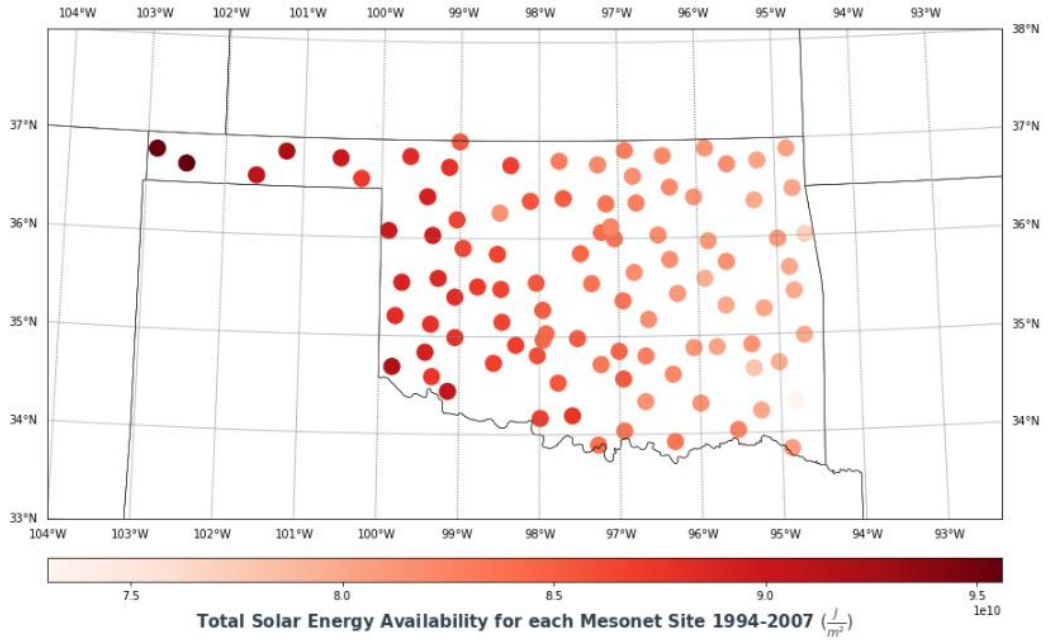
Predicting Short Term Solar Energy Production

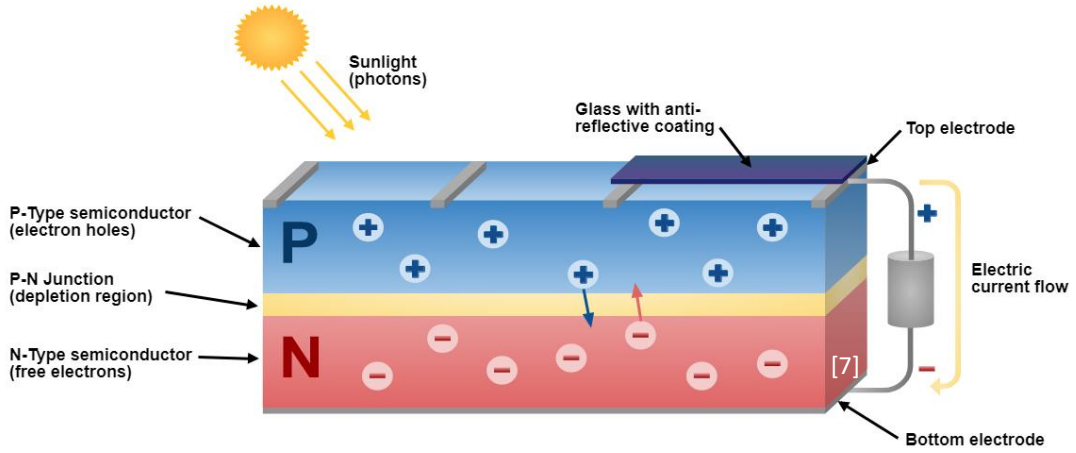

[6]



CONNOR MCANUFF - SEPTEMBER 2019

Project Overview – Solar Energy

- Global solar energy generation capacity increasing exponentially since 2006 [1]
- A form of solar energy generation is using photovoltaic (PV) panels which produce electricity through interaction with photons from the sun [2]



Project Overview – Solar Energy Availability

- Solar energy availability at a given location can be described by the level of solar irradiance (W/m^2)
- Solar irradiance can be measured using a pyranometer [3]
- **Solar irradiance is used as a measure of the energy available to enter a solar PV system [4]**

Project Overview – Predicting solar energy generation

- The system configuration and energy losses through a PV system are well known
- **Thus, if the solar irradiance is known, solar energy production can be accurately determined**
- **i.e. Predicting solar energy production can be done by proxy by predicting solar energy availability**
- Solar energy availability is in part determined by weather conditions [4]

Project Overview – Value to Client

- Utility companies sourcing solar energy are presented with the unique challenge of variable input to the grid as solar energy generation varies in part depending on weather conditions
- **Energy generation must be accurately forecasted to prevent energy shortages and surpluses**
- Utility companies must be able to make informed grid-balancing decisions to minimize costs and operational complications

Data Wrangling – Data Overview

Raw data has been provided in the following format (source – Kaggle [5]):

1) station_info.csv
   - Array of station ID, latitude, longitude, elevation
   - 98 rows (stations) x 4 columns

- 98 Mesonet weather monitoring stations are acting as solar farm stations spread across U.S. state of Oklahoma
- **The project objective is to predict the solar energy availability daily at each of these 98 stations given daily weather forecasts**

Data Wrangling – Data Overview (2)

Raw data has been provided in the following format (source – Kaggle):

2) train.csv
   - Array of dates and recorded total daily solar energy at each of the 98 Mesonet Solar Farms from 1994-01-01 to 2007-12-31
   - 5113 rows (days) x 98 columns (stations)

- **The total daily solar energy is the target variable for the machine learning models**

Data Wrangling – Data Overview (3)

Raw data has been provided in the following format (source – Kaggle):

3) Weather forecast variables
   - 15 NETCDF4 files (one for each weather variable) listing the forecast value for each of the 11 predictive models, 5 daily forecast hours, 9 latitudes, and 16 longitudes for each of the 5113 forecast days
   - Dimensions 11 x 5 x 9 x 16 x 5113 (x15 files)

- **The weather forecast variables are used as machine learning features i.e. the predictors of solar energy availability**

Data Wrangling – Data Overview (4)

Raw data has been provided in the following format (source – Kaggle):

3) Weather forecast variables:

3-Hour accumulated precipitation at the surface	Maximum Temperature over the past 3 hours at 2 m above the ground
Downward long-wave radiative flux average at the surface	Minimum Temperature over the past 3 hours at 2 m above the ground
Downward short-wave radiative flux average at the surface	Current temperature at 2 m above the ground
Air pressure at mean sea level	Temperature of the surface
Precipitable Water over the entire depth of the atmosphere	Upward long-wave radiation at the surface
Specific Humidity at 2 m above ground	Upward long-wave radiation at the top of the atmosphere
Total cloud cover over the entire depth of the atmosphere	Upward short-wave radiation at the surface
Total column-integrated condensate over the entire atmosphere	

Data Wrangling – Cleaning

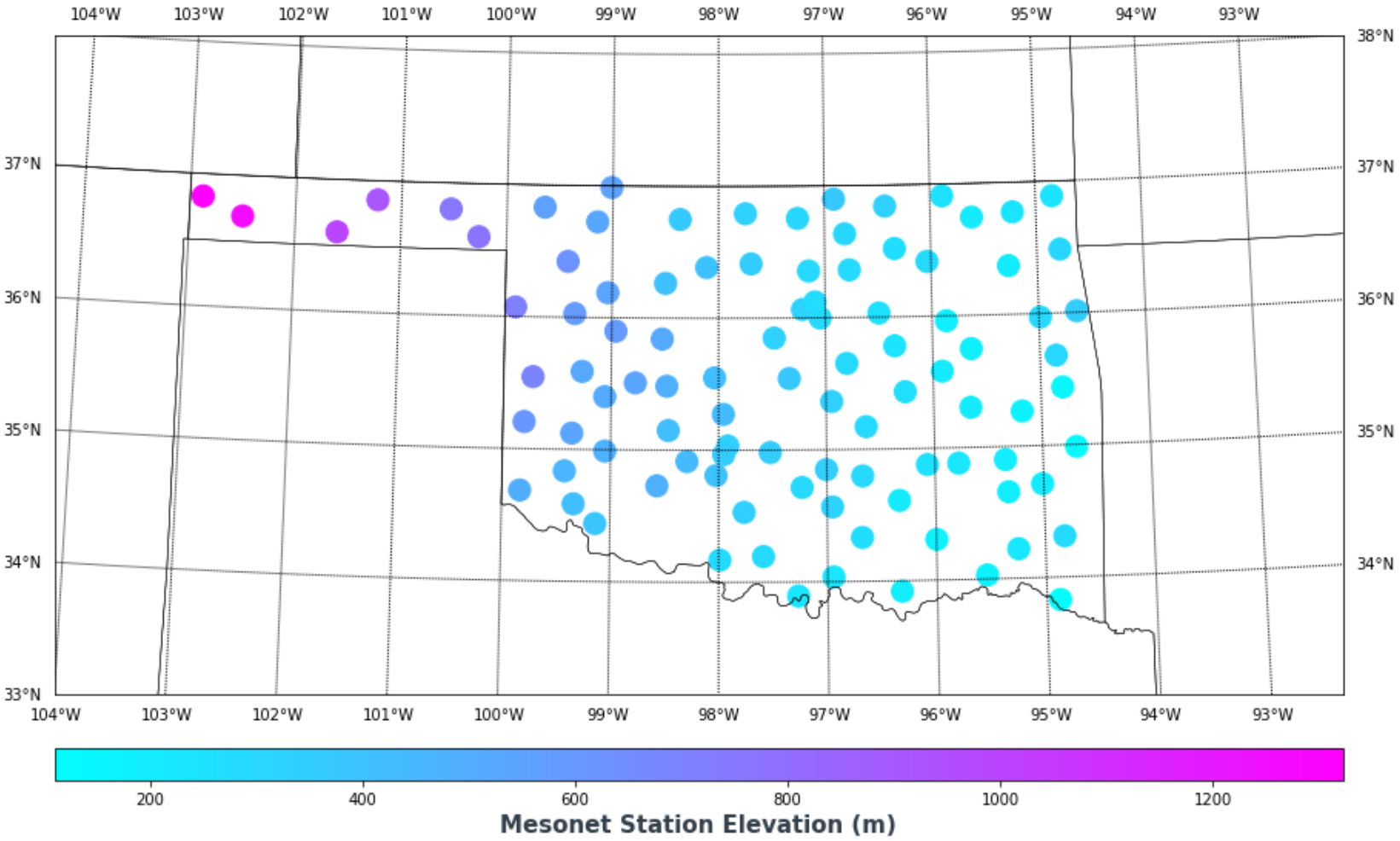
- No null values in dataset
- Fictional values
  - Client stated that the pyranometers occasionally stopped functioning correctly
  - Fictional values were input to train.csv (daily energy data)
  - These fictional values have been located and removed from the dataset (< 1% of total energy data)

