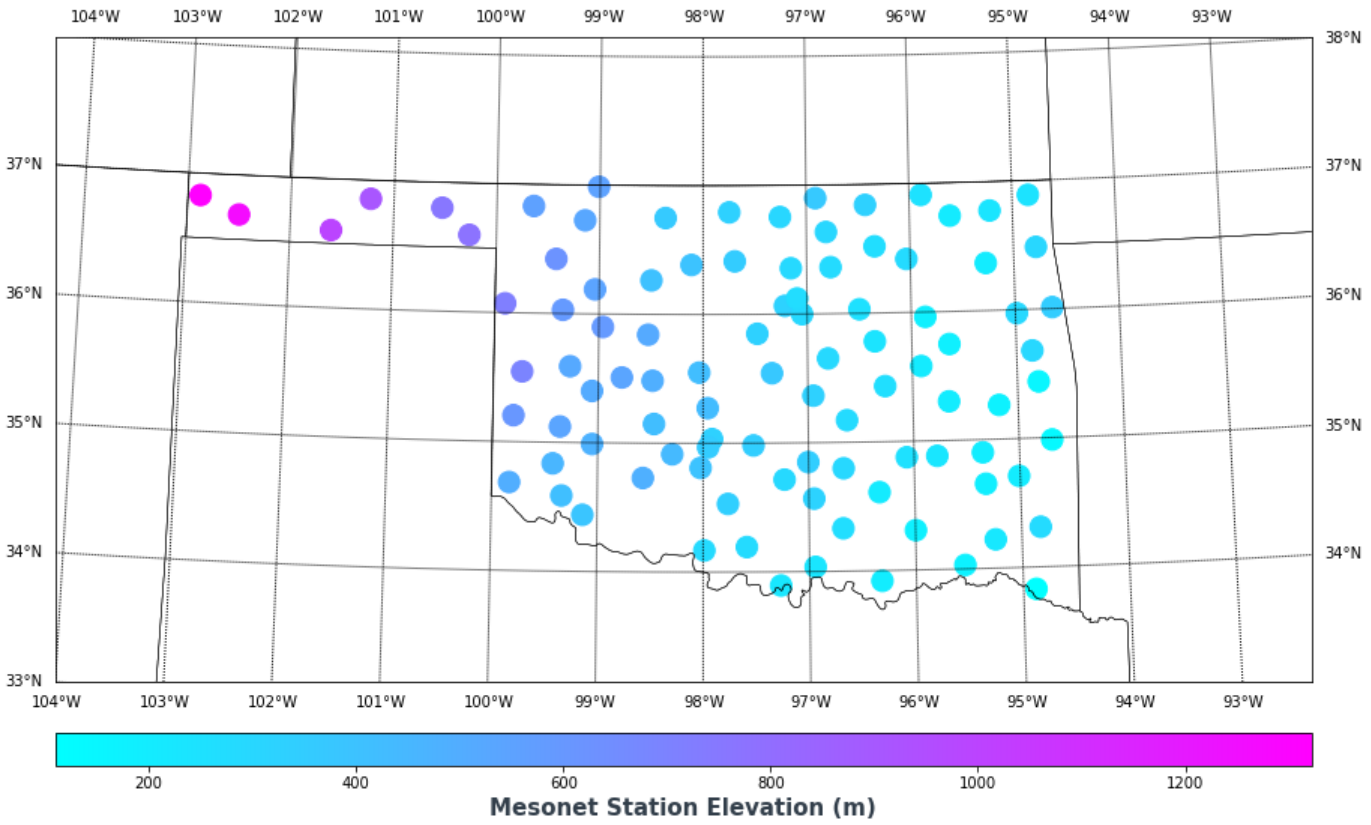

Figure 2: Mesonet station location and elevation.

3.1.3 Energy over time and space

Visualizing the daily energy data for a single station from 1994-2007 (alongside a 14-day rolling average) on a time-series scatterplot reveals that it is cyclical (Figure 3). Energy peaks yearly during June/July and troughs during December/January. This trend is expected, as Oklahoma is significantly north of the equator, thus from December 21 to June 21 of each year, the sun's time and angle above the horizon increase (and vice versa).

Year-to-year, the data does not appear to vary significantly. The energy in the winter months has a consistent maximum but also appears to be more variable than the summer months. Additionally, there appears to be significant variability in the data within short periods of time along the entire 14-year span.

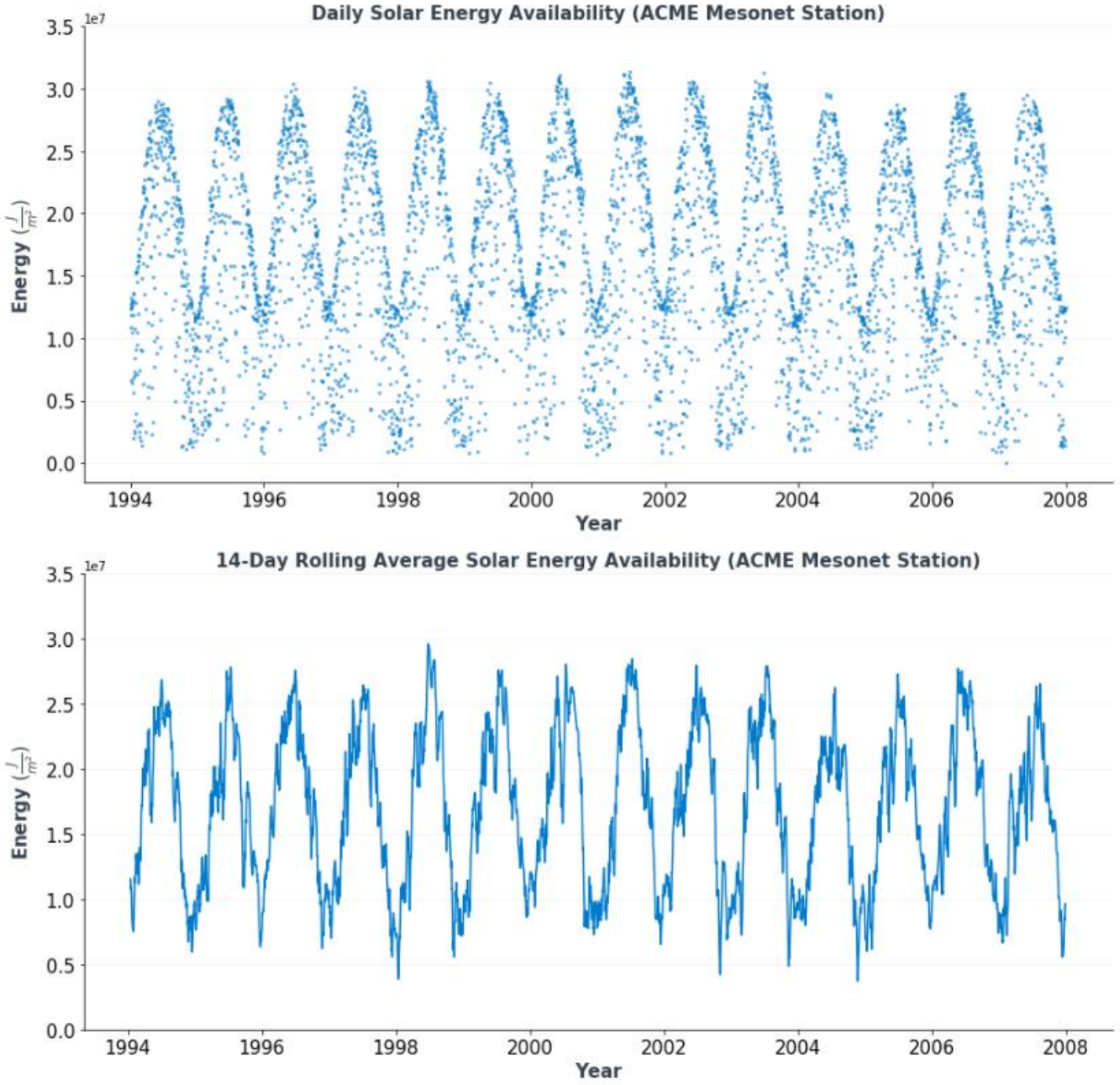

Figure 3: Daily and 14-day rolling average solar energy availability at ACME station.

Comparing the yearly and monthly energy totals for all stations combined, it is seen that there is little variation year-to-year (Figure 4). The combined station monthly data shows the same cyclical trend as the single station data. Yearly distributions shown by box and whisker plots and violin plots show that the yearly distributions are very similar year-to-year (Figure 5). 1994 appears to have a greater spread than any other year – this is due to a single datapoint from 1994-04-07 at IDAB Mesonet station.

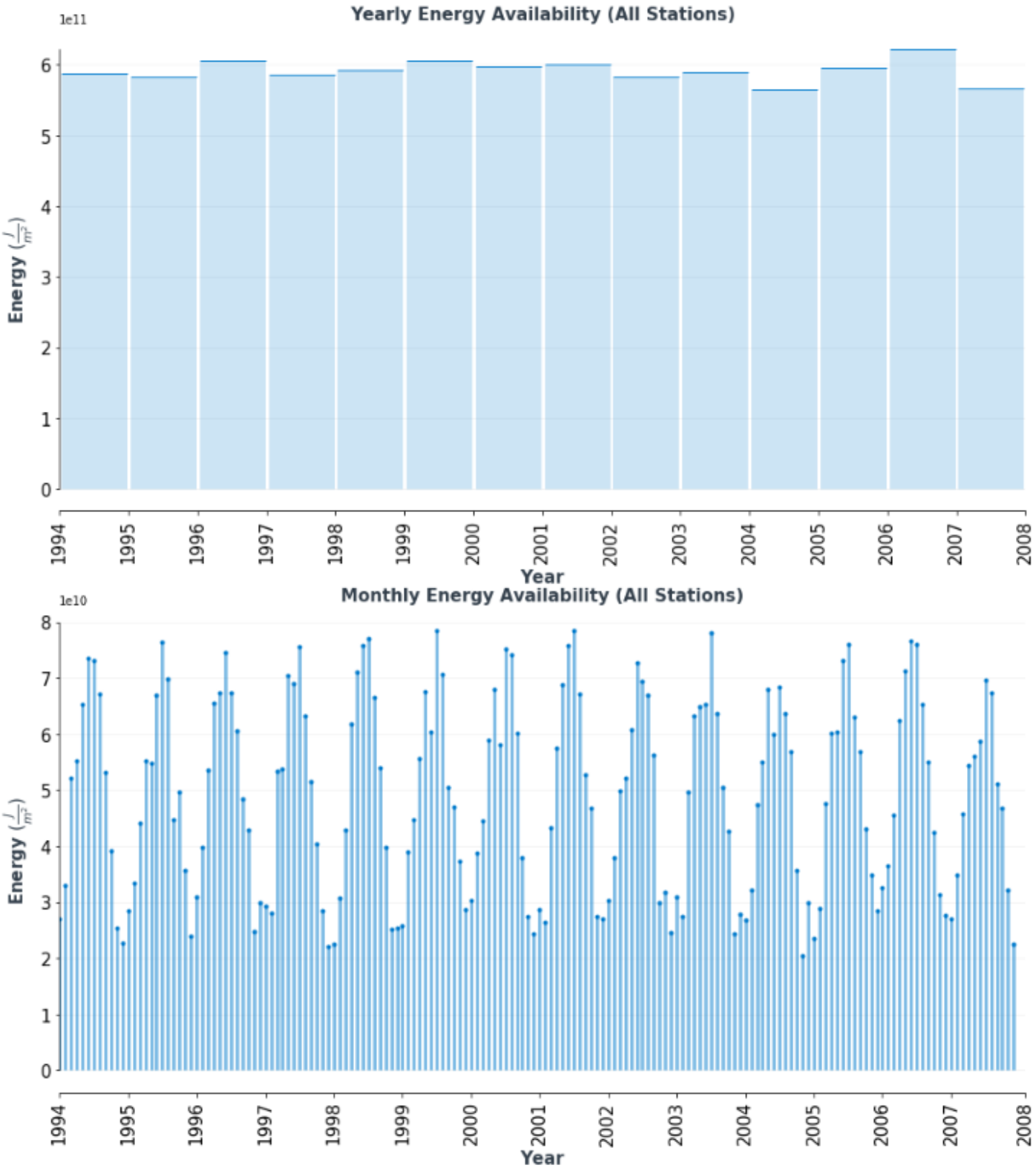

Figure 4: Yearly and monthly energy availability for all stations combined.

