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

Connor McAnuff

August 18, 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 is used as a measure of the power available to enter a solar PV system. The system configuration and power 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.

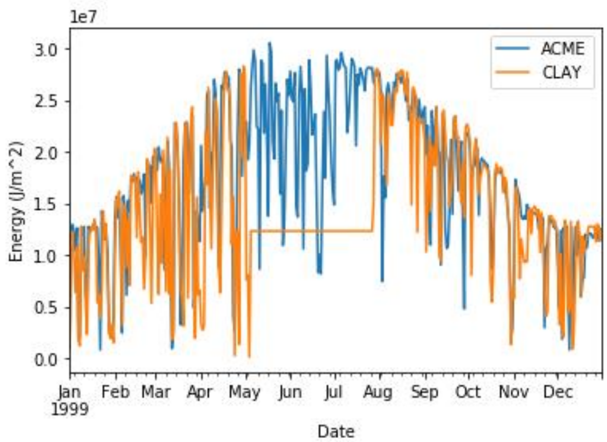

Figure 1: Daily available solar energy at two stations for the year 1999. CLAY has fictional data for May-July.

The fictional values were removed from the dataset after data formatting. They comprise 0.42% 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.

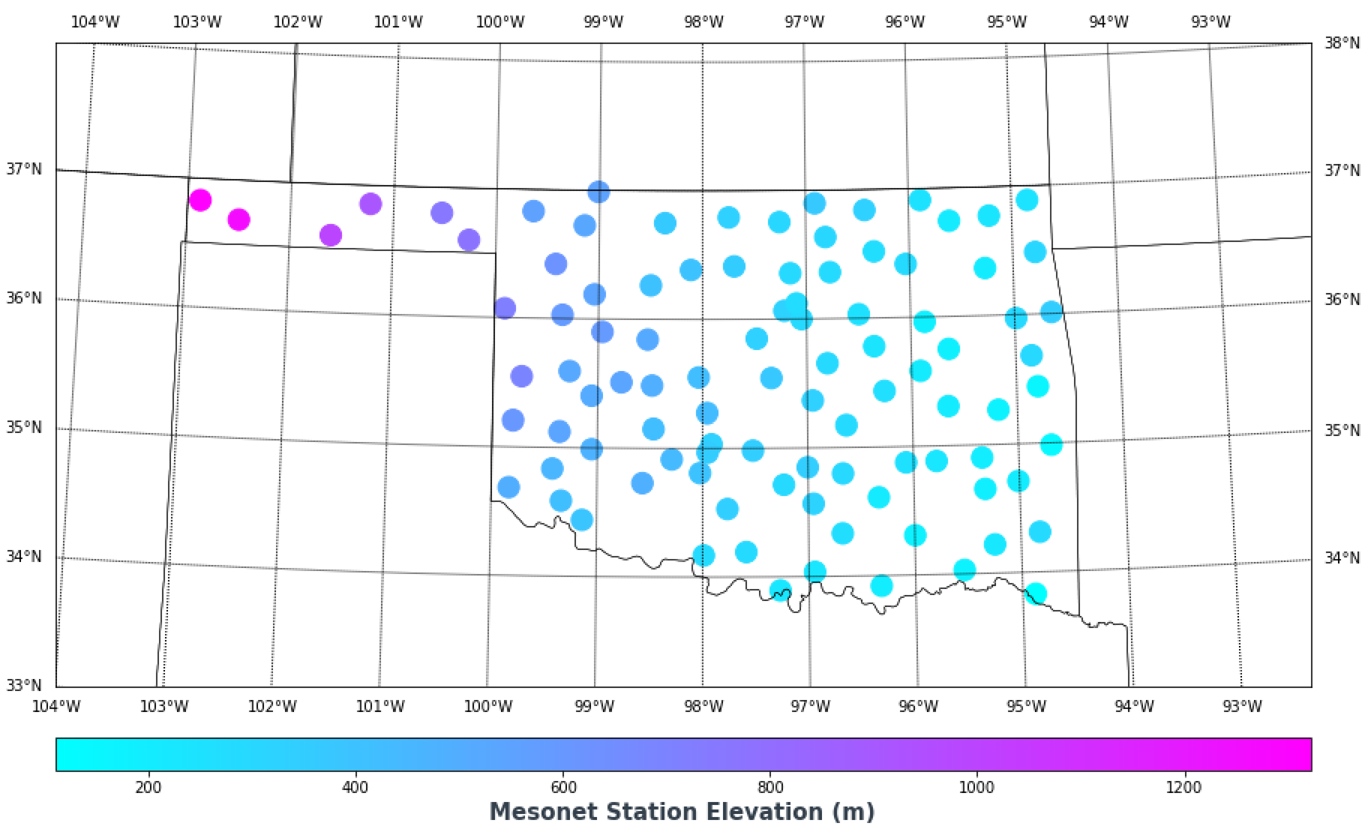

Figure 2: Mesonet station location and elevation.

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.

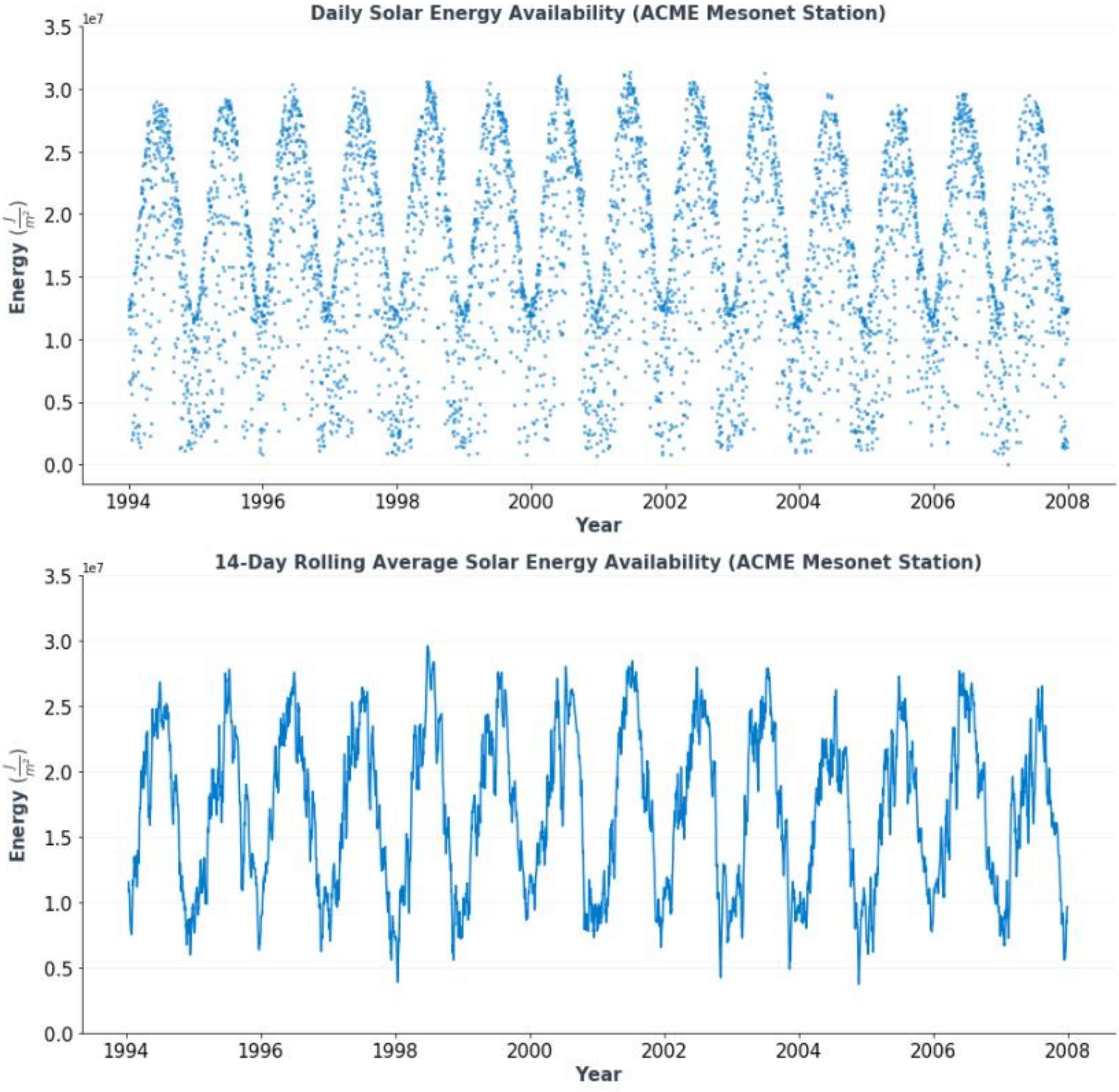

Figure 3: Daily and 14-day rolling average solar energy availability at ACME station.