Data Wrangling – Formatting

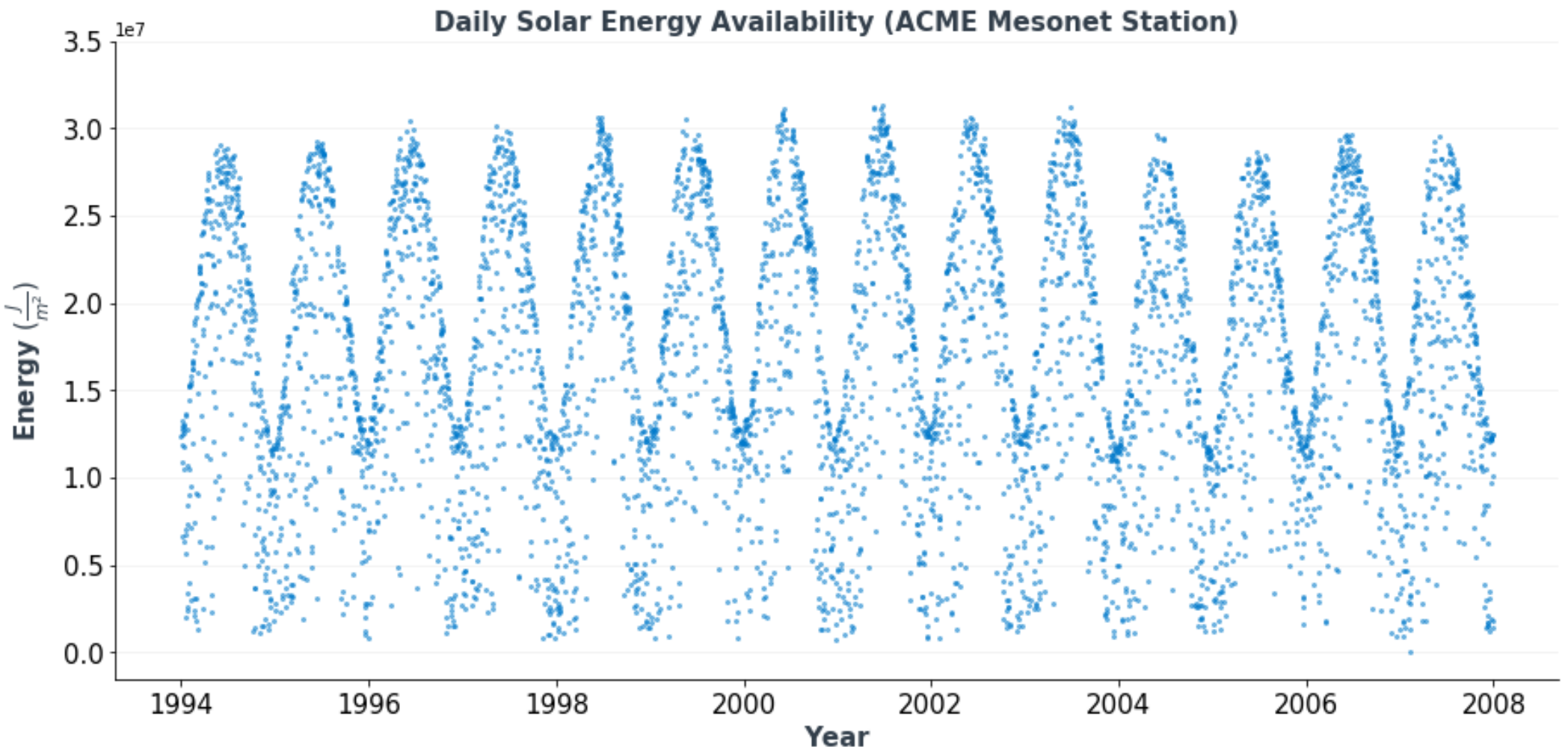
Goal: Create list of observations that include the date, station, solar energy (target variable) and the weather forecast variables for that day and station

1) Merge the list of energy data for a specific station with the list of stations
2) Determine the closest weather forecast grid point (by longitude and latitude) to a given station and get the weather forecast variables for that grid point
3) Use the median value of the 11 predictive models as a single weather forecast
4) Pivot the 5 daily forecast hours to be considered 5 different variables (for each of the 15 weather variables)
5) Merge the weather variable forecast data with the daily energy availability for each day and station

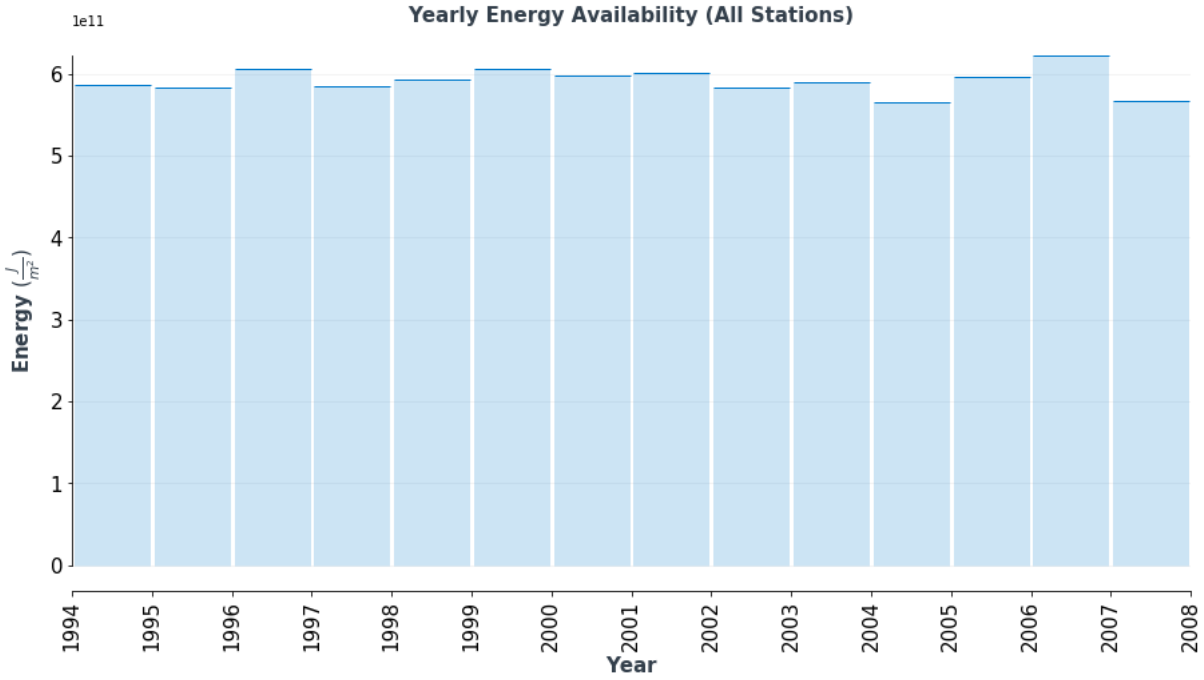
Exploratory Data Analysis – Mesonet Stations

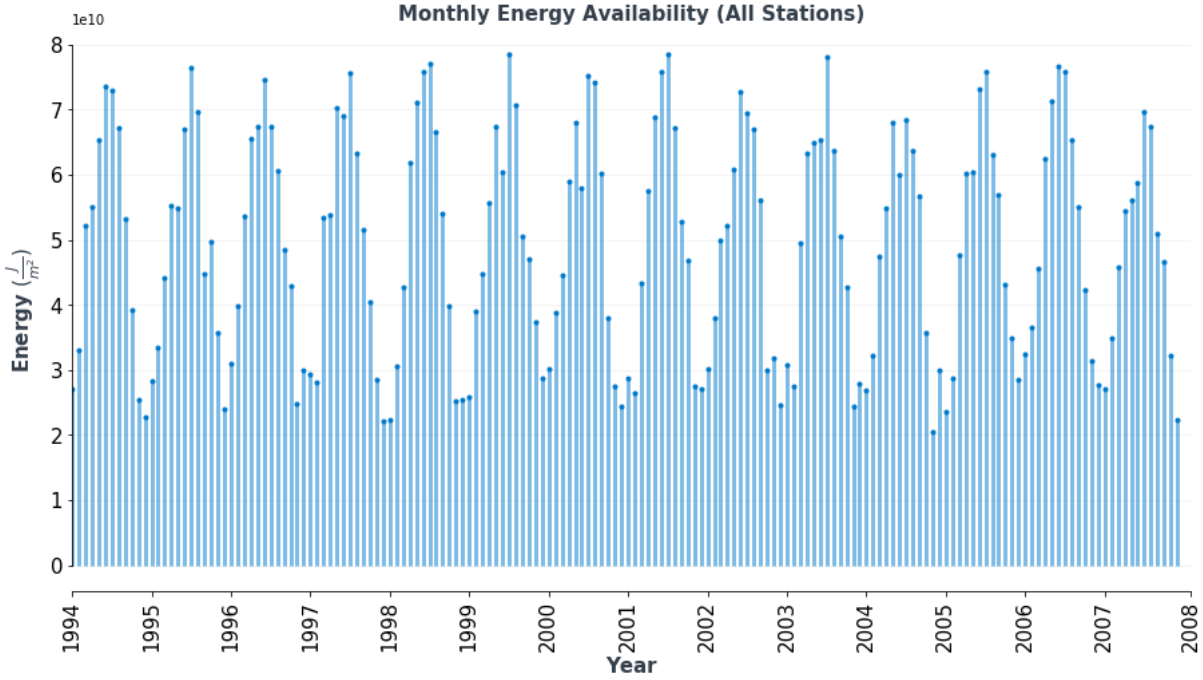


Exploratory Data Analysis – Energy over Time

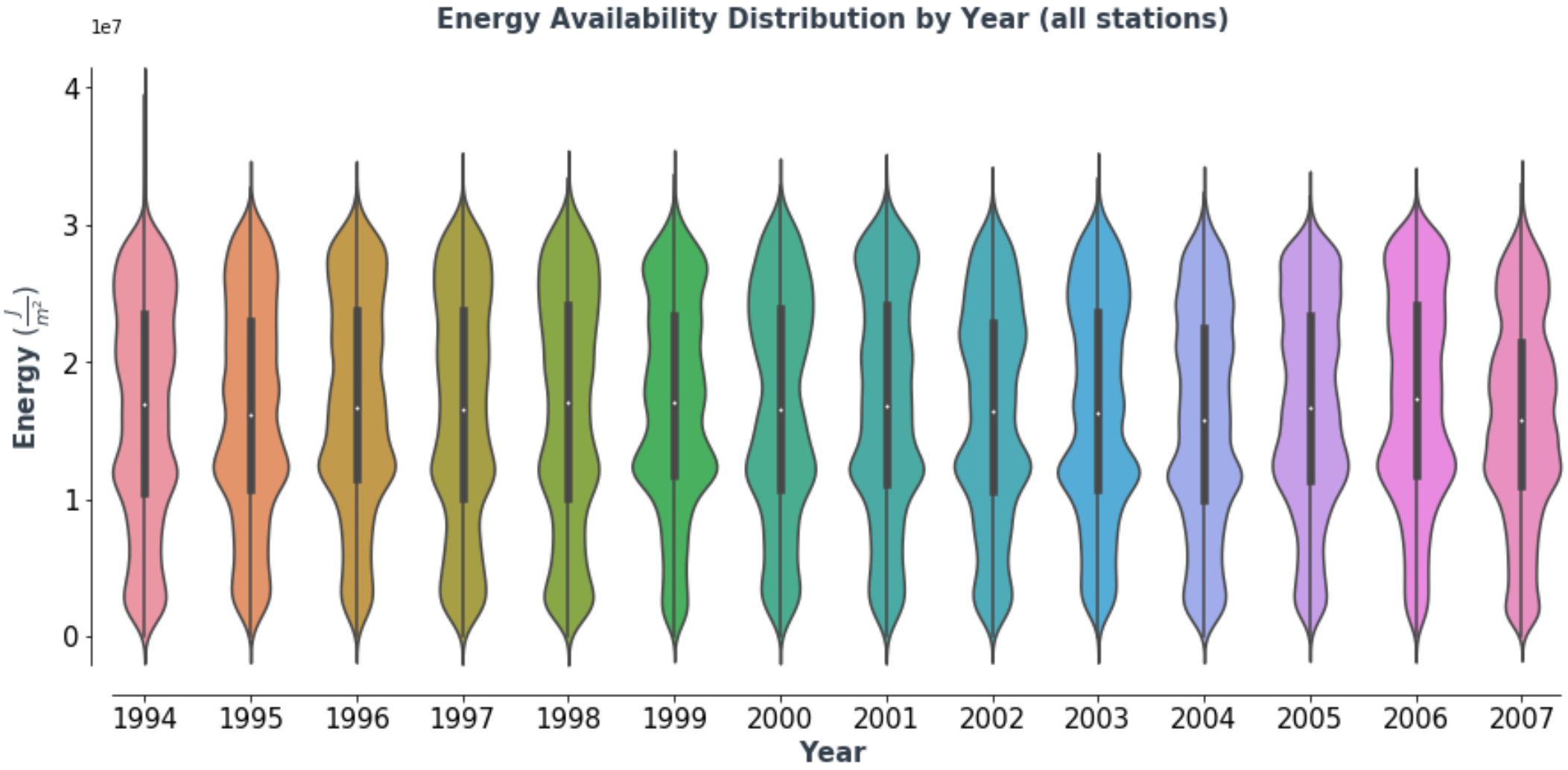


Exploratory Data Analysis – Energy over Time (2)