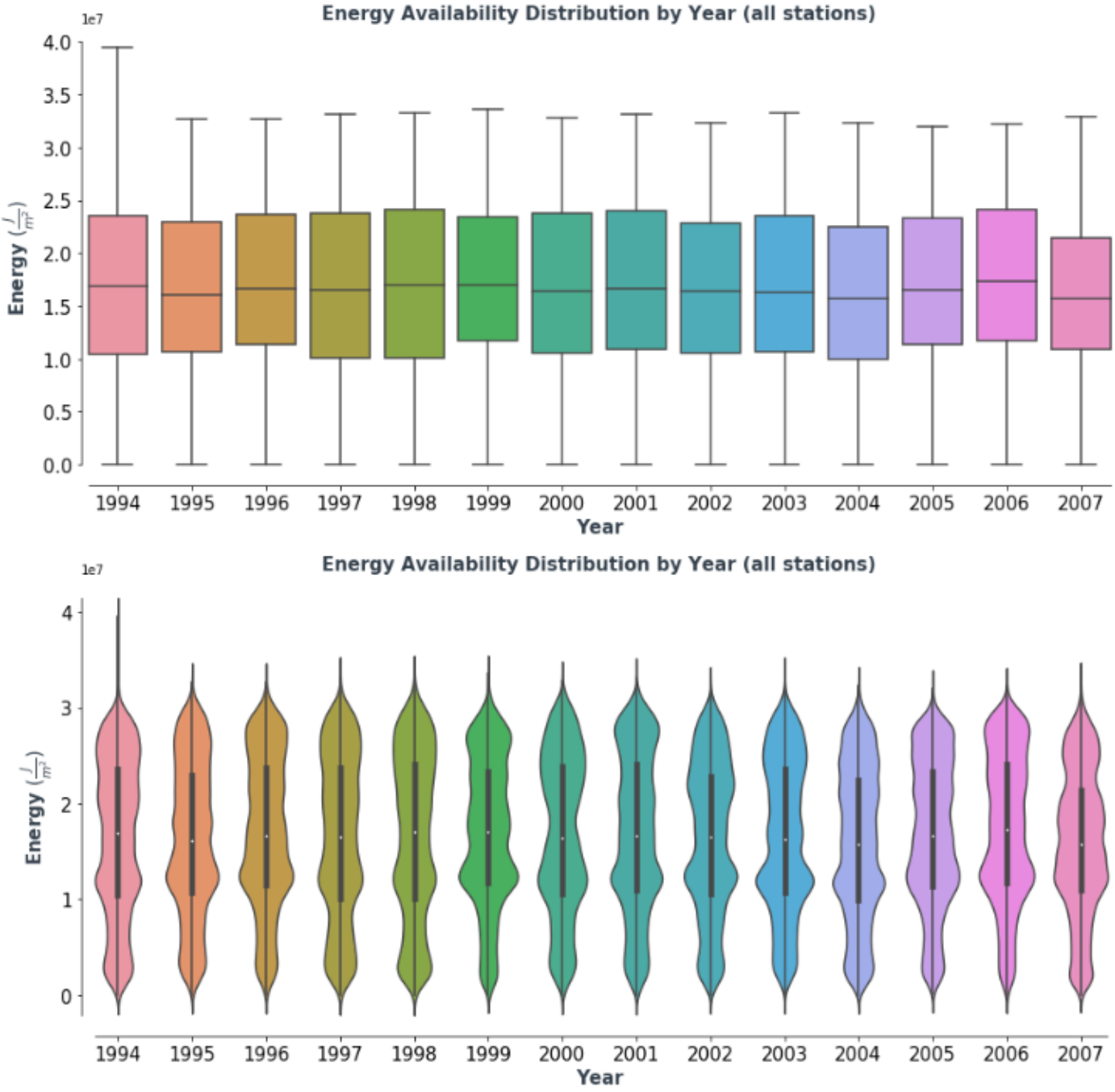

Figure 5: Daily energy availability distributions by year for all stations combined.

The monthly distributions (1994-2007) allow the cyclical trend to be further explored (Figure 6). The median value peaks in June/July and troughs in December/January. June-September have fewer low values than the remaining months - low values are shown as outliers as opposed to the box whiskers extending to 0. Overall, the summer months have a distribution heavily concentrated at the peak, with a long tail extending to 0, while the winter months are bimodal, having a peak at the top and a smaller peak at the bottom of the distribution. The bimodal distribution is most prevalent in November-January.

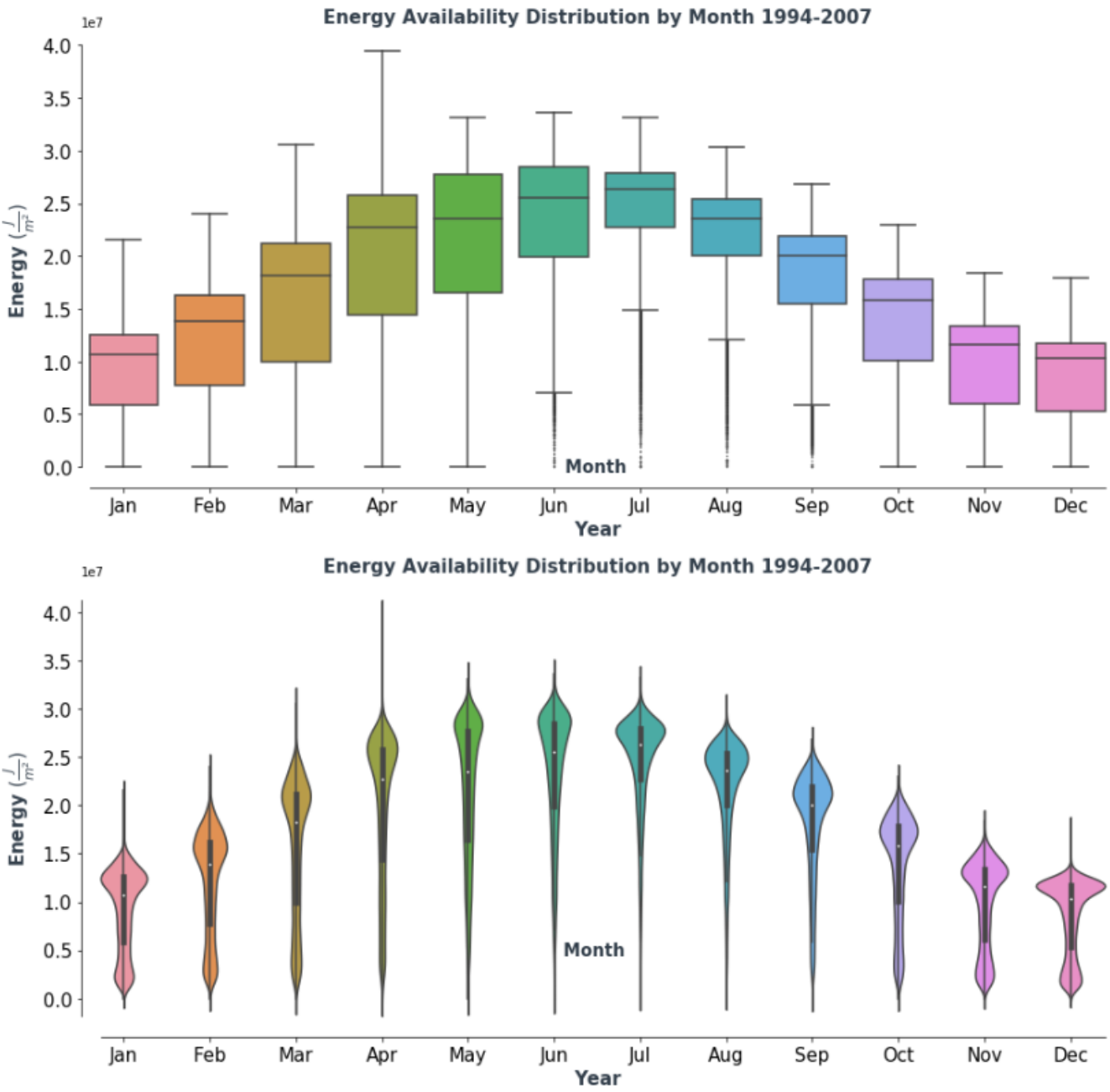

Figure 6: Daily energy availability distributions by month for all stations combined.

The total energy available at each station from 1994-2007 has been plotted on the Oklahoma map to analyze the distribution of energy across the state (Figure 7). There is a clear trend of increasing energy from east to west. As discussed above, the geography (and therefore, likely climate) changes from east to west.

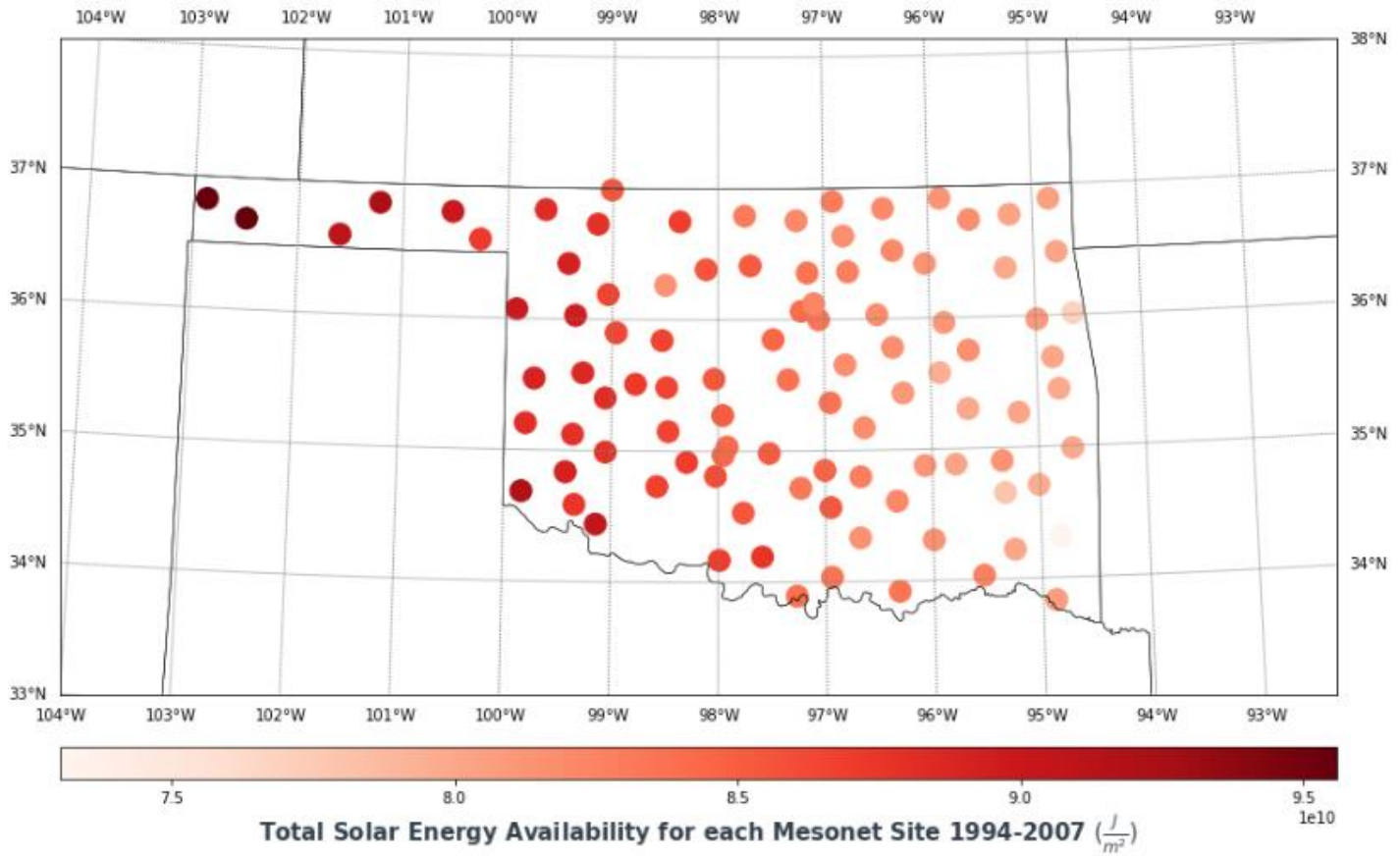

Figure 7: Total solar energy availability for each station.