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.

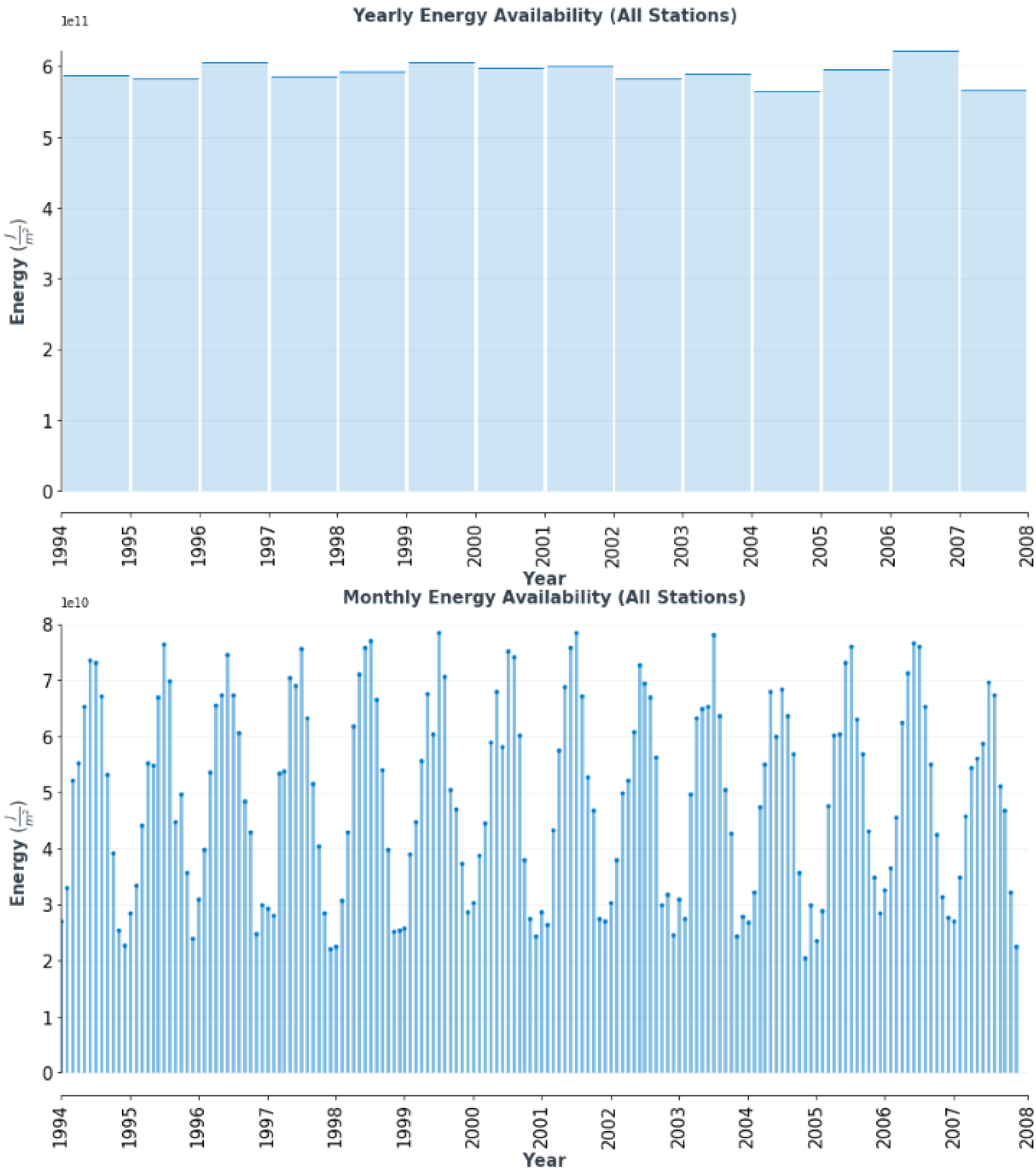

Figure 4: Yearly and monthly energy availability for all stations combined.