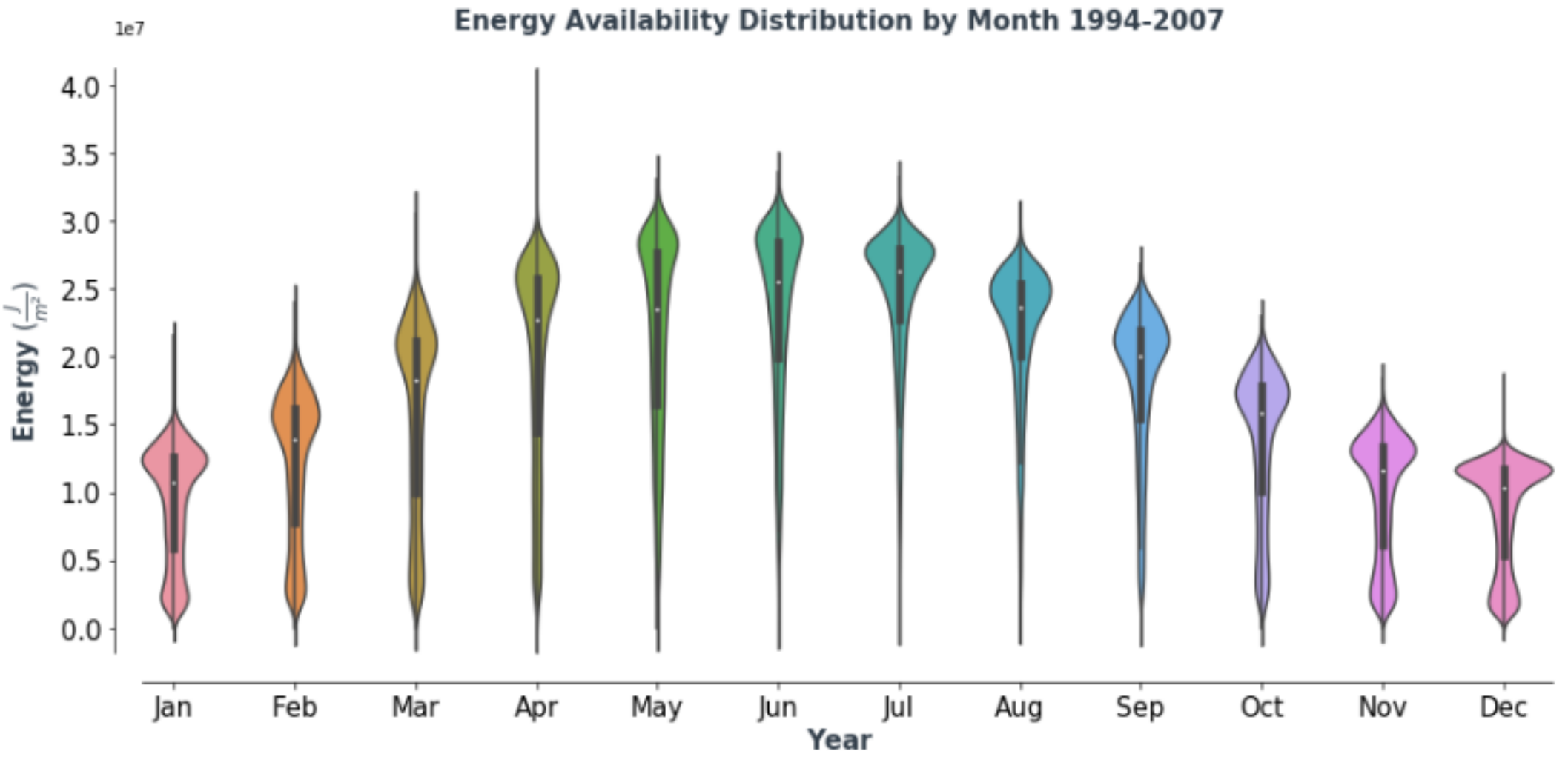




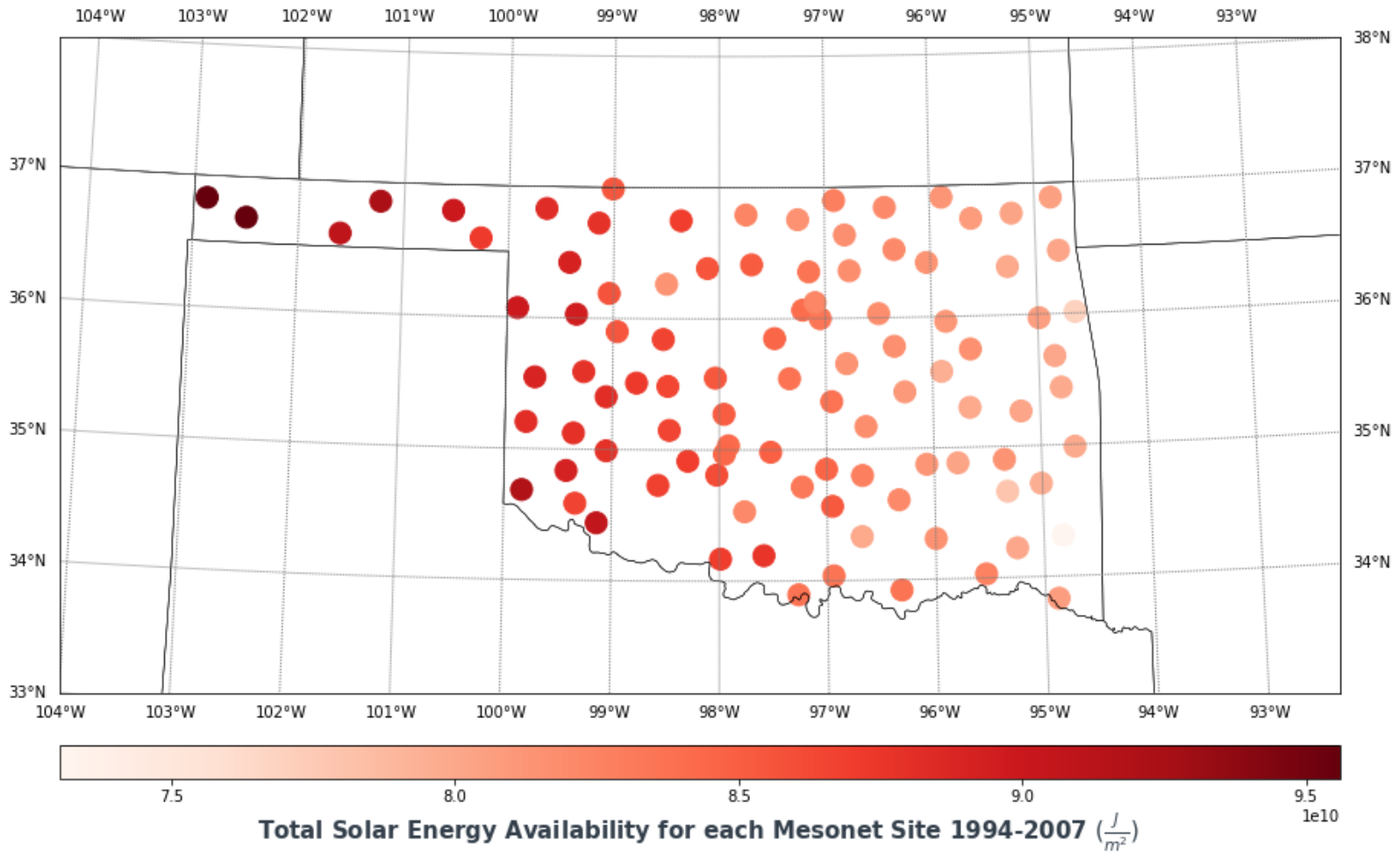
Exploratory Data Analysis – Energy over Time (3)



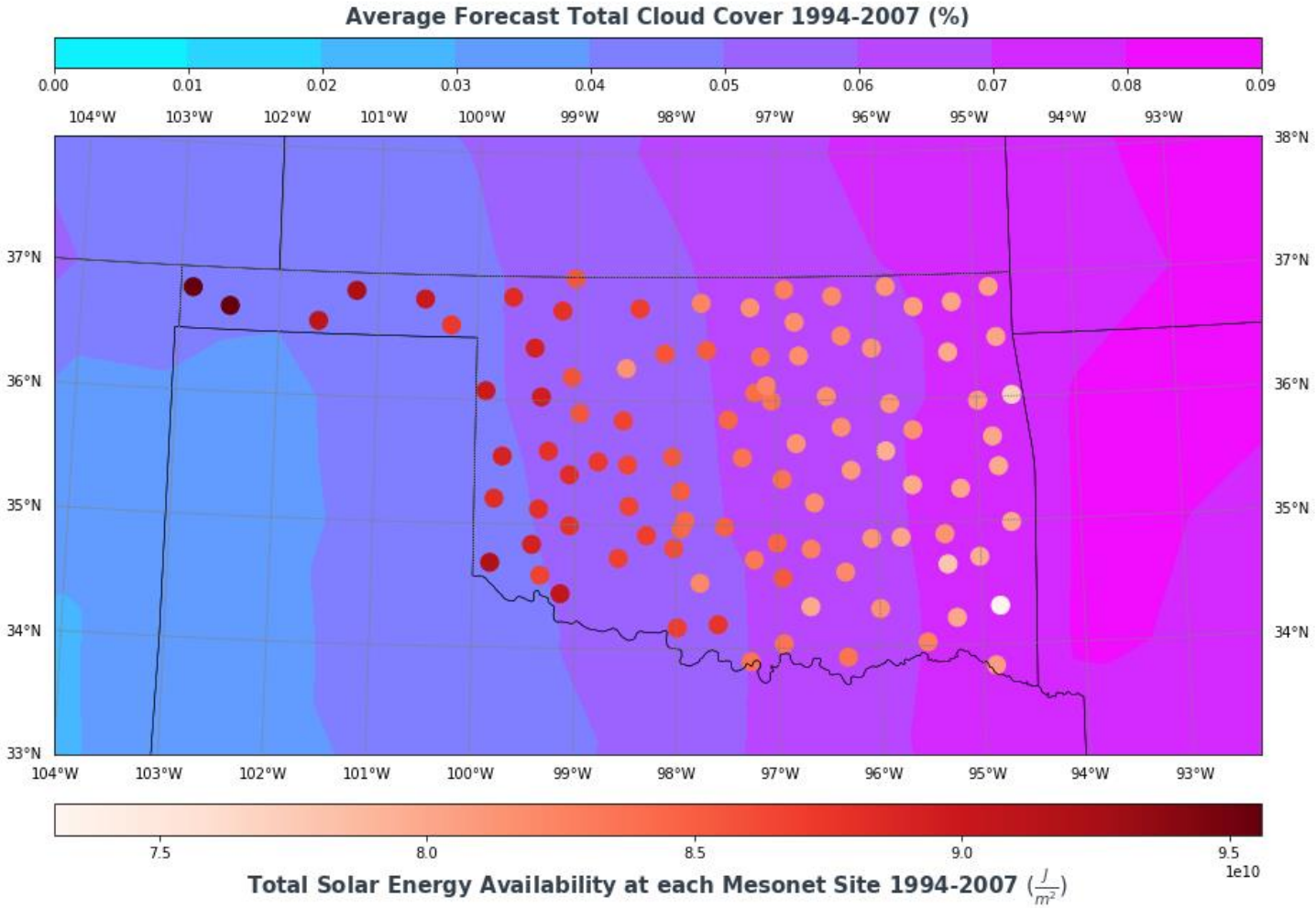
Exploratory Data Analysis – Energy over Time (4)