Stations were sorted into two groups, those east of 98°W, and those west of 98°W. The stations in the west had greater average energy for all years. The differences in total energy and energy distribution between the two groups appears to be consistent for all years.

3.1.4 Weather forecast variables

Weather forecast variables have been visualized by summing or averaging the forecast values from 1994-2007, using the median value of the 11 predictive models, and contouring those values over a map of Oklahoma. On the same plot, the total energy for each station is shown to explore relationships between the weather forecast variables and energy.

The total forecast precipitation contours show that it increases from west to east (Figure 8). The east side of Oklahoma had an order of magnitude greater precipitation forecast than the west side from 1994-2007. Unsurprisingly, the stations to the east had less energy available in the same time frame relative to the stations in the west. The same trend is seen with average forecast cloud cover percentage.

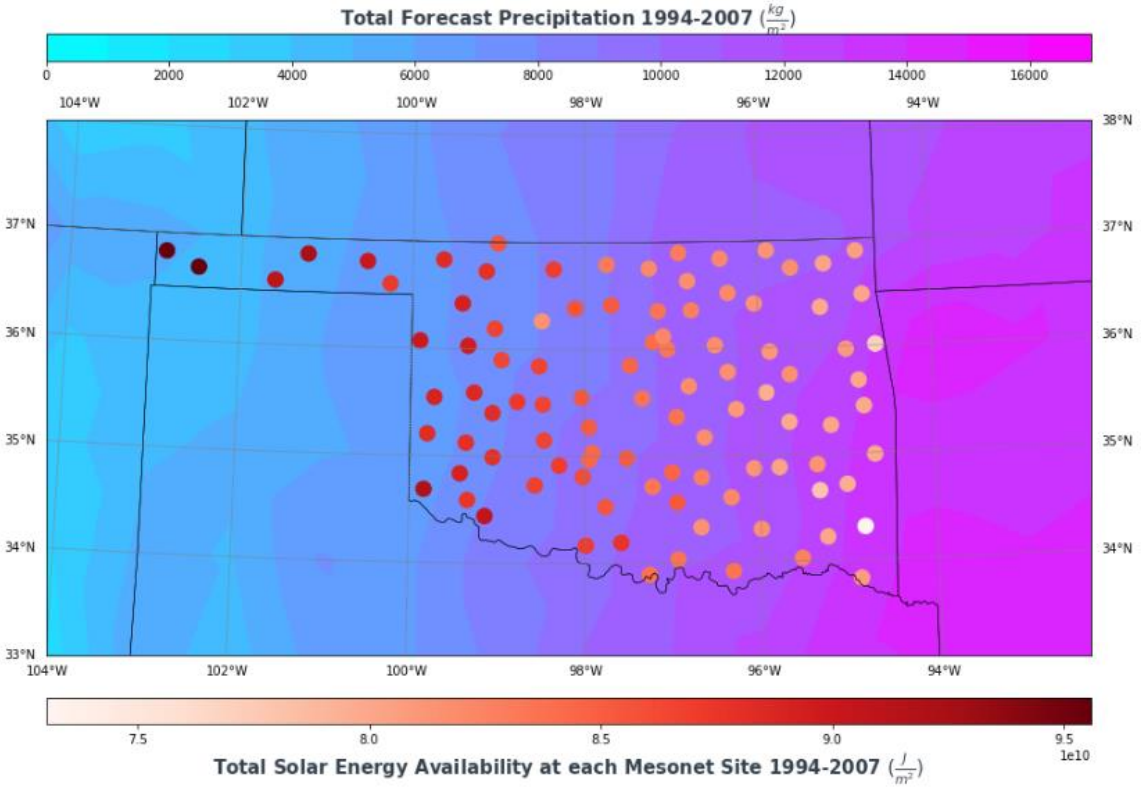

Figure 8: Total forecast precipitation and total solar energy availability for each station.

Total forecast downward short-wave radiative flux at surface generally increases from east to west, with a slight rotation to increase to the south (Figure 9). The radiative flux at surface is likely a function of other weather variables such as precipitation and cloud cover. The radiative flux appears to correlate with energy.

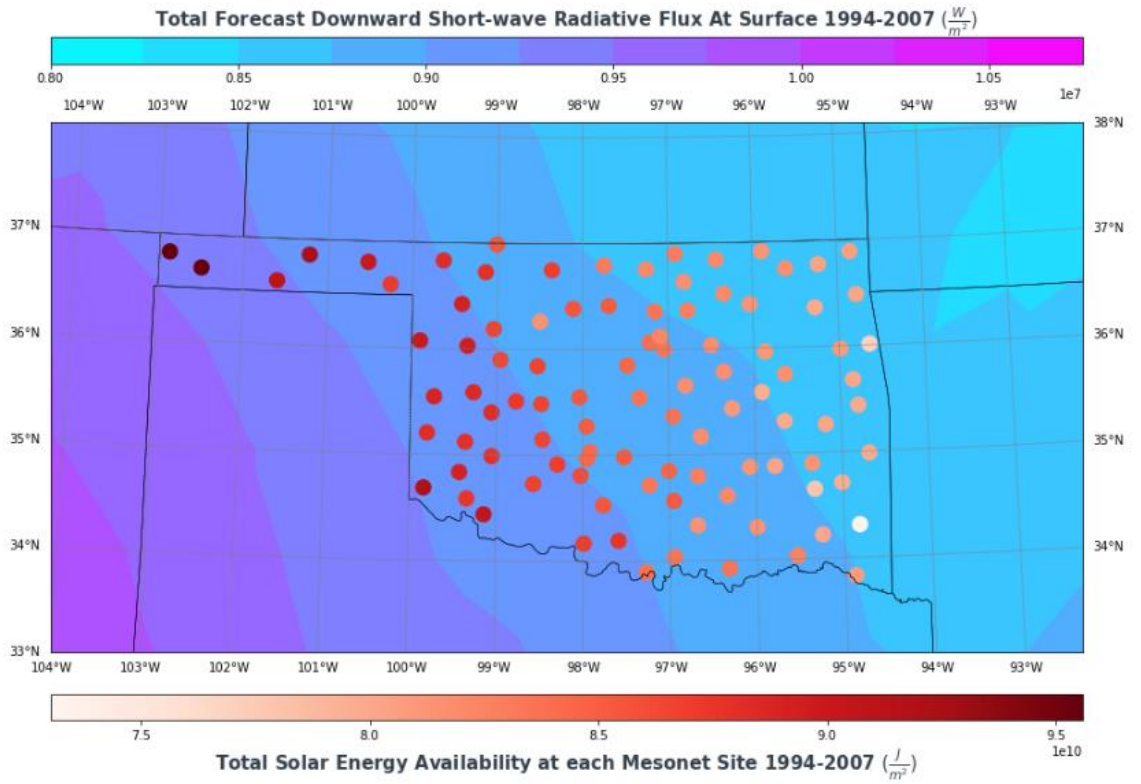

Figure 9: Total forecast short-wave radiative flux and total solar energy availability for each station.

Lastly, average forecast surface temperature increases from north to south (Figure 10). It does not appear to be strongly correlate with energy.

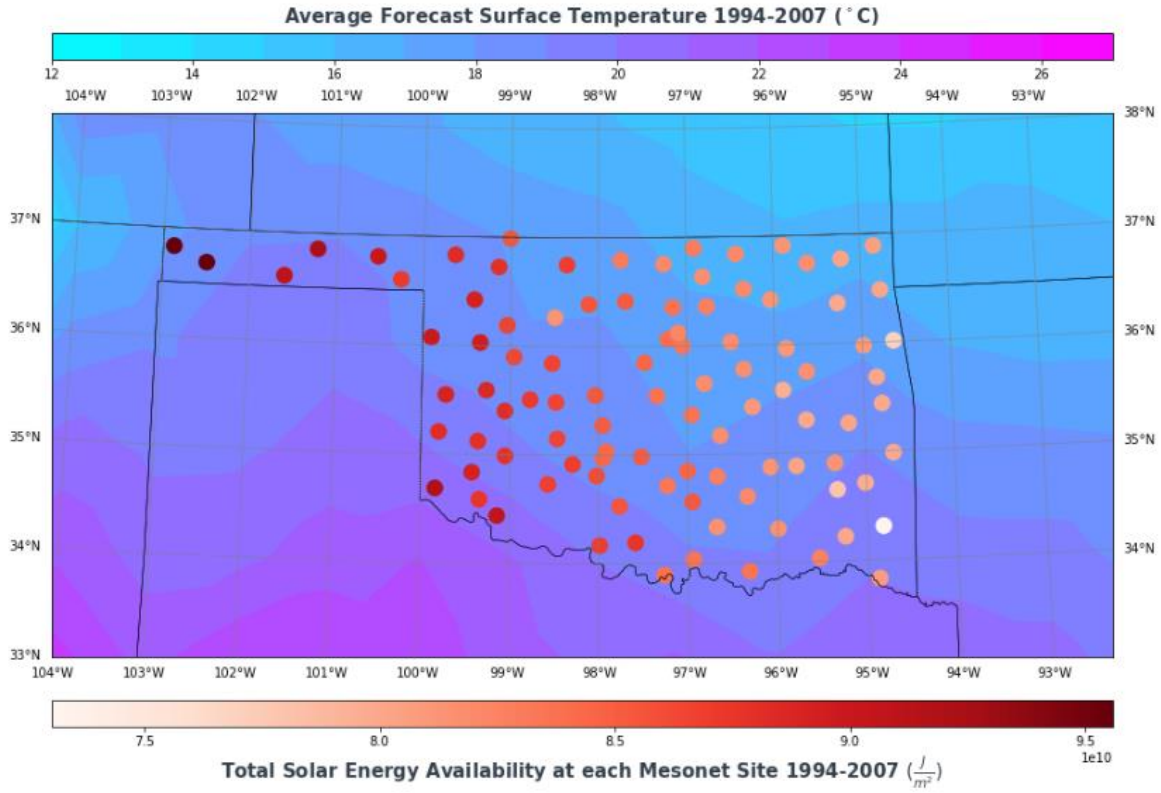

Figure 10: Average forecast temperature and total solar energy availability for each station.

3.2 Statistical data analysis

3.2.1 Bootstrap inference CHER Station July 1994-2007

As shown discussed in the exploratory data analysis section, the energy availability varies greatly by month. It may be desired by a utility company to estimate the daily and monthly expected energy availability at a certain station for long term planning. Bootstrap inference can be used to determine confidence intervals on the mean daily energy availability for a given month. The distribution of daily energy for all days in July from 1994-2007 at CHER Station is shown in Figure 11.

The distribution is certainly non-normal. There are many values concentrated near the maximum of the distribution, and a long tail of values from the concentration down to 0. 10,000 sets of samples of the same size as the distribution have been taken and the mean of those sets of samples calculated as bootstrap replicates.

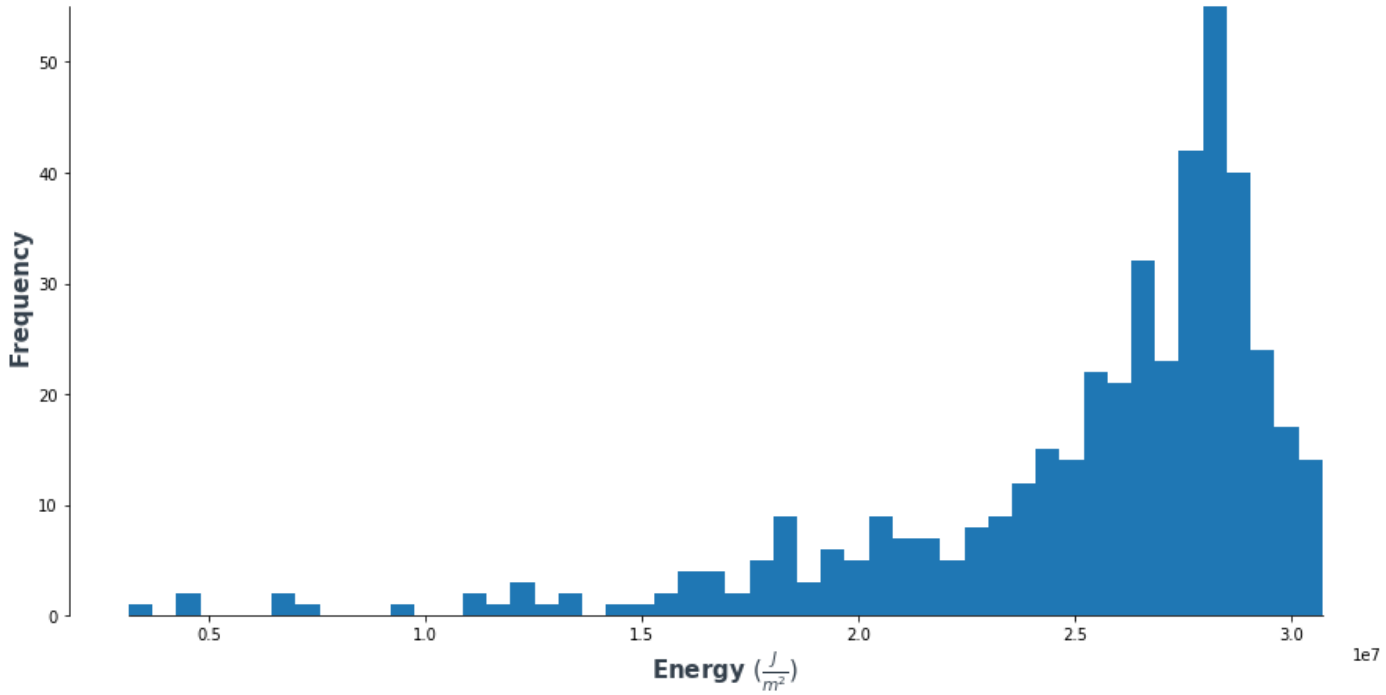

Figure 11: CHER station July daily solar energy availability 1994-2007

The distribution of bootstrap replicates of the mean is shown in Figure 12. The 95% confidence intervals represent the boundaries within which 95% of means will fall when July daily energy values at CHER station are sampled with replacement. The confidence intervals could be used be a utility company when forecasting expected daily values and monthly totals.

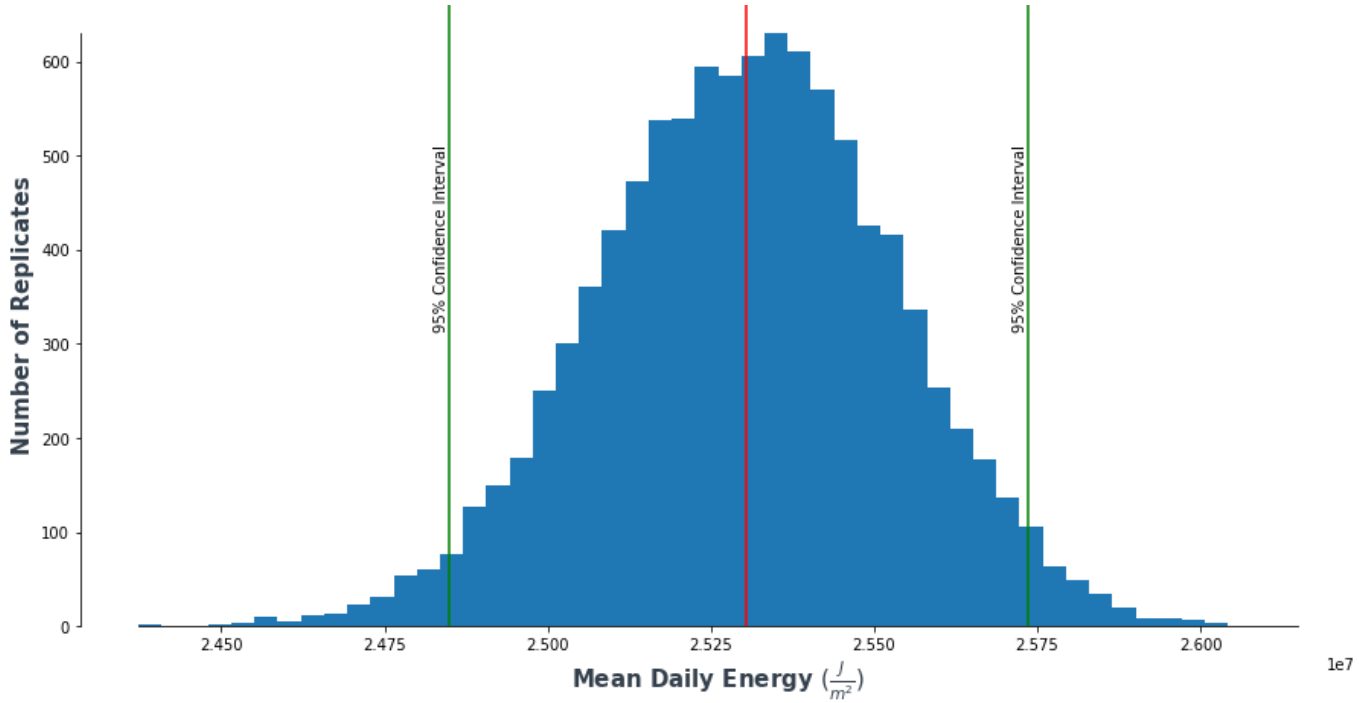

Figure 12: CHER station July daily solar energy availability bootstrap replicates of mean.

3.2.2 Bootstrap inference CHER Station January 1994-2007

The same bootstrap inference performed on the CHER station July energy data has been performed on the CHER station January energy data. The distribution of the data is again non-normal (Figure 13).

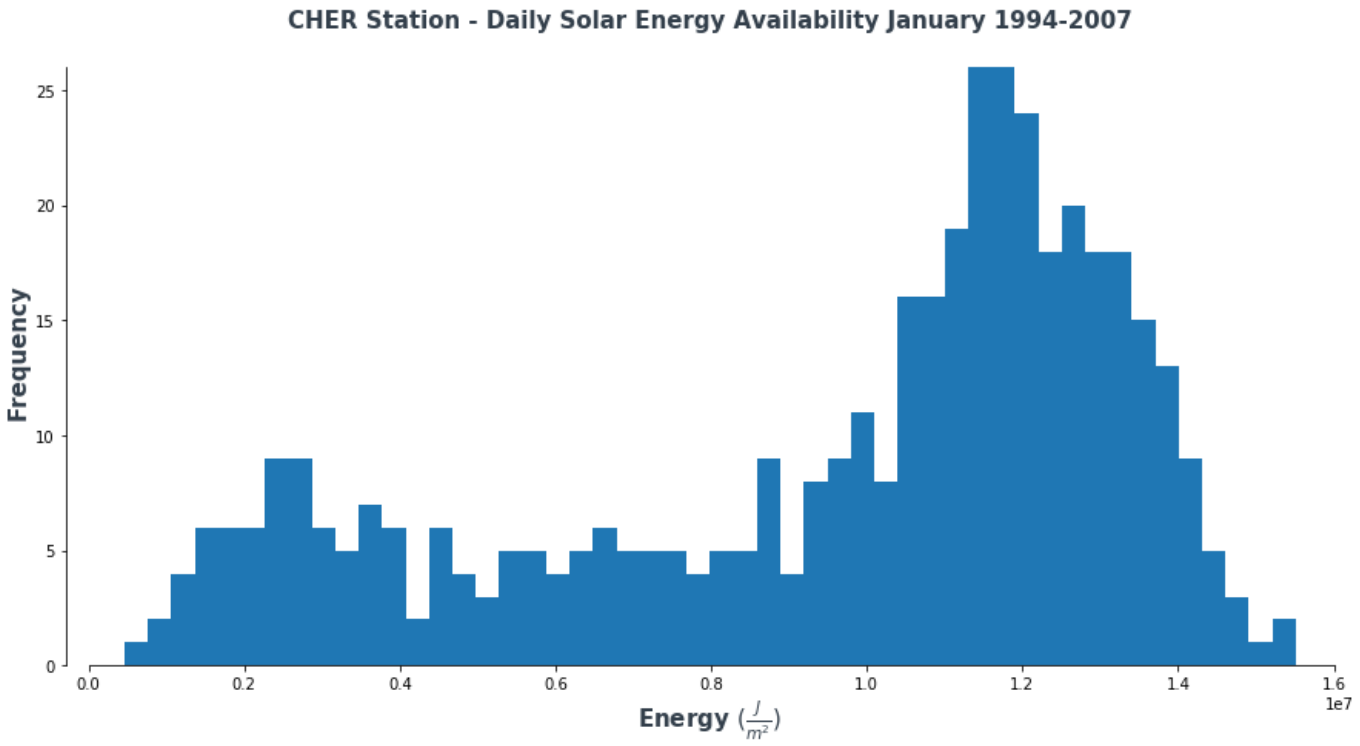

Figure 13: CHER station January daily solar energy availability 1994-2007

Figure 14 compares the bootstrap replicate means of July and January. The July means are much greater than the January means, as expected. Repeating this process for all months of the year would provide a utility company with confidence intervals on the expected values of daily energy for each month. Additionally, an estimate of the population standard deviation can be made, which could provide a utility company with information on which months will have the most variation from the mean.

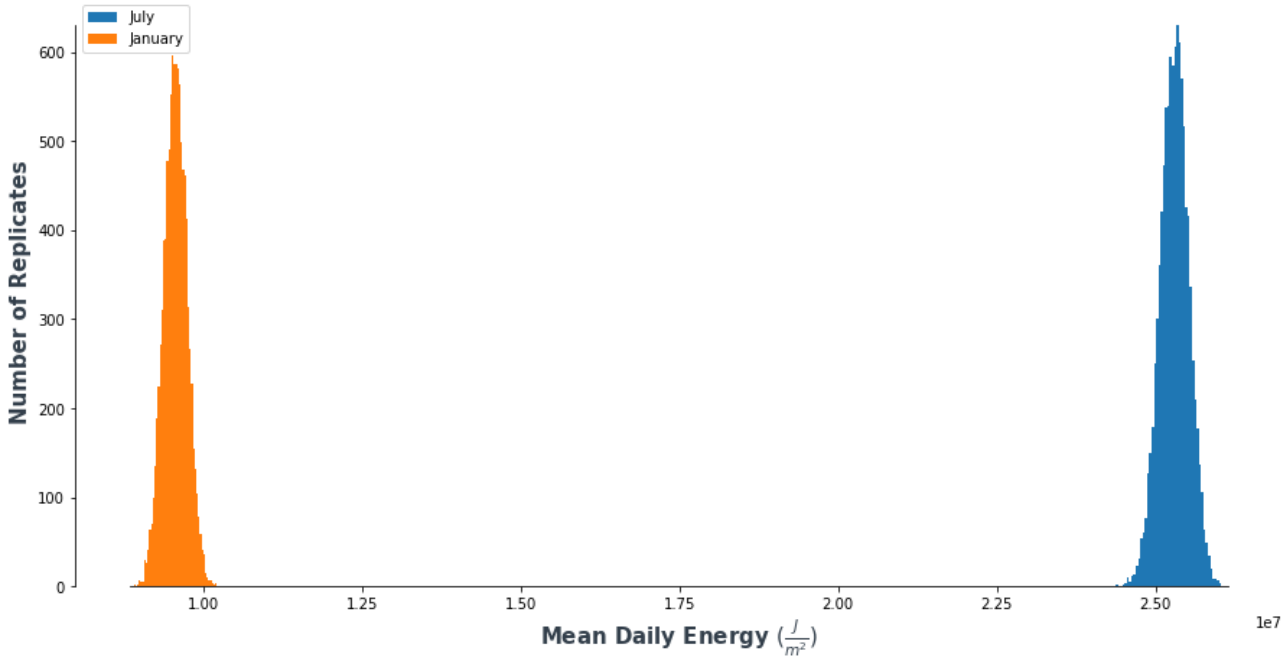

Figure 14: CHER station January and July daily solar energy availability bootstrap replicates of mean.