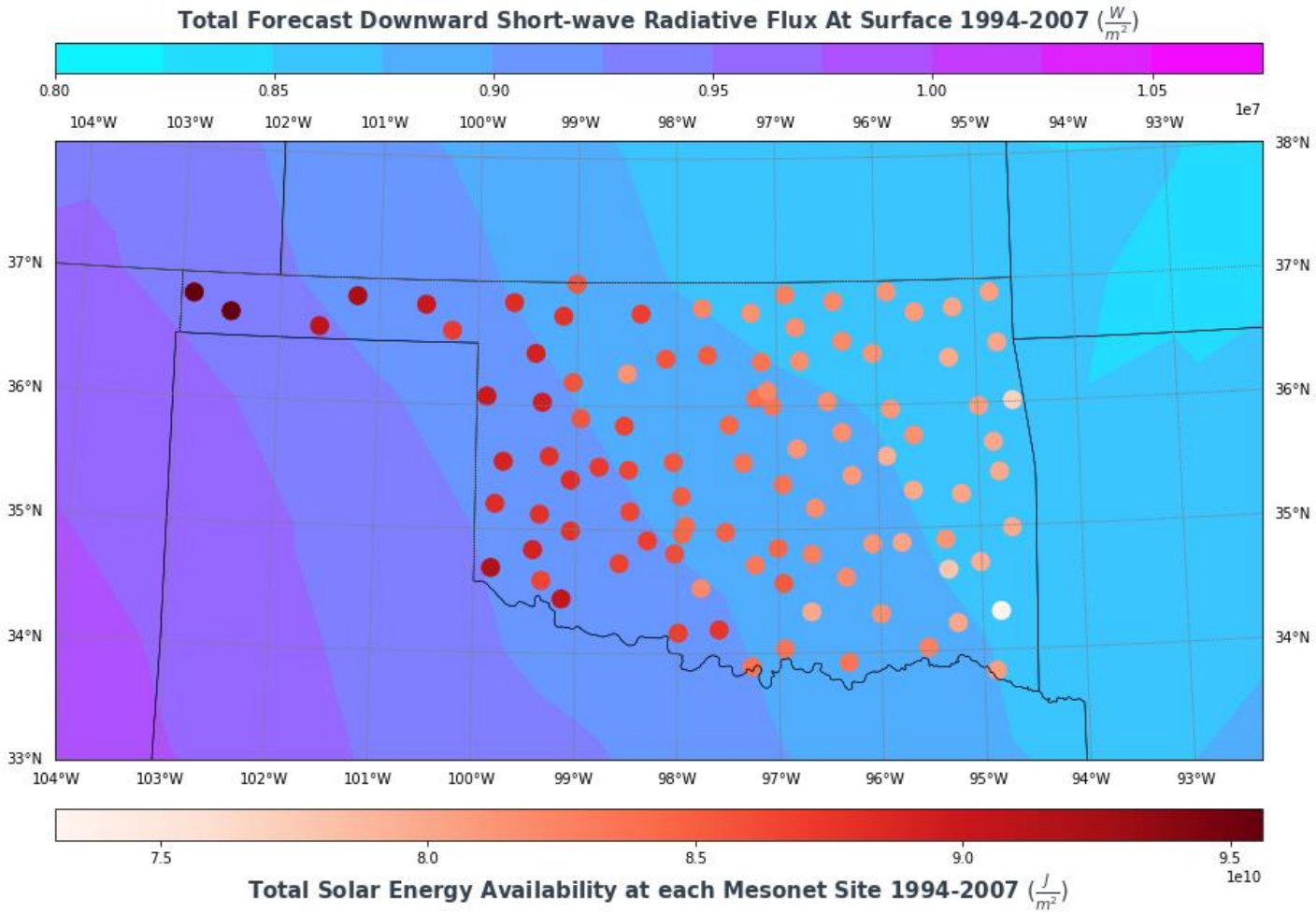
Exploratory Data Analysis – Energy over Space



Exploratory Data Analysis – Weather Forecast Variables

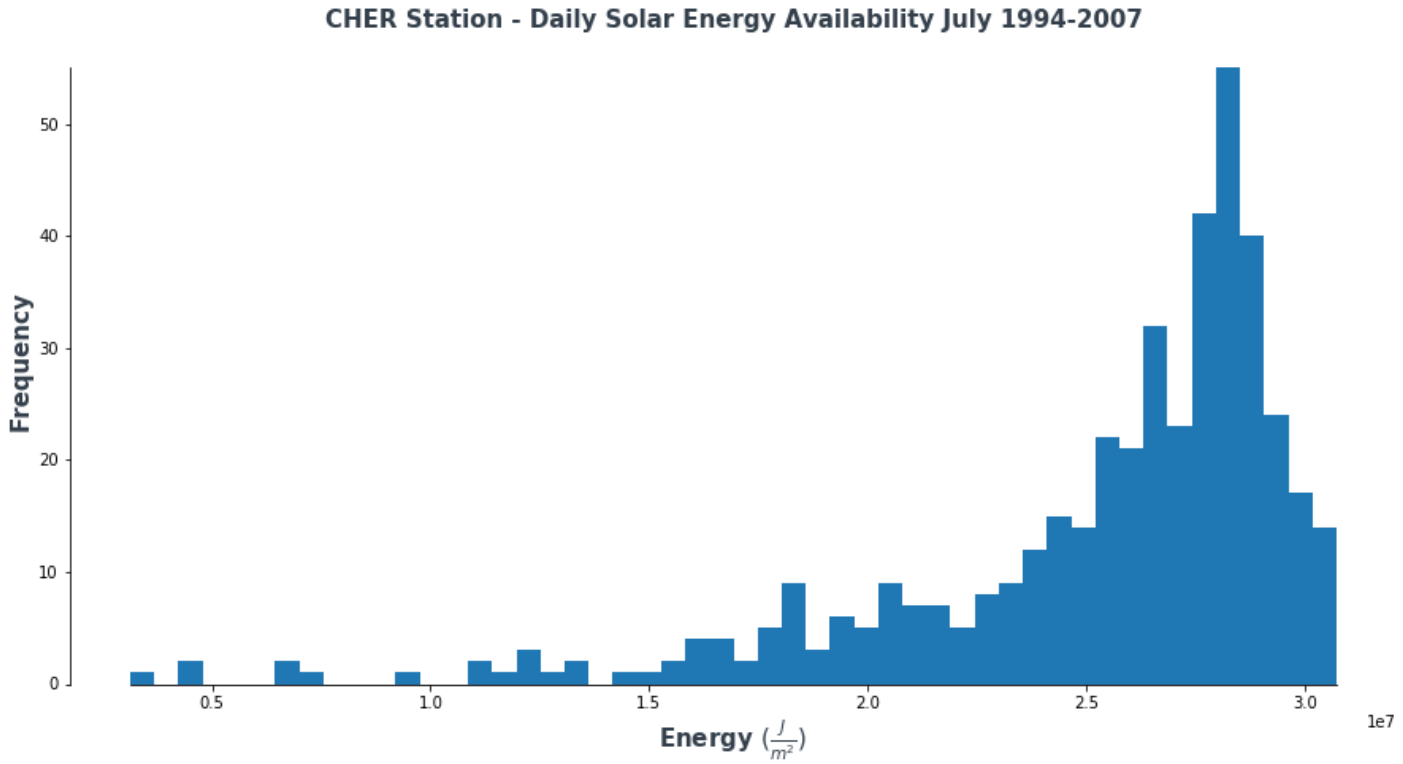


Exploratory Data Analysis – Weather Forecast Variables (2)

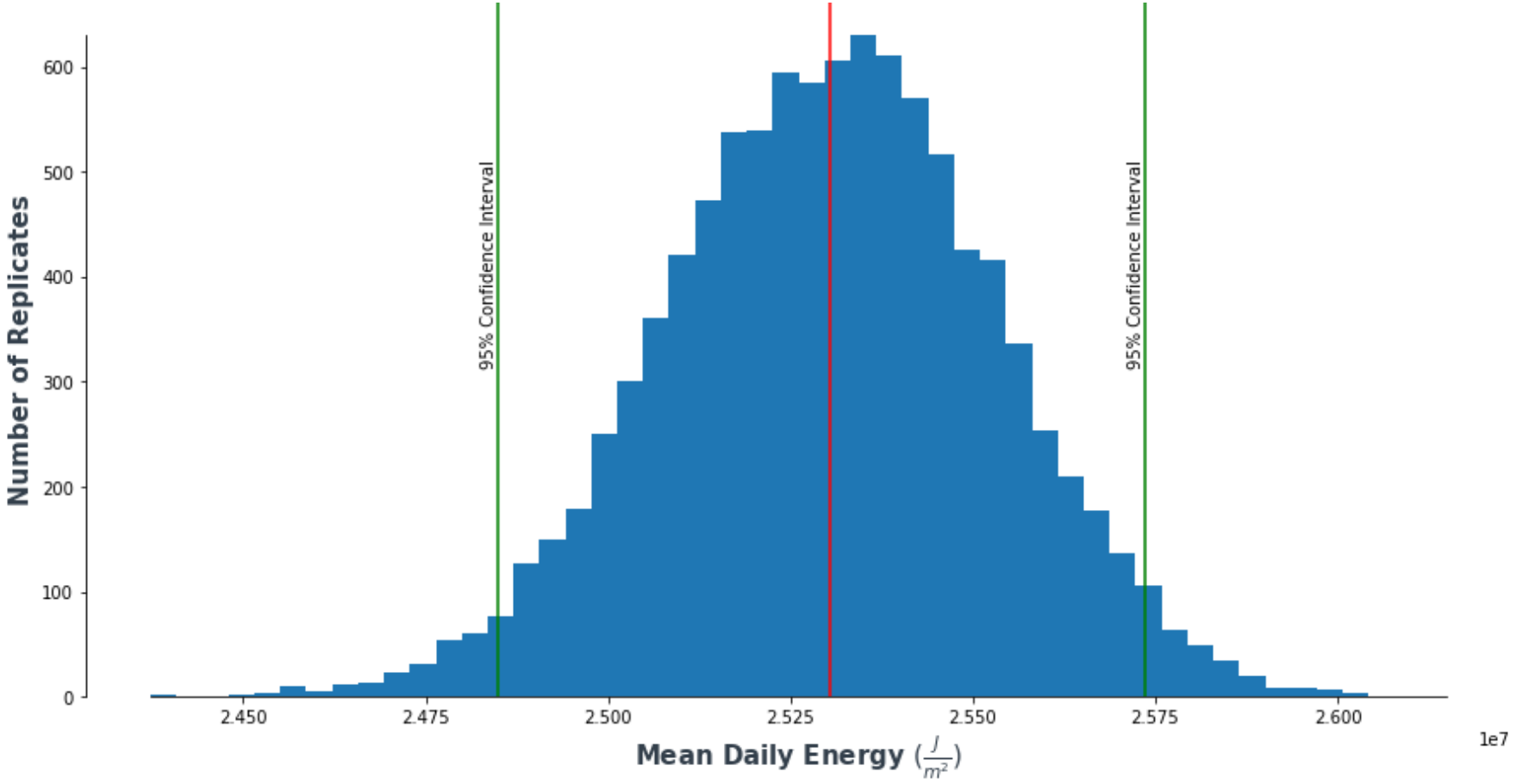


Statistical Data Analysis – Bootstrap Inference

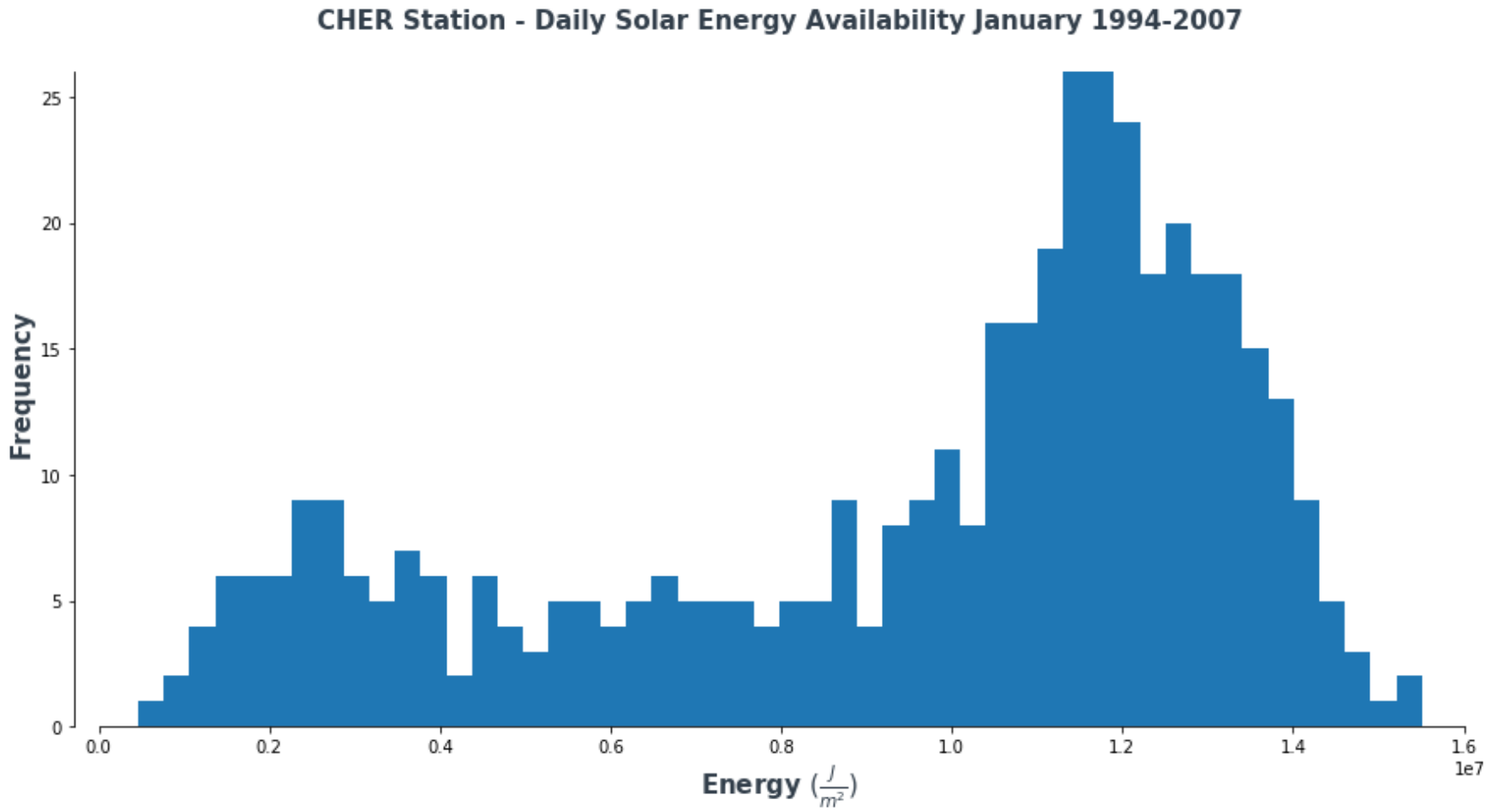
Goal: Determine a confidence interval for the mean daily energy availability for a single station for a given month.