3.2.3 Bayesian inference CHER Station January 1994-2007

Bayesian inference can be used to model distributions of data and determine constraints on parameters of the model. An attempt at modelling the CHER station January energy data (Figure 13) has been made using Bayesian inference. It is proposed that the data could be constructed by sampling from 3 normal distributions with centers at approximately 0.25e7 J/m^2, 0.7e7 J/m^2, and 1.2e7 J/m^2. Pymc3 has been used to model the assignment of data into 3 normal distributions, and the mean and standard deviation of those distributions.

The parameter traces show good convergence for distributions 1 and 3, however they show poor convergence for distribution 2. This result is understandable, as distribution 2 is in between the other two distributions, and thus many means and standard deviations could be possible. Nonetheless, the distributions of posteriors show that distribution 2 was still delimited.

The posterior-mean parameters have been used to visualize the clusters (Figure 15) and also simulate data (Figure 16). It appears that the posterior-mean parameters construct a model that can simulate the data well. This model could be of use to utility companies looking to determine the frequency of differing amounts of energy availability for January at CHER station.

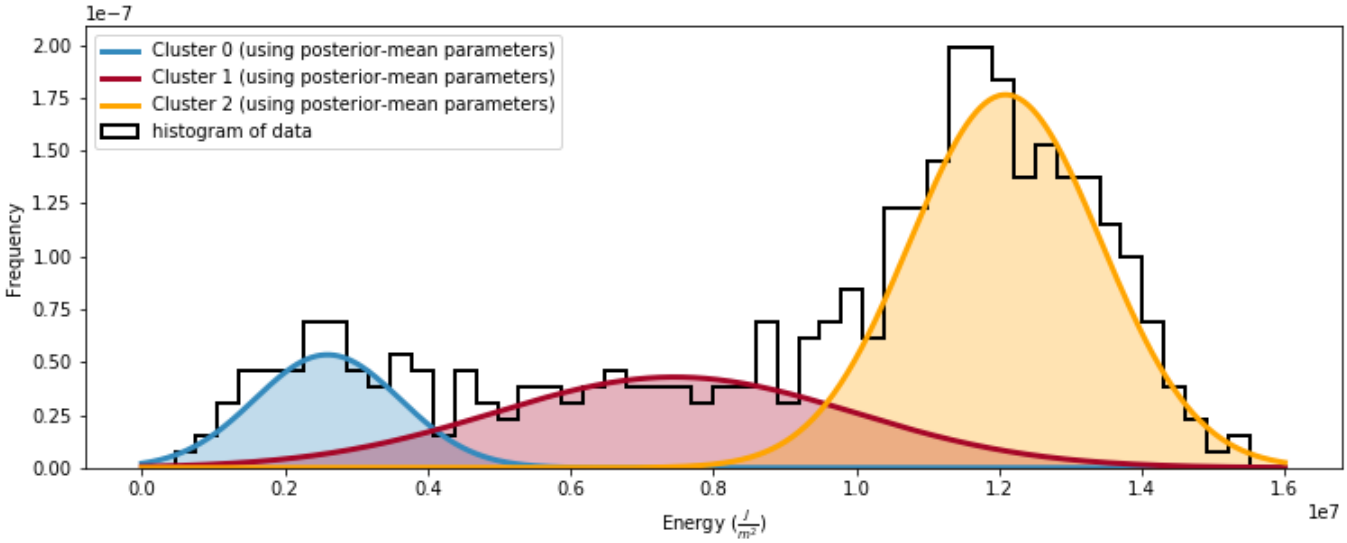

Figure 15: Posterior-mean parameter clusters visualized alongside a histogram of the data.

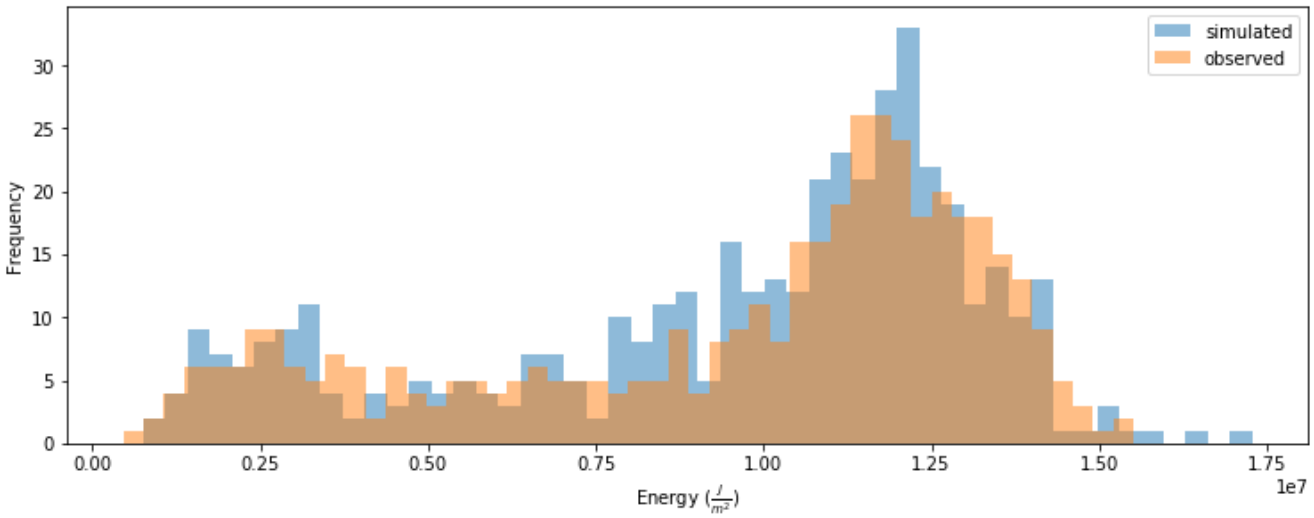

Figure 16: Data simulated using posterior-mean parameters compared with observed data.

4. Machine Learning

4.1 Overview

The goal of the machine learning models is to predict the daily solar energy availability at each of the 98 stations given daily weather forecasts. It is a regression task. Multiple models have been constructed for evaluation and comparison:

1) Ordinary Least Squares (OLS) linear regression
    a. No Feature Selection
    b. Reduced Features
    c. Further Reduced Features
2) Stochastic Gradient Descent (SGD)
3) Gradient Boosting Regressor (GBR)

4.2 Model Evaluation

The data is first split into training and test data – the training data is then further split into 5 contiguous folds for cross-validation (CV). The folds and train/test split are done at year end/beginning to account for autocorrelation in the energy data from year-to-year:

- 1994-1995: Contiguous Fold 1
- 1996-1997: Contiguous Fold 2
- 1998-1999: Contiguous Fold 3
- 2000-2001: Contiguous Fold 4
- 2002-2003: Contiguous Fold 5
- 2004-2007: Test Data

Models will be evaluated using the Mean Absolute Error (MAE) which is commonly used in the energy industry to measure prediction accuracy according to the client. First, an average CV MAE will be calculated by running each of the 5 contiguous folds as the CV test data, leaving the remaining 4 folds as the CV train data. Next, all 5 contiguous folds are used together as the training data, and the 2004-2007 test data is used as a final validation (evaluated using MAE). Additionally, Adjusted Pearson's Correlation Coefficient (Adj. R^2) will be used for model evaluation.

4.3 Linear Regression (OLS)

4.3.1 Model Overview

Linear regression fits a hyperplane to a set of points in N-dimensions (N = number of features). It generally requires the following assumptions:

- Linear relationship between the target variable and features exists.
- Data should not exhibit multicollinearity.
- Homoscedasticity (constant variance) of residuals.
- Normally distributed residuals.

4.3.2 OLS Model 1a – No Feature Selection

The baseline model will use no feature selection. Each of the 15 weather forecast variables at 5 forecast hours are used as 75 different features. Additionally, latitude, longitude, elevation, and the month of the year are used as features to total 79 features. The baseline model evaluation criteria scores are given in Table 1.

Table 1: OLS Model 1a – No Feature Selection Evaluation Criteria Scores

Evaluation Criteria	Score
CV MAE (W/m^2)	2,353,201
Test MAE (W/m^2)	2,195,538
Test Adj. Data R^2	0.915

Assumption #1: Linear relationship between the target variable and features exists.

The coefficients of the regression are non-zero and thus show that there are linear relationships between the target variable and the features. The p-value of the t-test scores are close to 0, indicating that it is very likely that these features have predictive power. The relationship between a single feature (current temperature at 2m above the ground) and energy is shown in Figure 17. The data can by modelled by a linear relationship but would have significant and biased residuals.

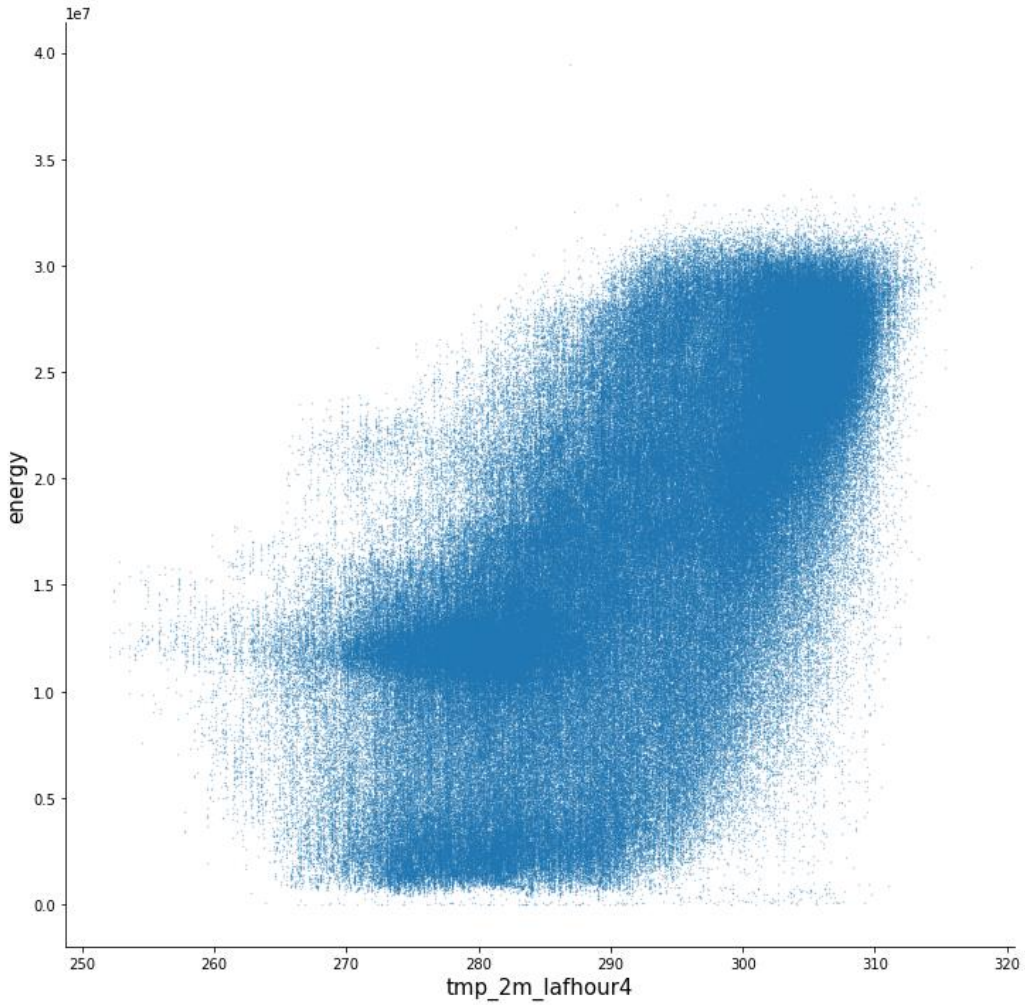

Figure 17: Current temperature at 2m above the ground forecasts vs energy.

Assumption #2: Data should not exhibit multicollinearity.

Features highly correlated with other features can cause an increase in the standard errors of the coefficients. In turn, coefficients for other features may not be found to be significantly different from 0, and thus they will be labelled statistically insignificant. Multicollinearity can be tested using a correlation matrix (calculating Pearson's Correlation Coefficient between each set of features and between each feature and the target variable). To start, only forecast hour 0 will be checked for multicollinearity between features – the correlation matrix is shown in Figure 18.