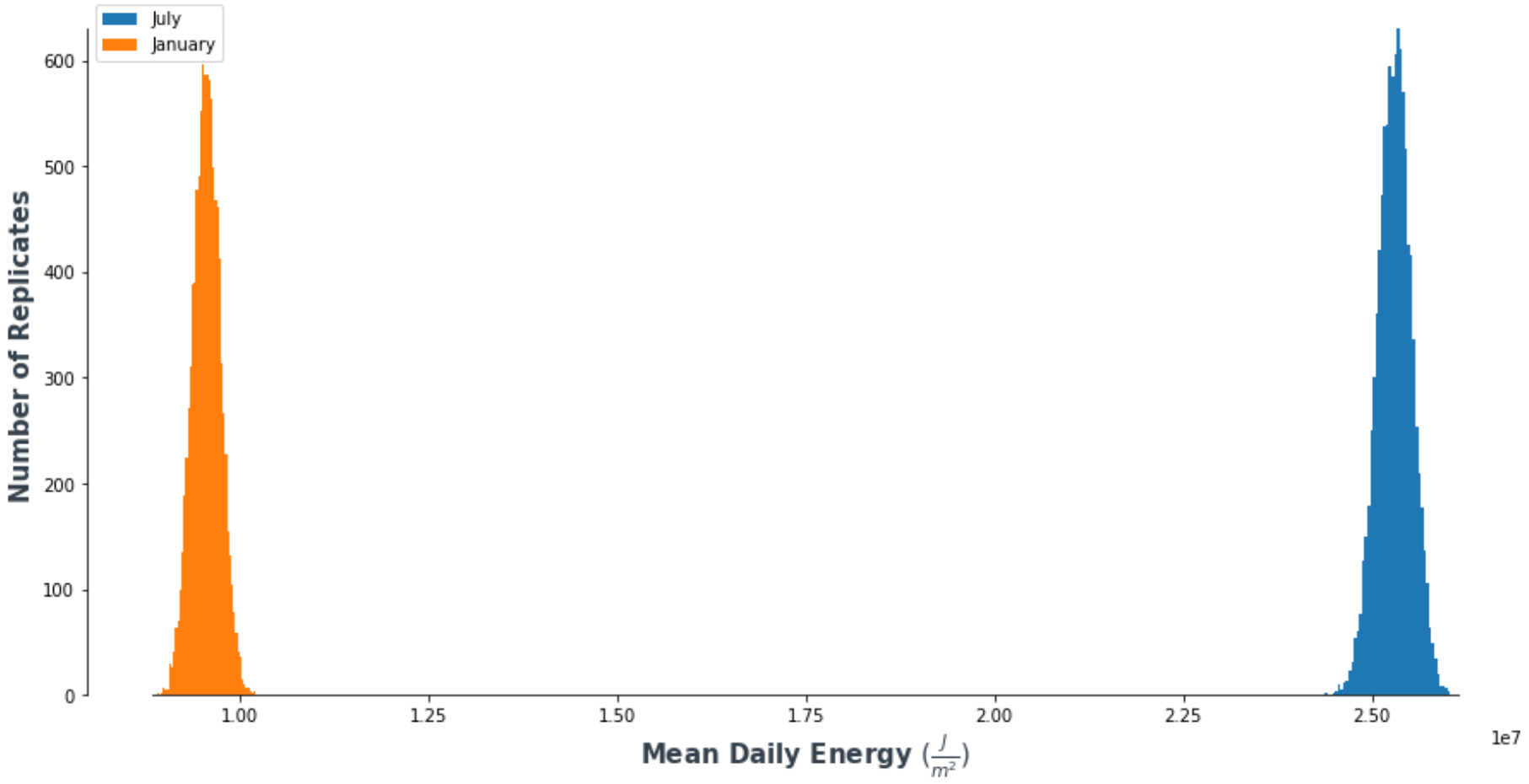
Statistical Data Analysis – Bootstrap Inference (2)



Statistical Data Analysis – Bootstrap Inference (3)

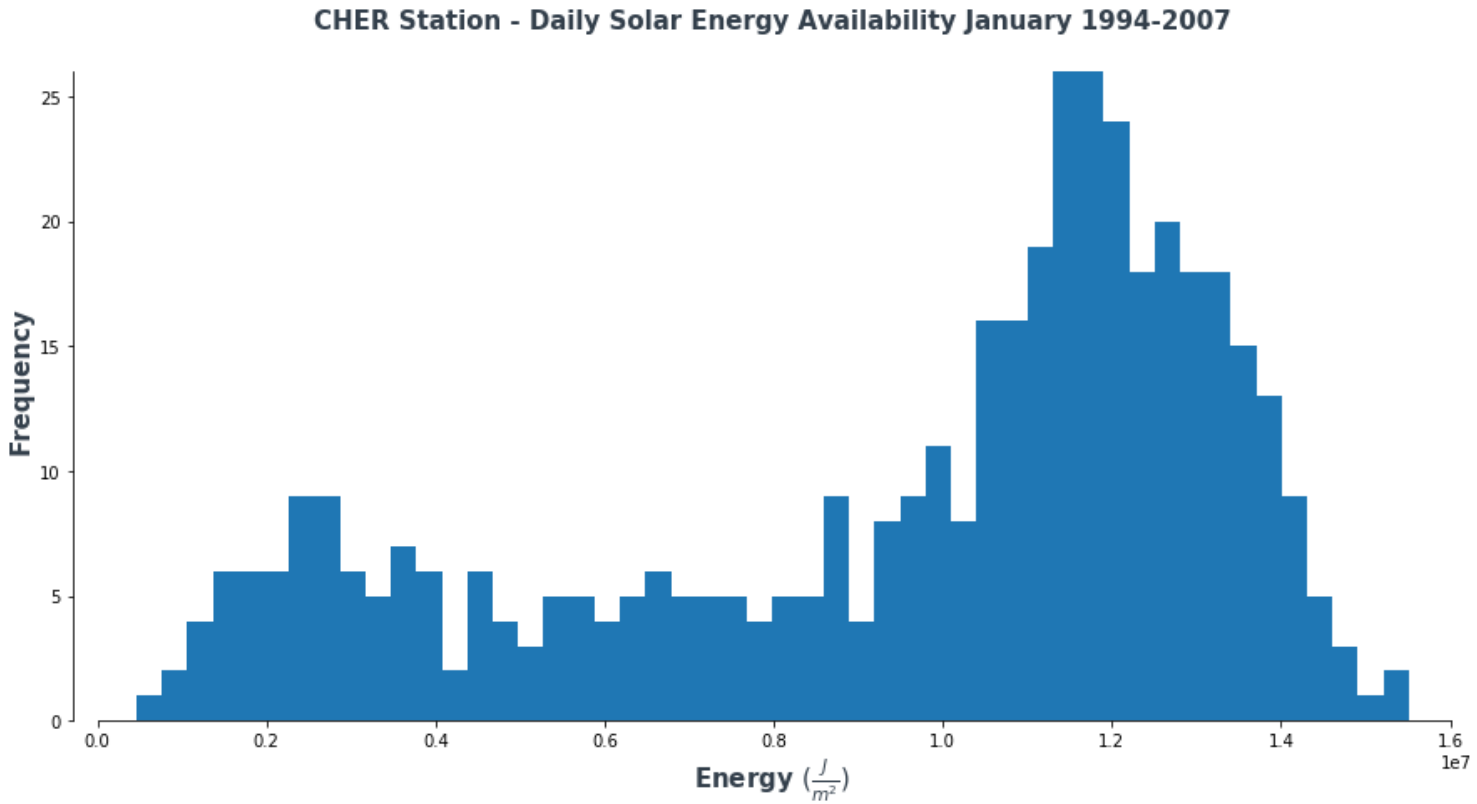


Statistical Data Analysis – Bootstrap Inference (4)

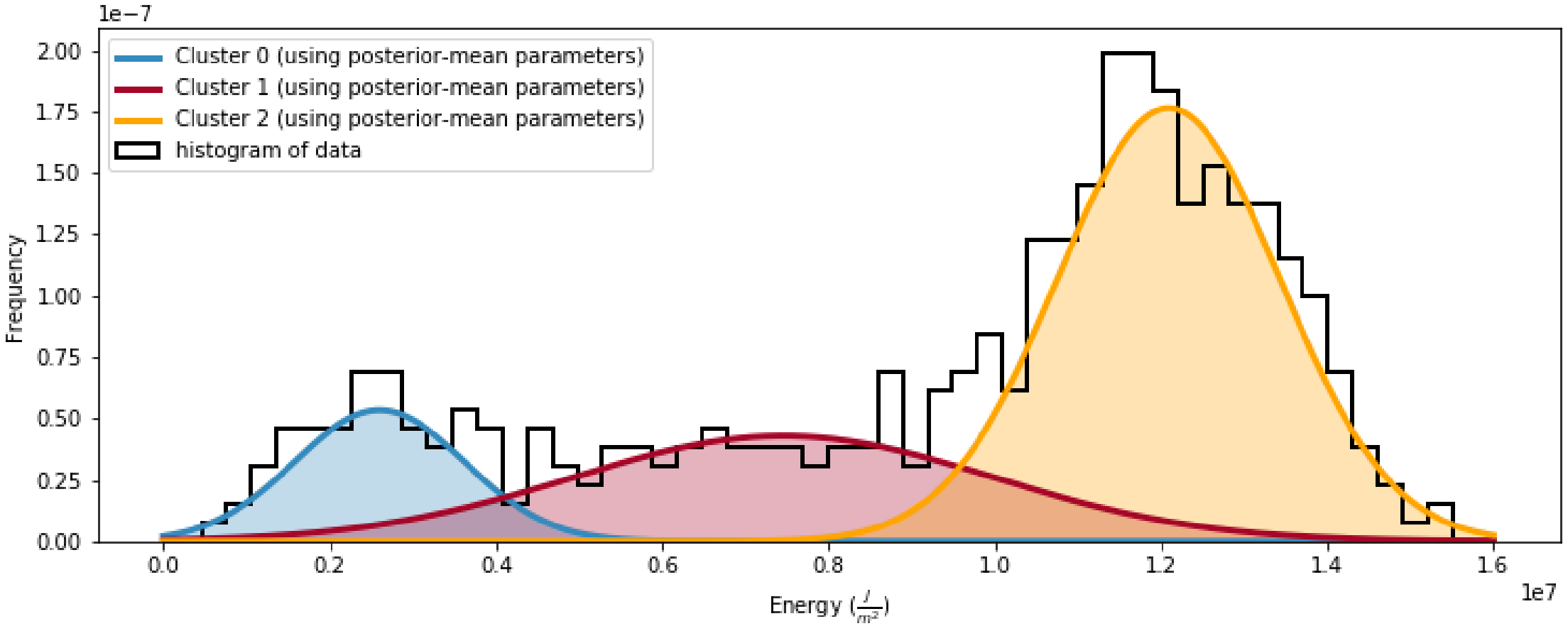


Statistical Data Analysis – Bayesian Inference

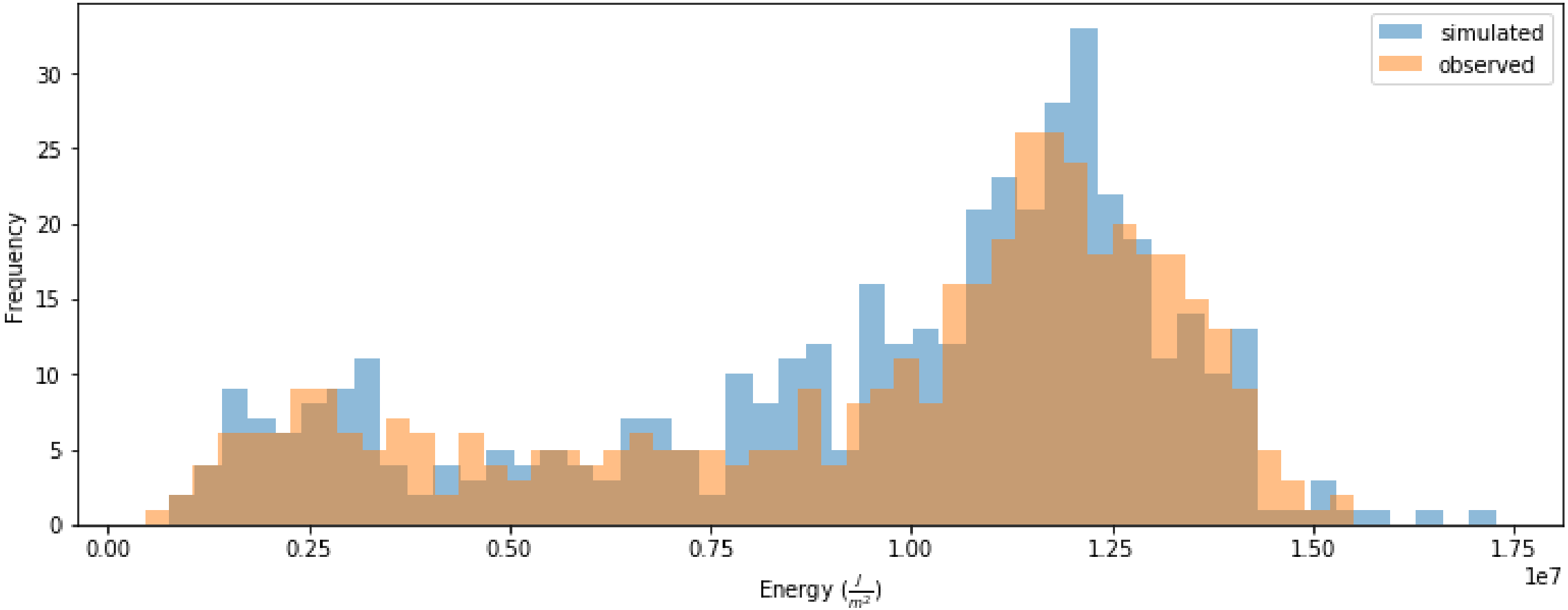
Goal: Model the distribution of daily energy availability for a single station for a given month.



Statistical Data Analysis – Bayesian Inference (2)



Statistical Data Analysis – Bayesian Inference (3)



Machine Learning – Overview

Goal: Predict the solar energy availability daily at each of the 98 stations given daily weather forecasts.

Machine learning models built:

- Ordinary least squares linear regression
  - Varying features
- Stochastic Gradient Descent (SGD)
  - Hyperparameter optimization
- Gradient Boosting Regressor (GBR)
  - Hyperparameter optimization

Machine Learning – Model Evaluation

Cross-validation:

- Time series energy data displays auto-correlation year-to-year
- Data split into contiguous folds for cross-validation:
  - 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

Test Data is not used in hyperparameter optimization or cross-validation.
Each model is evaluated on average CV **Mean Absolute Error (MAE),** Test Data MAE, and Test Data Adjusted R^2

Machine Learning – Linear Regression

Model Theory:

- Fitting a hyperplane to a set of points in N-dimensions (N = number of features)
- Generally requires the following assumptions:
  1) Linear relationship between the target variable and features exists
  2) Data should not exhibit multicollinearity
  3) Homoscedasticity (constant variance) of residuals
  4) Normally distributed residuals

Machine Learning – OLS Model 1a

Baseline Model 1a: No Feature Selection

- 15 weather forecast variables at 5 different timestamps = 75 features
- Latitude, longitude, elevation, month of the year = 4 features
- Total = 79 features

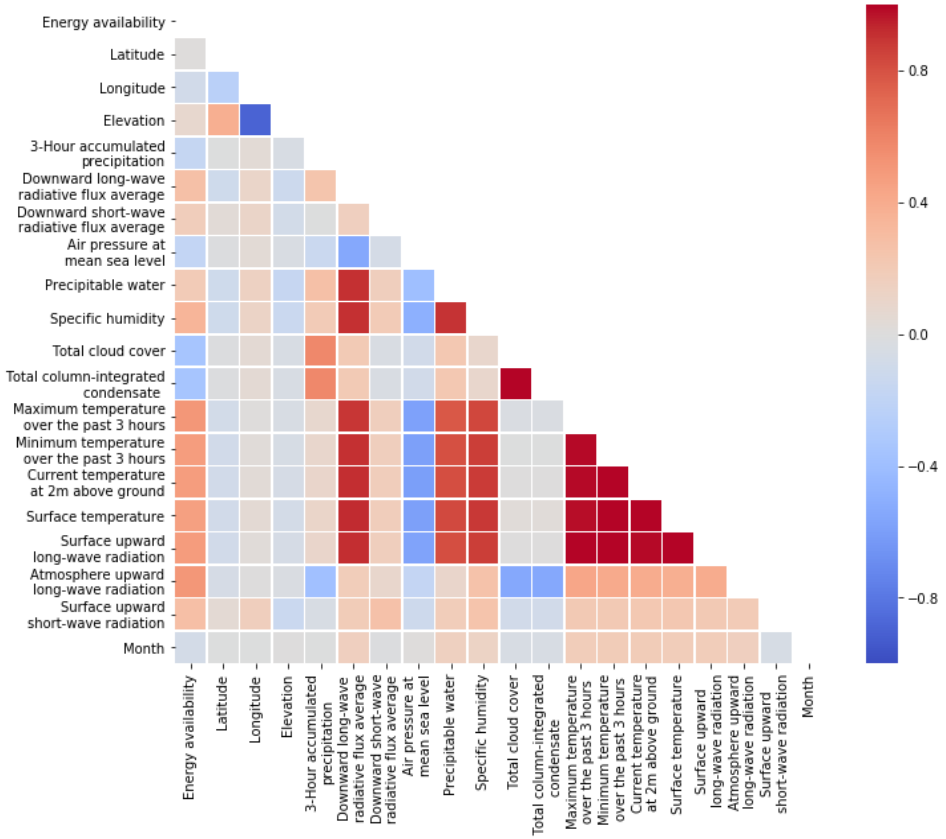
Machine Learning – OLS Model 1a (2)

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

- Regression coefficients non-zero
- T-test p-values very close to 0, indicating that the features have predictive power

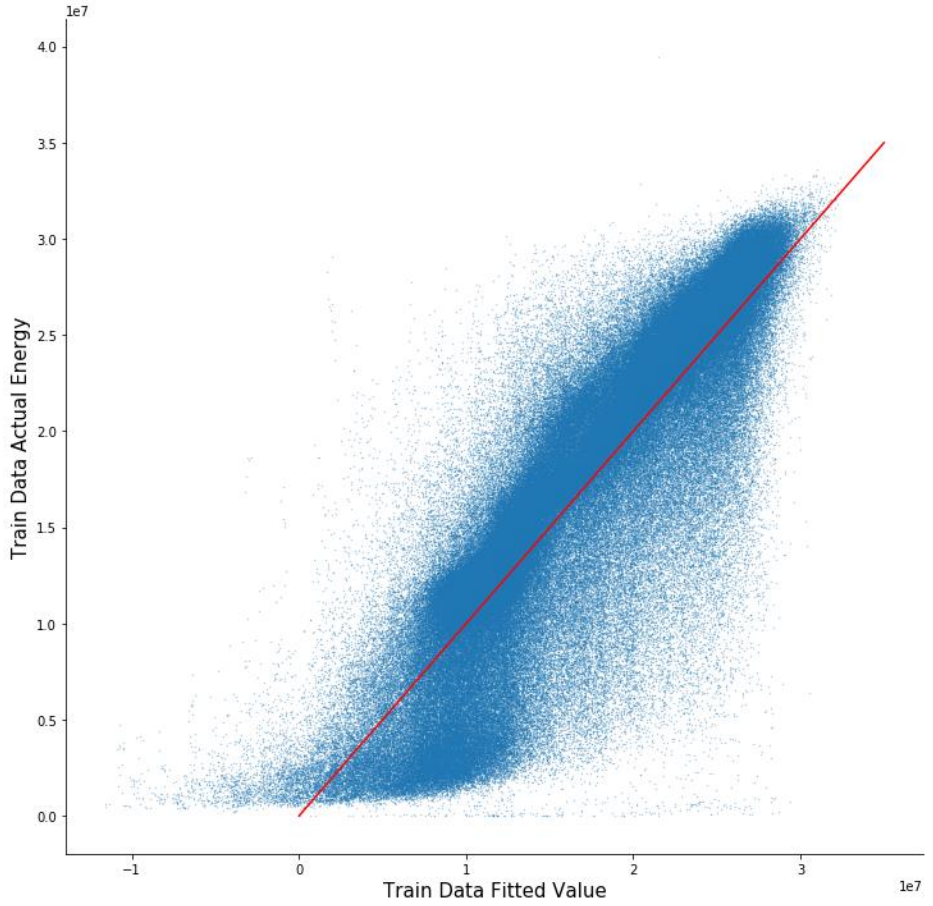
Machine Learning – OLS Model 1a (3)

Assumption #2: Data should not exhibit multicollinearity



Machine Learning – OLS Model 1a (4)

Assumption #3: Homoscedasticity (constant variance) of residuals

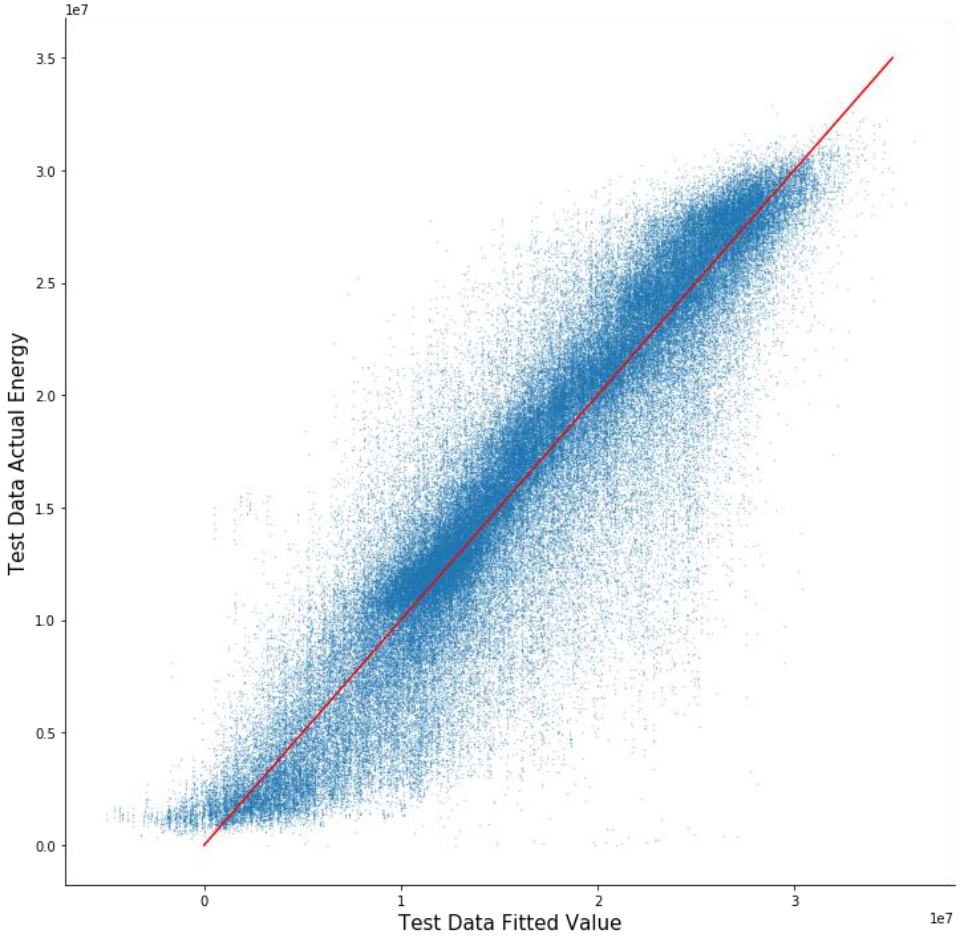


Breusch-Pagan Test:

- Null hypothesis of homoscedasticity
- Null hypothesis rejected due to near-zero p-value of the test statistic

Machine Learning – OLS Model 1a (4)