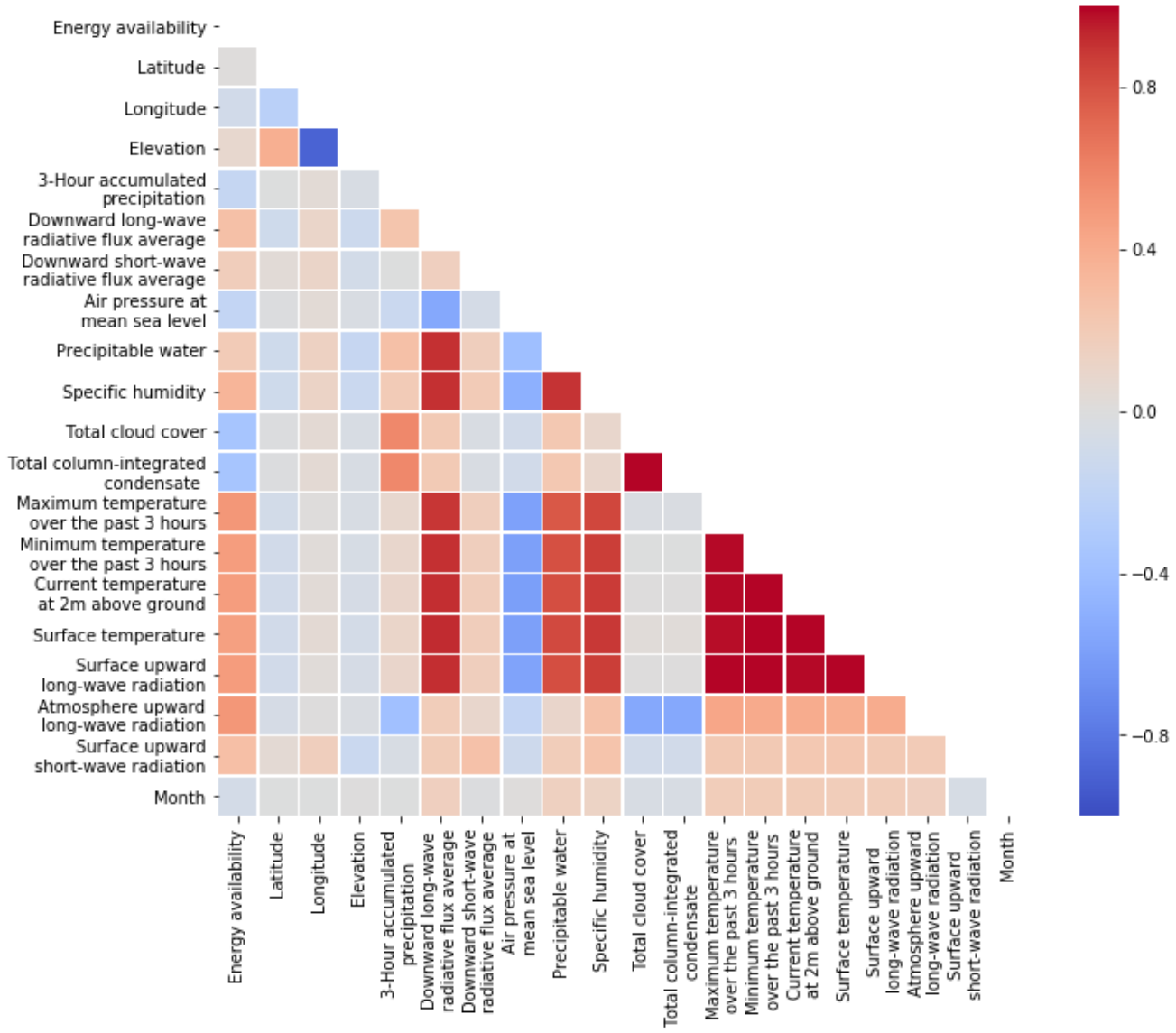

Figure 18: Correlation Matrix for all features (weather variables forecast hour 0 only)

Many features exhibit multicollinearity. A reasonable cut-off for labelling a correlation 'strong' is $|R^2| > 0.7$. When choosing which of a set of features to remove, it makes sense to remove the feature that has a weaker correlation with the target variable, energy. This logic has been used to determine that the following features should be removed from the data to lower the risk of multicollinearity causing issues with the linear regression:

- Elevation
- Downward long-wave radiative flux average at the surface (all forecast hours)
- Precipitable Water over the entire depth of the atmosphere (all forecast hours)
- Specific Humidity at 2 m above ground (all forecast hours)
- Total column-integrated condensate over the entire atmosphere (all forecast hours)
- Minimum Temperature over the past 3 hours at 2 m above the ground (all forecast hours)
- Current temperature at 2 m above the ground (all forecast hours)
- Temperature of the surface (all forecast hours)
- Upward long-wave radiation at the surface (all forecast hours)

Assumption #3: Homoscedasticity (constant variance) of residuals.

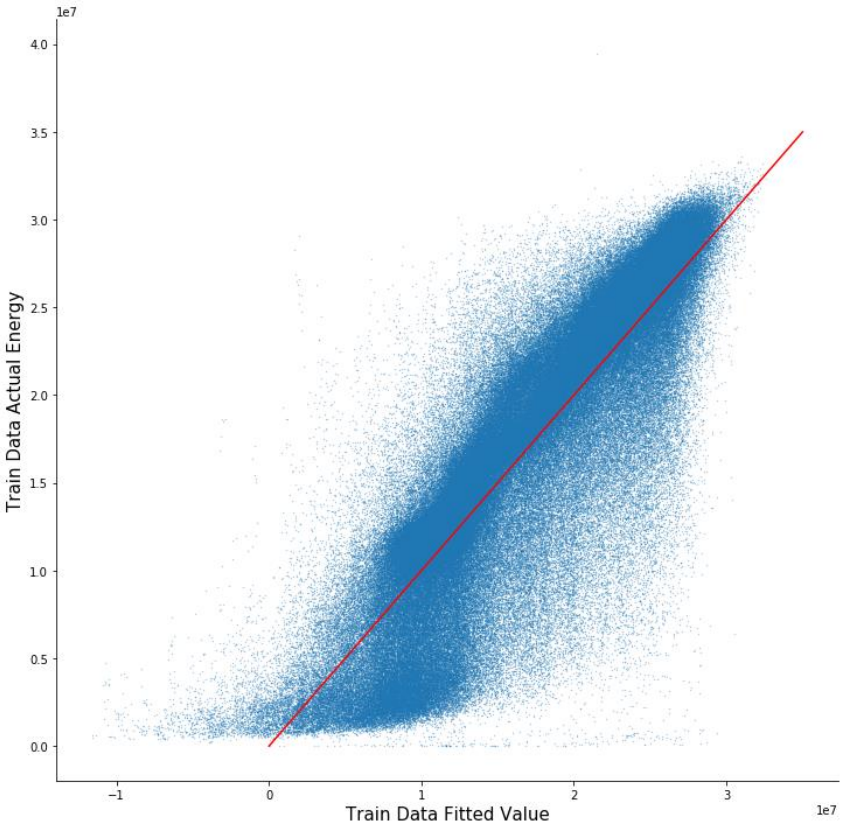

Figure 19: Model 1a train data fitted vs actual energy values.

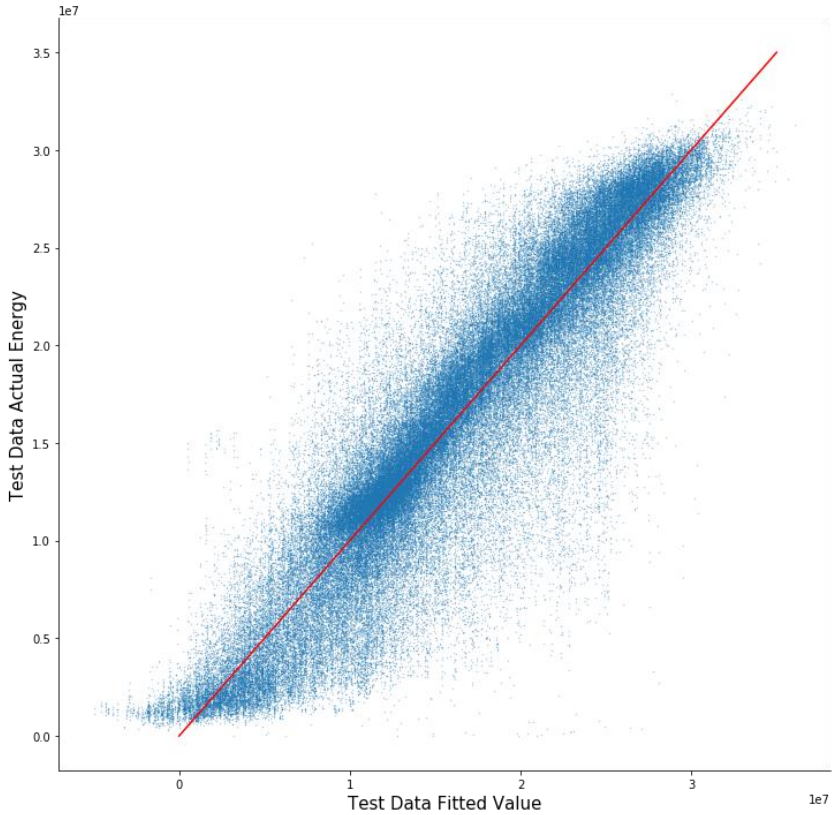

Figure 20: Model 1a test data fitted vs actual energy values.

Visually, it can be seen in Figure 19 and Figure 20 that the residuals are not distributed evenly along the line of a theoretical perfect prediction. Low actual energy values are over-predicted while higher actual energy values are under predicted. Statistically, homoscedasticity can be tested using the Breusch-Pagan (BP) test, which takes the residuals and features as inputs. The null hypothesis of this test is that there is no violation of homoscedasticity. The BP test gave a p-value of near 0, indicating the null hypothesis is rejected, and there is likely a violation of homoscedasticity, confirming what was seen visually.

Assumption #4: Normally distributed residuals.

A histogram of residual values is shown in Figure 21. The distribution appears somewhat normally distributed, but it appears that the left side has more mass than the right side. The Anderson-Darling (AD) Normality test has the null hypothesis that the data is normally distributed. The p-value of the test statistic is near 0, thus we reject the hypothesis that the residuals are normally distributed.

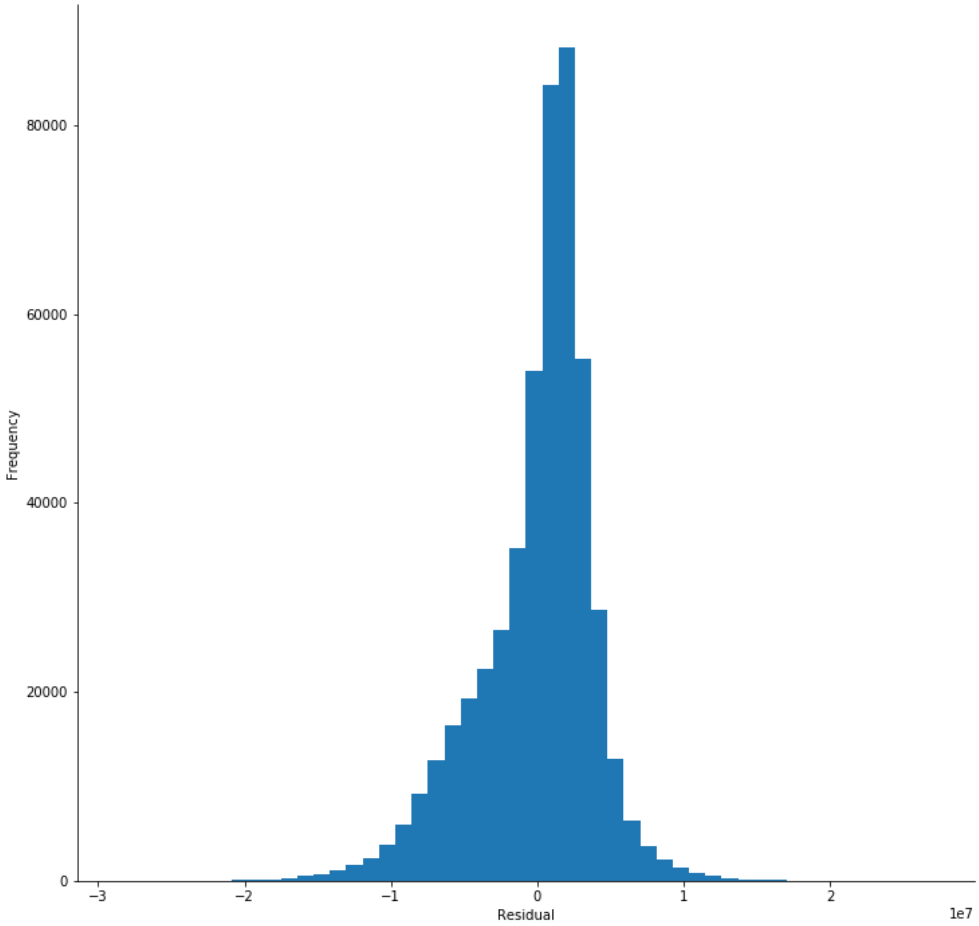

Figure 21: Model 1a residual histogram

Non-normality of residuals can also be visualized using Quantile-Quantile (Q-Q) plots. Ideally, the Q-Q trace will be along the red line (Figure 22). The trace clearly strays from the red line at the tails, indicating the linear regression is a poor fit at low and high values of energy.

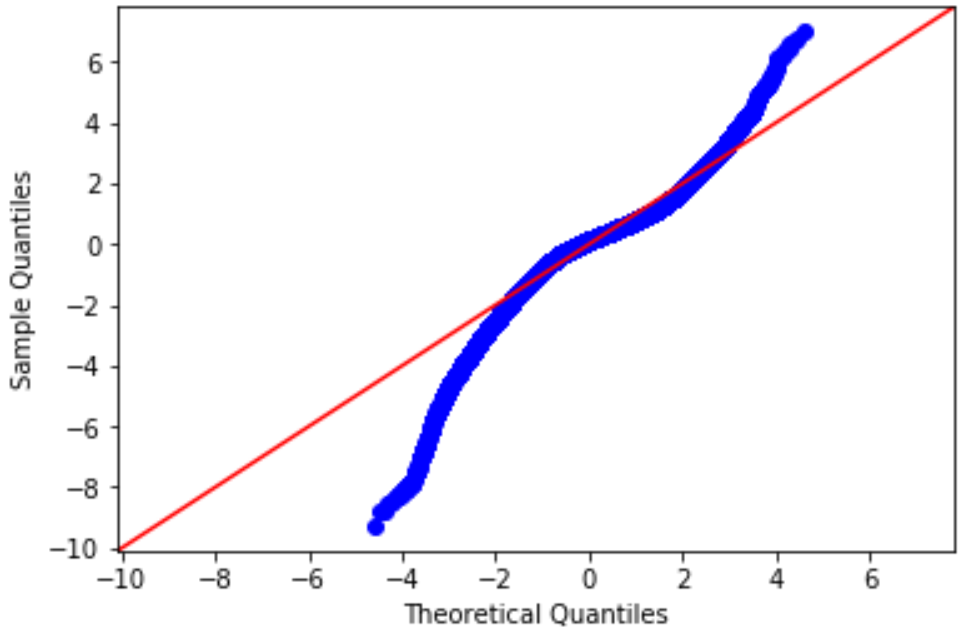

Figure 22: Model 1a Quantile-Quantile plot

4.3.3 OLS Model 1b – Reduced Features

The features identified for removal in model 1a discussion to reduce multicollinearity have been removed to construct model 1b. The model evaluation criteria scores are given in Table 2. CV MAE and test MAE are both greater than for model 1a. Additionally, test data adj. R^2 has increased.

Table 2: OLS Model 1b – Reduced Features Evaluation Criteria Scores

Evaluation Criteria	Score
CV MAE (W/m^2)	2,552,359
Test MAE (W/m^2)	2,359,393
Test Adj. Data R^2	0.901

This model exhibits the same violations of the basic assumptions as model 1a. Homoscedasticity and normally of residuals are both violated. Additionally, there is still multicollinearity present, as only forecast hour 0 was checked previously. A correlation matrix for all features in model 1b is shown in Figure 23. This correlation matrix allows us to identify further features for removal from the model to account for multicollinearity.

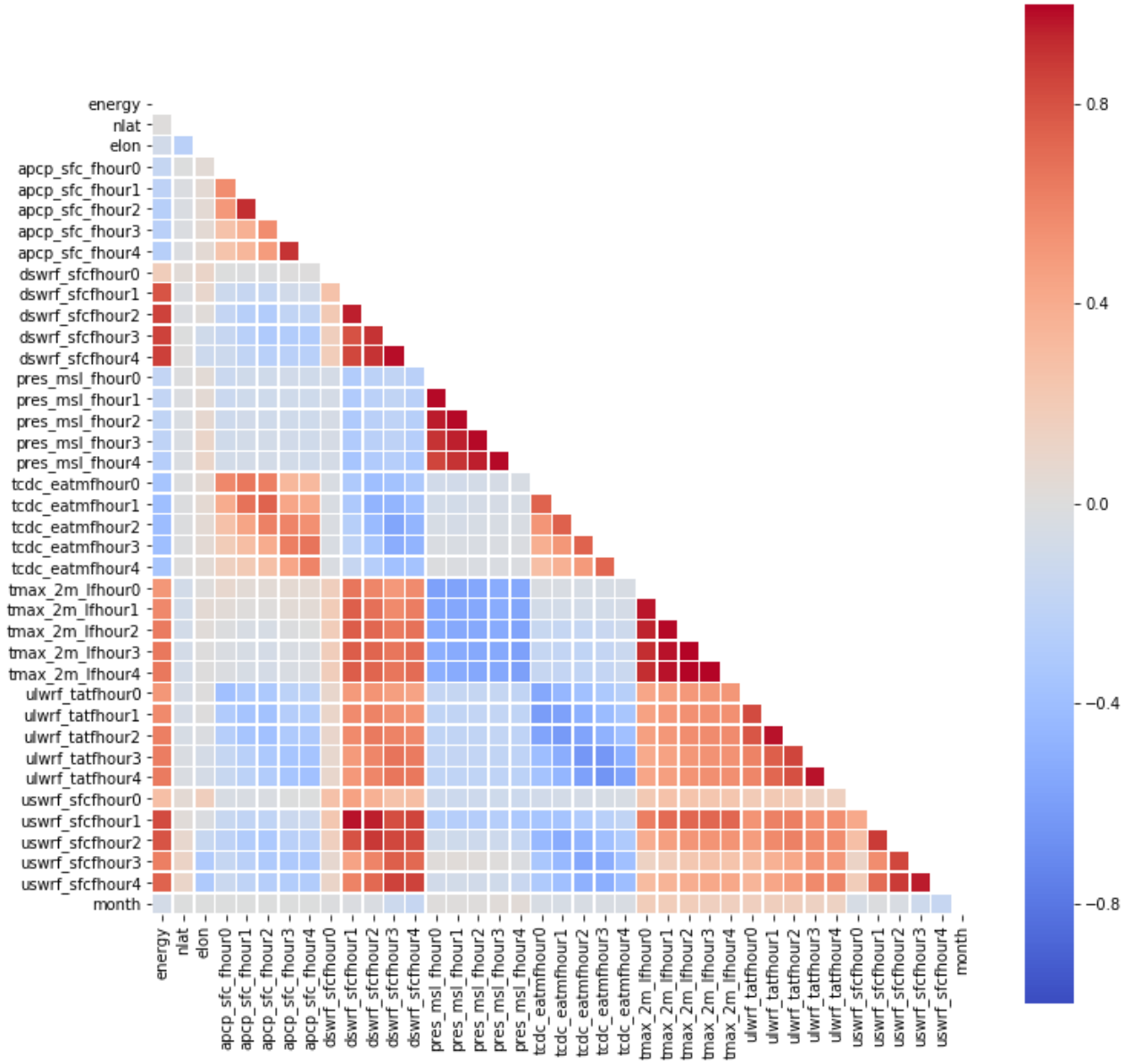

Figure 23: Correlation matrix for all features in model 1b.

4.3.4 OLS Model 1c – Further Reduced Features

The features identified for removal in model 1b discussion to reduce multicollinearity have been removed to construct model 1c. The model evaluation criteria scores are given in Table 3. CV MAE and test MAE are both greater than for model 1a and model 1b. Additionally, test data adj. R^2 has increased.

Table 3: OLS Model 1c – Further Reduced Features Evaluation Criteria Scores

Evaluation Criteria	Score
CV MAE (W/m^2)	3,021,995
Test MAE (W/m^2)	2,844,458
Test Adj. Data R^2	0.867

This model again violates homoscedasticity and normality of residuals, although there are no longer any ‘strong’ correlations between features.

4.3.5 OLS Model Comparison

Homoscedasticity and normality of the residuals could not be achieved by any OLS model. When multicollinearity was reduced, the predictive power of the model also decreased. This result implies that multicollinearity was not a problem for this dataset. Overall, the linear regression models do have predictive power, with model 1a having a test MAE of 2,195,538 (W/m^2). The mean daily energy is 16,567,964 (W/m^2), resulting in a test MAE that is ~13.3% of the mean daily energy. The models all have decreased prediction power when predicting at very low and high values.

4.4 Stochastic Gradient Descent

4.4.1 Model Overview

Stochastic Gradient Descent (SGD) is an iterative process that can be used to optimize the coefficients of a linear regression by minimizing a cost function. 'Stochastic' means that the process is linked with random probability – samples of the dataset are randomly selected for each iteration, as opposed to using the entire dataset, reducing computational requirements. SGD is sensitive to feature scaling – the constructed model uses a pipeline to scale the features prior to fitting.

4.4.2 Hyperparameter tuning

SGD models have several hyperparameters to be tuned. For this model, the following hyperparameters have been tuned using GridSearchCV - the resulting best parameters are also listed:

- Alpha (0.00001)
- Max iterations (60)
- Epsilon (0.1)
- Loss function (sum of squared differences)

There are additional hyperparameters that could be tuned, but for the purposes of this project, only those listed above were tuned – the remaining parameters have been left as the default values.

4.4.3 Results

The model performs very similarly to the simple OLS model (1a) with the same features. Again, the homoscedasticity and normality of residuals have been violated.

Table 4: SGD Model Evaluation Criteria Scores.

Evaluation Criteria	Score
CV MAE (W/m^2)	2,360,498
Test MAE (W/m^2)	2,204,851
Test Adj. Data R^2	0.914

4.5 Gradient Boosting Regressor

4.5.1 Model Overview

Gradient Boosting Regression (GBR) is an ensemble prediction process (i.e., it uses a collection of predictors to give a final prediction). The aim of this process is to reduce noise, variance, and bias. The process uses boosting, meaning the ensemble of predictors are made sequentially – observations with higher errors from previous predictors are more likely to appear in subsequent predictors. Each predictor is a regression tree.

4.5.2 Hyperparameter tuning

GBR models have several hyperparameters to be tuned. For this model, the following hyperparameters have been tuned using GridSearchCV - the resulting best parameters are also listed:

- Number of estimators (1000)

- Max features (10)
- Max depth (6)

The loss function used was least absolute distance. A subsample of 0.5 was used for each estimator. There are additional hyperparameters that could be tuned, but for the purposes of this project, only those listed above were tuned – the remaining parameters have been left as the default values. Computation time for fitting GBR models with many datapoints is high – the grid search was performed on an Amazon Web Service EC2 instance.

4.5.3 Results

The GBR model performs better than all other models. Again, the homoscedasticity and normality of residuals have been violated.

Table 5: GBR Model Evaluation Criteria Scores.

Evaluation Criteria	Score
CV MAE (W/m^2)	2,123,177
Test MAE (W/m^2)	1,975,215
Test Adj. Data R^2	0.924

4.6 Model Comparison

The three best-performing models have their evaluation criteria compared in Table 6. GBR has the lowest CV and Test MAE values, and a highest adj. R^2. Additionally, Figure 24 and Figure 25 show that the residuals of the GBR model have a much more even distribution. Figure 26 and compares distributions of fitted and actual values for each of the models. The GBR model produces values more like the actual data than the OLS model.