Assumption #3: Homoscedasticity (constant variance) of residuals

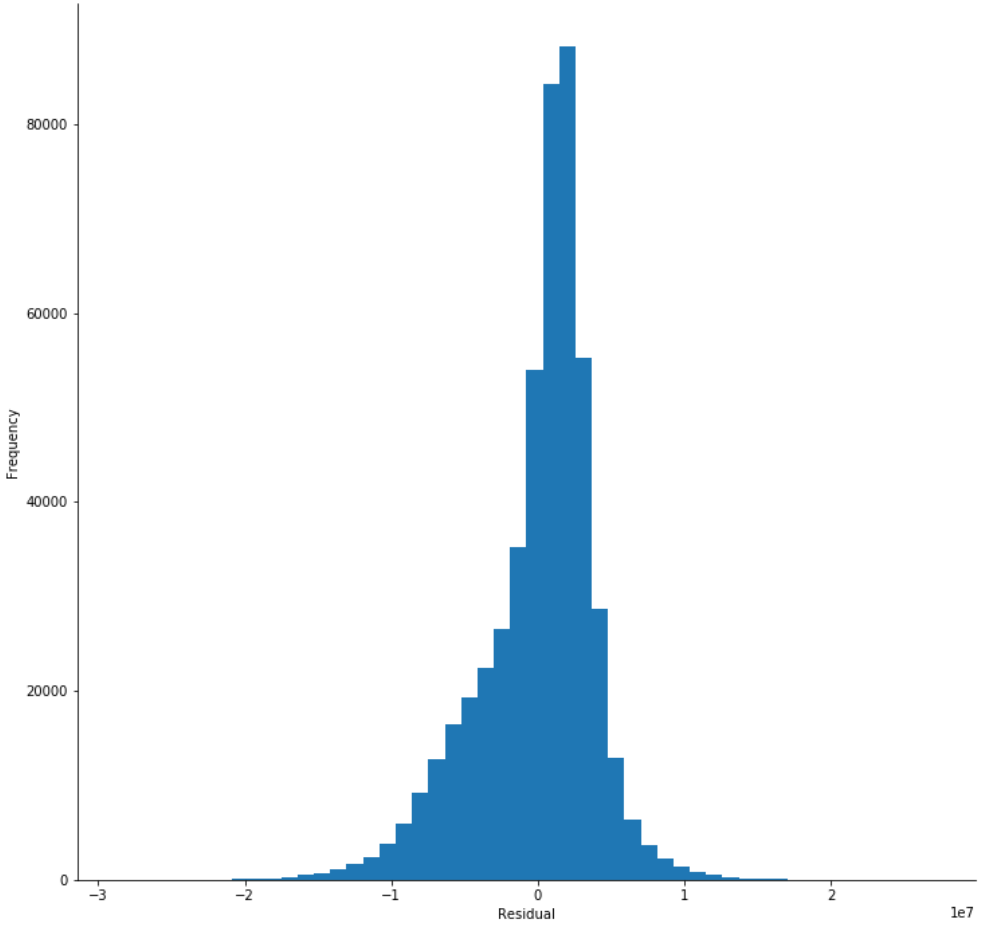


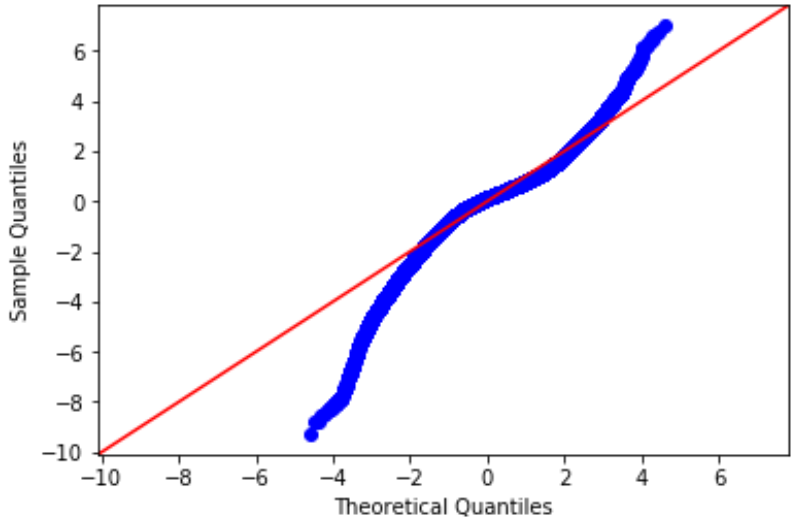
Breusch-Pagan Test:

- Null hypothesis of homoscedasticity
- Null hypothesis rejected due to near-zero p-value of the test statistic

Machine Learning – OLS Model 1a (5)

Assumption #4: Normally distributed residuals



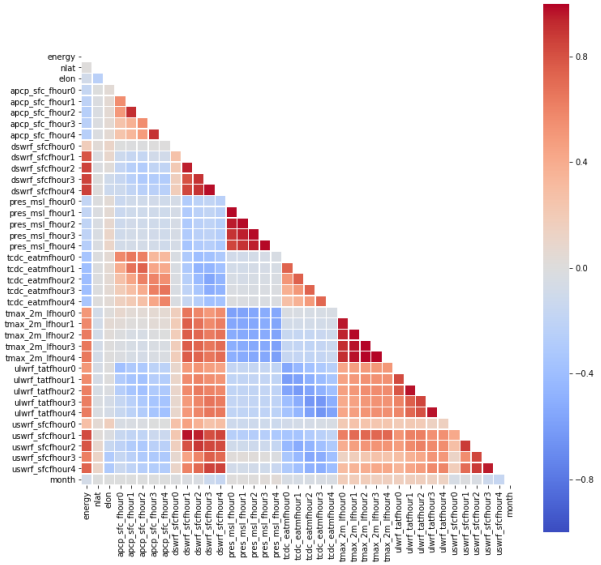


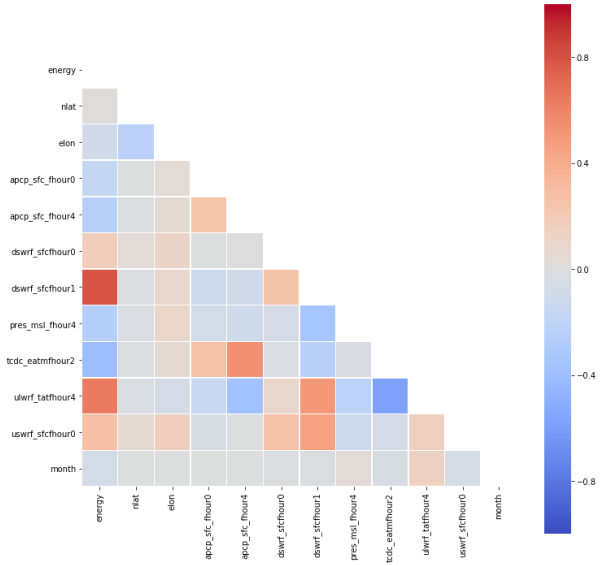
Anderson-Darling Test:

- Null hypothesis of normal distribution
- Null hypothesis rejected due to near-zero p-value of the test statistic

Machine Learning – Linear Regression Feature Reduction

- Two additional models constructed using reduced features to address multicollinearity:
  - Model 1b – features removed using forecast hour 0 correlation matrix (multicollinearity between forecast hours ignored)
  - Model 1c – all features exhibiting multicollinearity ($|R^2| > 0.7$) removed





Machine Learning – Linear Regression Conclusions

- All OLS models violate homoscedasticity and normality of residuals
- All OLS models have the greatest errors at low and high values of energy
- All models have predictive power:

	Model 1a	**Model 1b**	**Model 1c**
# Features	79	38	11
CV MAE (W/m^2)	2,353,201	2,552,359	3,021,995
Test MAE (W/m^2)	2,203,998	2,360,389	2,844,928
Test Data Adj. R^2	0.915	0.901	0.867

Machine Learning – Stochastic Gradient Descent

Model Theory:

- Process that can be used to optimize a linear regression by minimizing a cost function by following a gradient
- ‘Stochastic’ means that the process is linked with random probability
  - Samples of the dataset are randomly selected for each iteration
- Sensitive to feature scaling – features are scaled before fitting
- Several hyperparameters to be tuned

Machine Learning – Stochastic Gradient Descent (2)

Hyperparameter Tuning:

- 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)

Results:

- Residuals violate homoscedasticity and normality

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

Machine Learning – Gradient Boosting Regressor

Model Theory:

- Ensemble prediction process (i.e., uses a collection of predictors to give a final prediction)
  - Reduce noise, variance, and bias
- 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
- Several hyperparameters to be tuned

Machine Learning – Gradient Boosting Regressor (2)

Hyperparameter Tuning:

- 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)

Results:

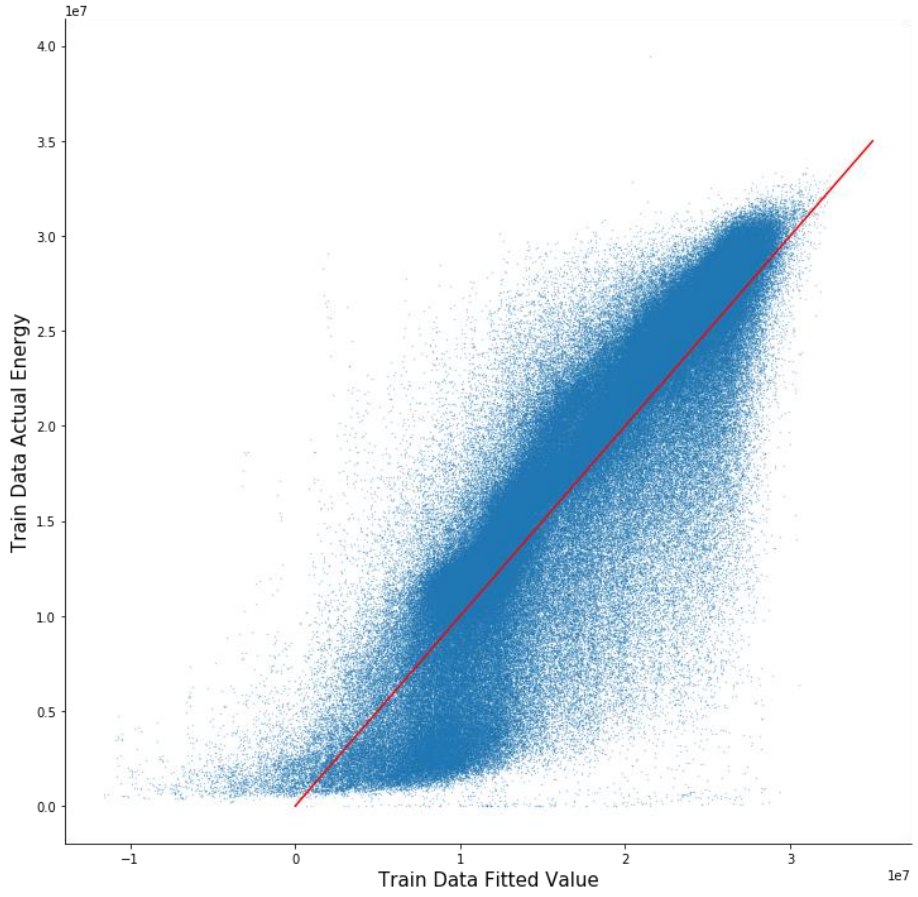
- Residuals violate homoscedasticity and normality

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

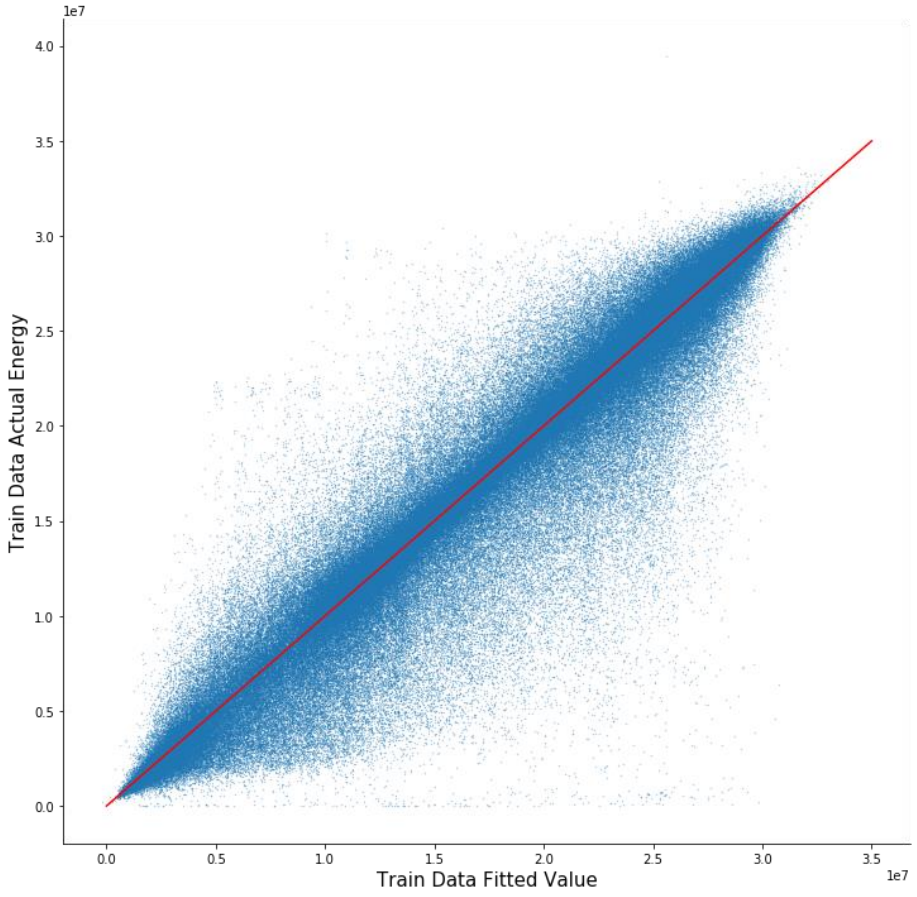
Machine Learning – Model Comparison

	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
Train Data Adj. R^2	0.915	0.914	0.924