Table 6: Model evaluation criteria comparison.

	OLS Model 1a	SGD	GBR
CV MAE (W/m^2)	2,353,201	2,360,498	2,123,177
Test MAE (W/m^2)	2,203,998	2,204,851	1,975,215
Test Adj. Data R^2	0.915	0.914	0.924

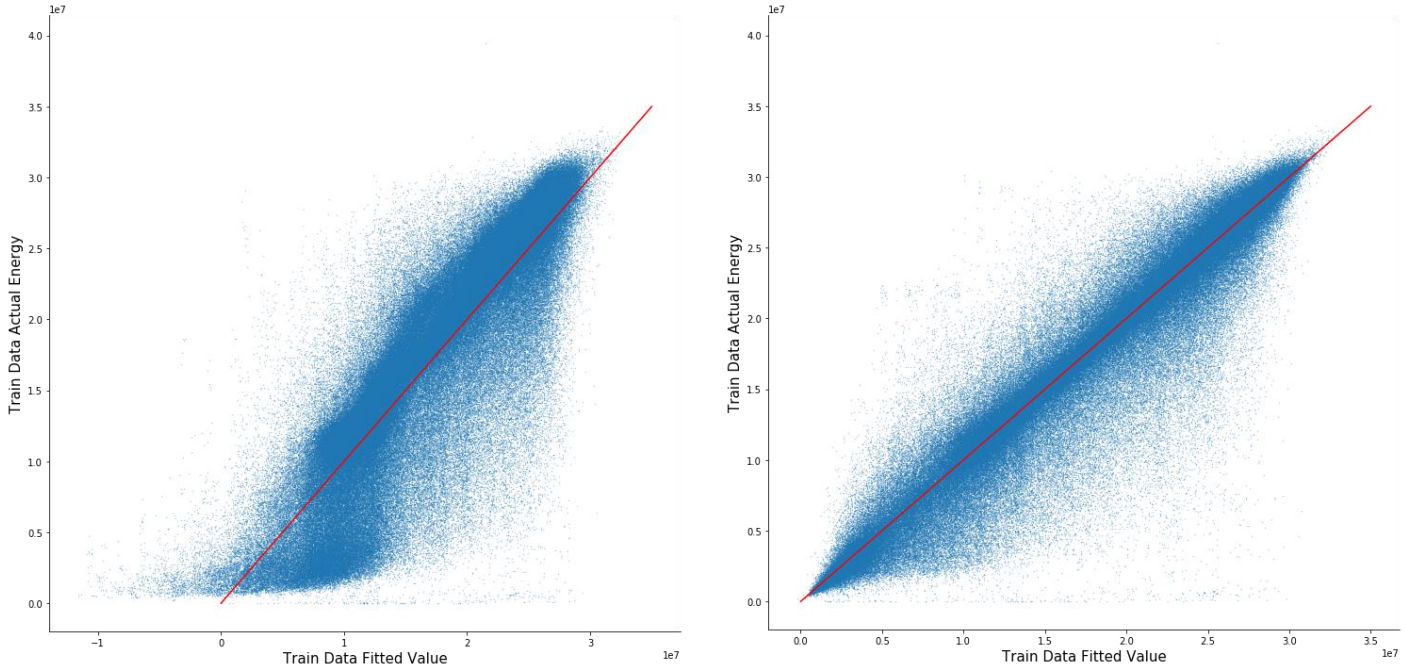

Figure 24: Fitted vs actual energy value scatterplots for (left) OLS model 1a and (right) GBR model (train data)

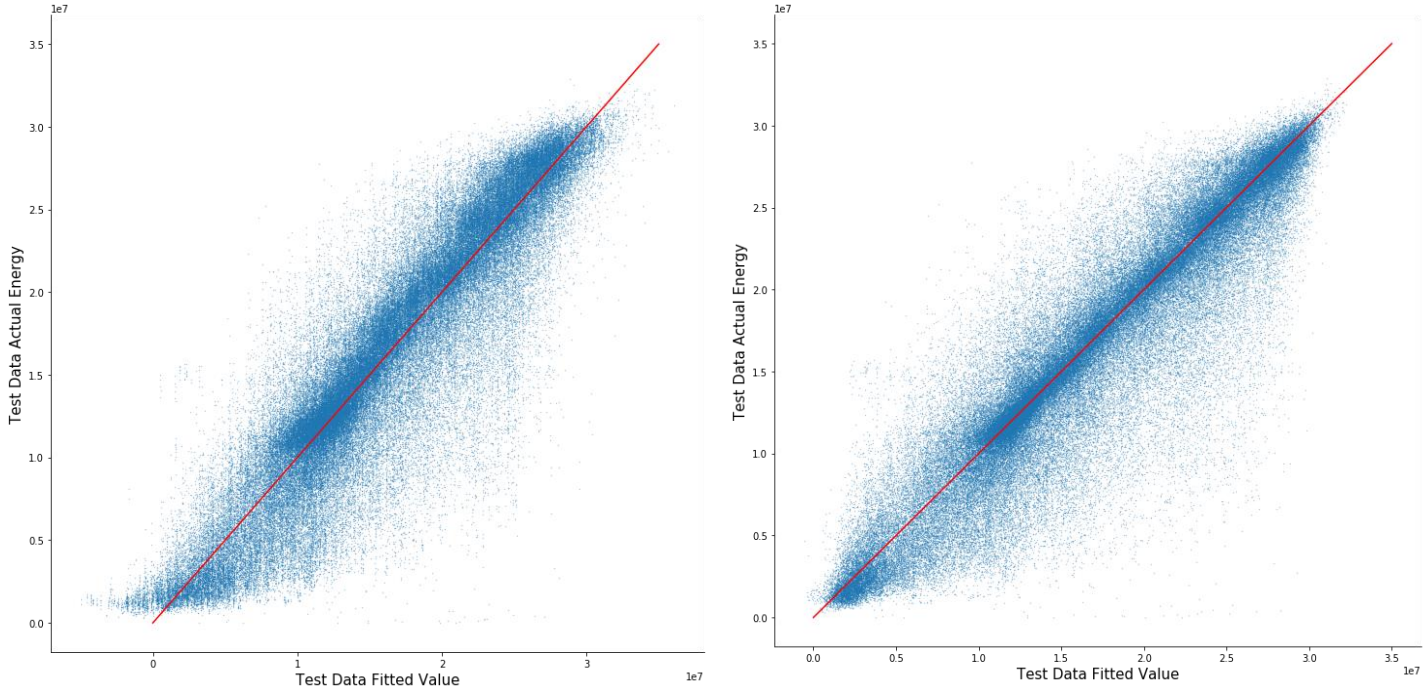

Figure 25: Fitted vs actual energy value scatterplots for (left) OLS model 1a and (right) GBR model (test data)

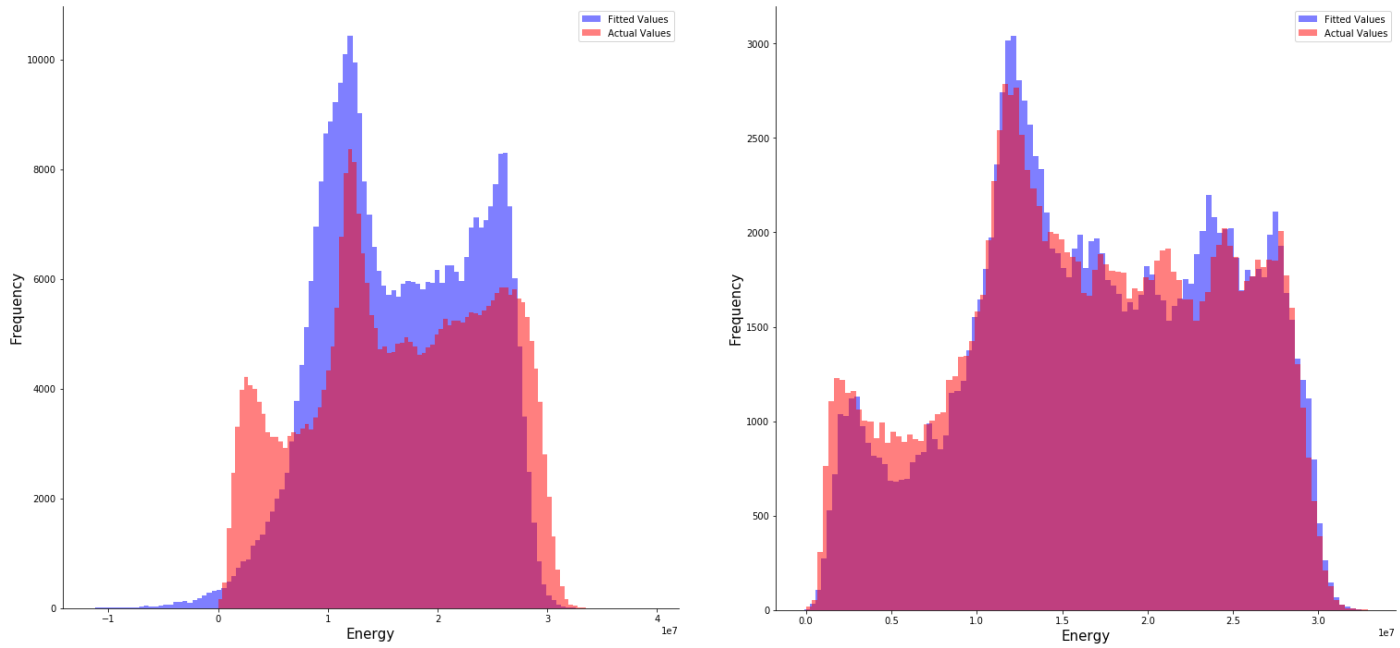

Figure 26: Fitted and actual energy value histograms for (left) OLS model 1a and (right) GBR model

4.7 Conclusions and Further Work

Linear regression resulted in models with predictive power but many shortcomings. Low values were consistently over-estimated while higher values were consistently under-estimated. The residuals therefore violate homoscedasticity and normality, indicating that linear regression may not be the best choice of model for this dataset. Additionally, the linear regression models predict negative values, as an implicit artifact of the method.

The GBR model performed better than the linear regression models according to visual assessment of the data and the evaluation criteria of CV MAE, Test Data MAE, and Test Data adj. R^2. The GBR model was able to predict on a test dataset (separate from fitting and hyperparameter tuning) the total daily energy with a mean absolute error of 1,975,215 (W/m^2), which is approximately 11.9% of the mean total daily energy value.

Further work on this project would be to spend more time tuning the GBR hyperparameters using randomized and grid searches. Additionally, it would be interesting to use all 11 of the weather forecast models to train 11 regression models and average the predictions. Lastly, it could be explored to use a distance-weighted average of weather forecasts (at multiple grid points) as the weather forecast variable features.

5. References

[1] Forecast International - Powerweb , "Renewable Energy," 2016. [Online]. Available: http://www.fi-powerweb.com/Renewable-Energy.html.

[2] Live Science, "How Do Solar Panels Work?," Live Science, 2017. [Online]. Available: https://www.livescience.com/41995-how-do-solar-panels-work.html. [Accessed 2019].

[3] Hukseflux, "Pryanometers," Hukseflux, 2019. [Online]. Available: https://www.hukseflux.com/products/solar-radiation-sensors/pyranometers. [Accessed 2019].

[4] J. Lago, "Forecasting in the Electrical Grid," Incite, 2018. [Online]. Available: http://www.incite-itn.eu/blog/forecasting-in-the-electrical-grid/. [Accessed 2019].

[5] http://news.mit.edu/2016/mit-neutralize-17-percent-carbon-emissions-through-purchase-solar-energy-1019, "MIT to neutralize 17 percent of carbon emissions through purchase of solar energy," MIT News, 2016. [Online]. Available: http://news.mit.edu/2016/mit-neutralize-17-percent-carbon-emissions-through-purchase-solar-energy-1019. [Accessed 2019].

[6] J. Slingo and T. Palmer, "Uncertainty in weather and climate prediction," *Philos Trans A Math Phys Eng Sci,* vol. 369, no. 1956, p. 4751–4767, 2011.