Machine Learning – Model Comparison (2)

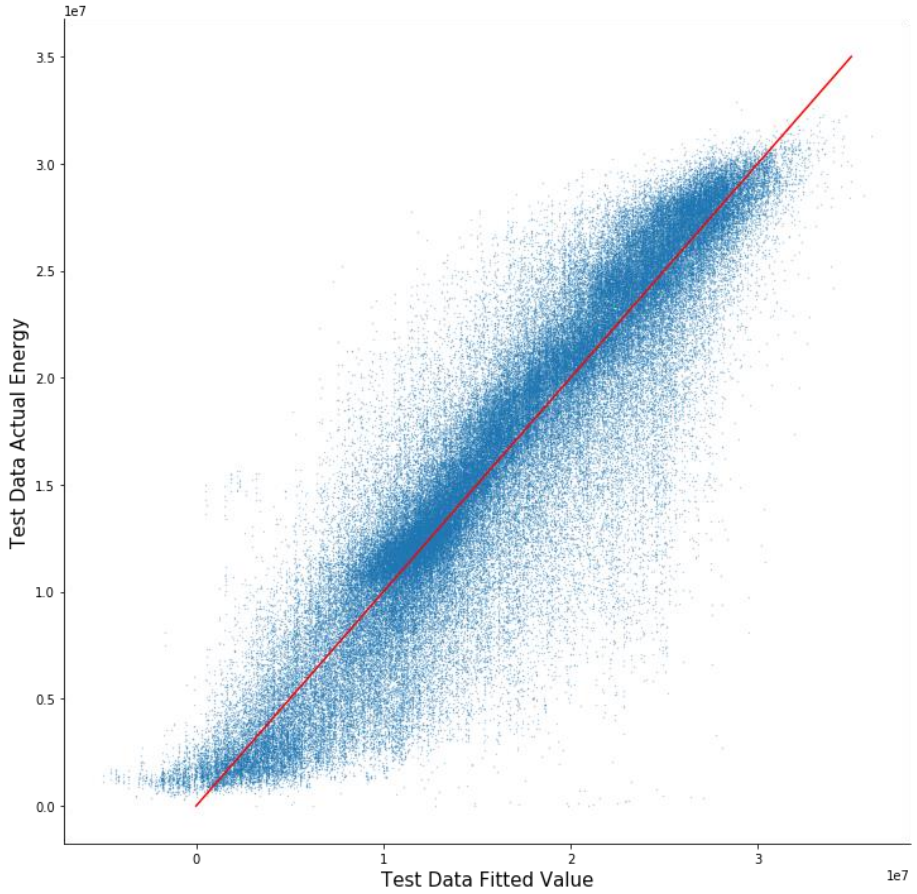


OLS Model 1a

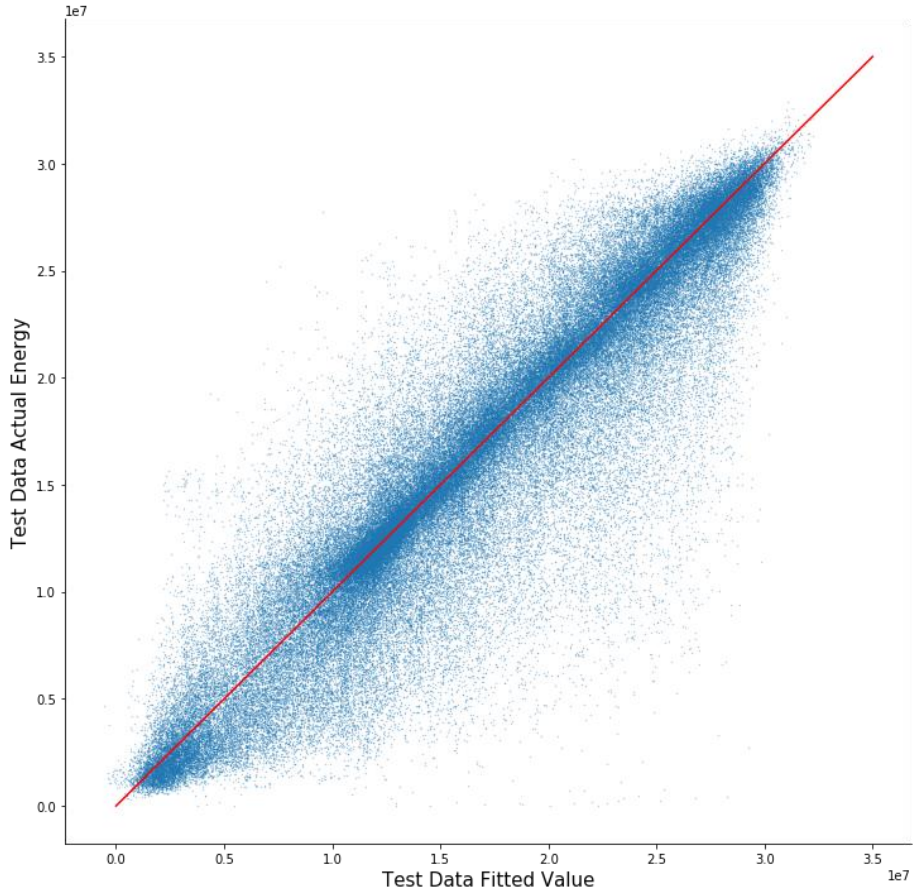


GBR Model

Machine Learning – Model Comparison (2)

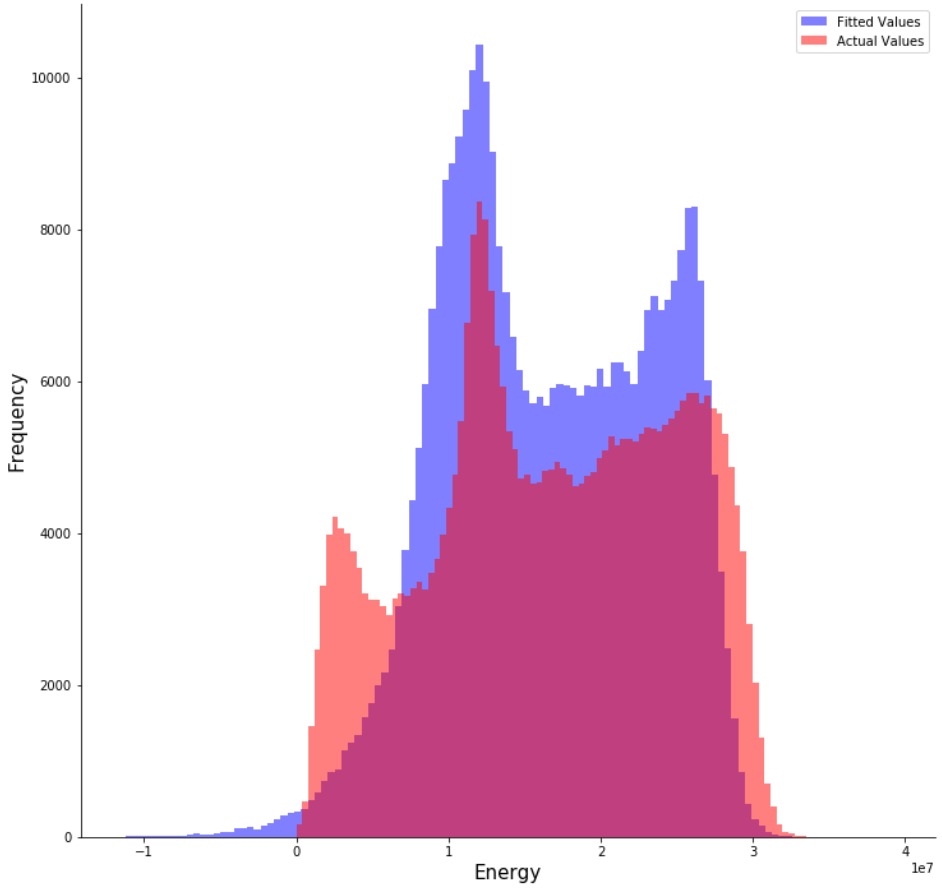


OLS Model 1a

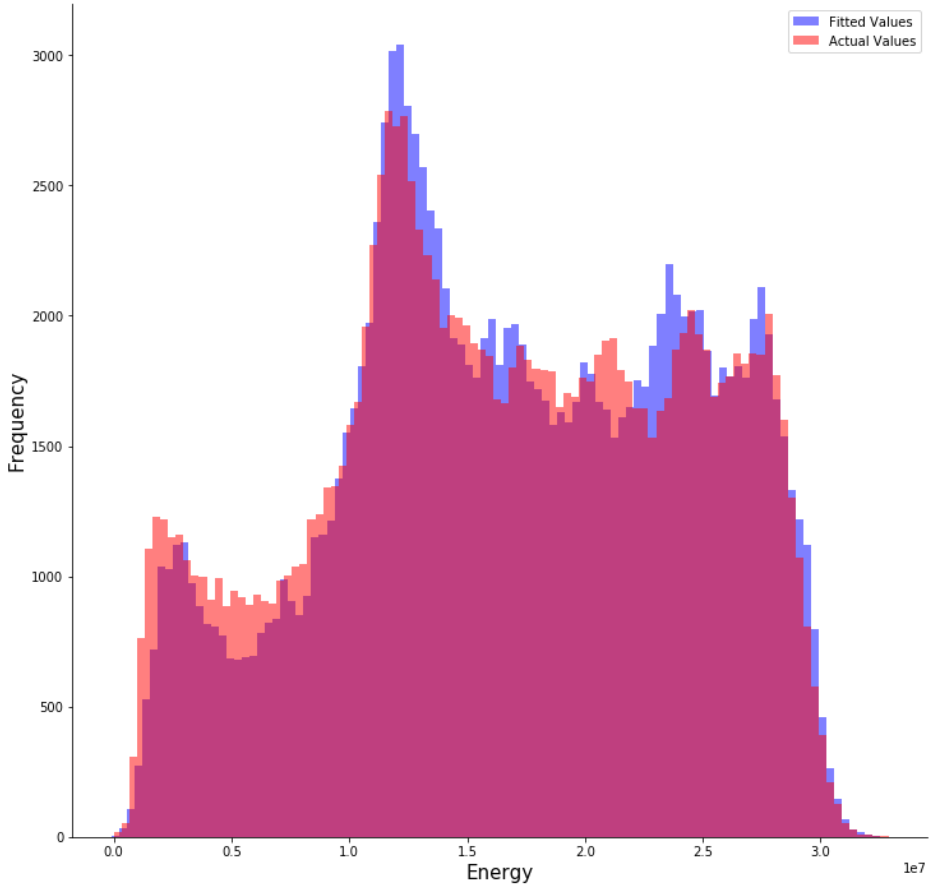


GBR Model

Machine Learning – Model Comparison (3)



OLS Model 1a



GBR Model

Conclusions

- Linear regression models have predictive power but also many shortcomings
  - Low values were over-estimated while high values were under-estimated
  - Residuals violate homoscedasticity and normality
- **GBR model performed best** according to visual assessment and the evaluation criteria
  - On test dataset, model predicts total daily energy with a mean absolute error of 1,975,215 (W/m^2), approximately 11.9% of the mean total daily energy value

Further Work

- More time tuning GBR hyperparameters using randomized and grid searches
- Use all 11 weather forecast models to train 11 regression models and average the predictions
- Use distance-weighted average of weather forecasts (at multiple grid points) as the weather forecast variable features

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] https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/overview

[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.

[7] http://www.apricus.com/solar-pv-systems-pv-panels-19.html#.XYkWZyhKiUk

[8] https://en.wikipedia.org/wiki/Solar_panel#/media/File:Photovoltaik_Dachanlage_Hannover_-_Schwarze_Heide_-_1_MW.jpg