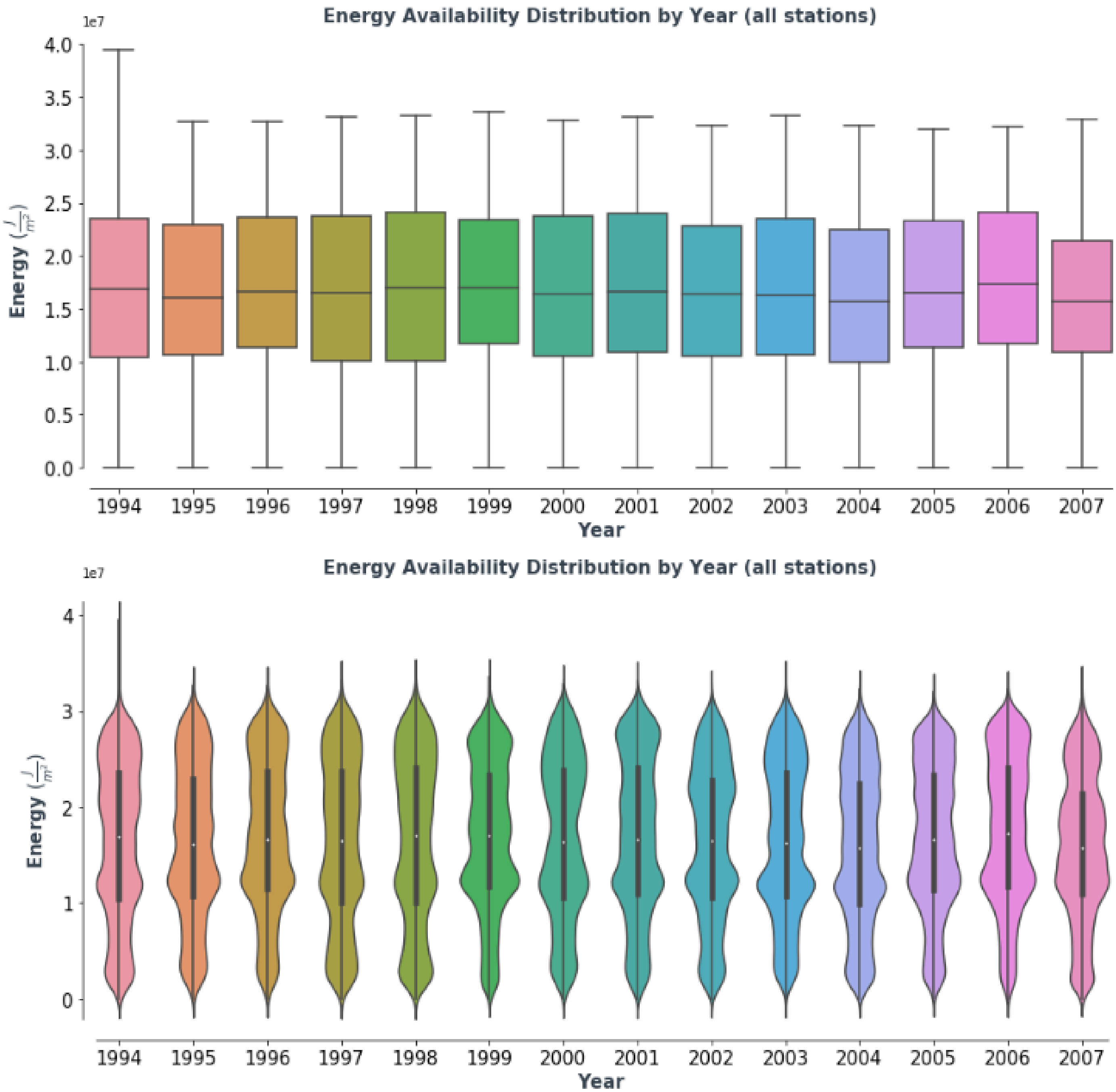

Figure 5: Daily energy availability distributions by year for all stations combined.

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.

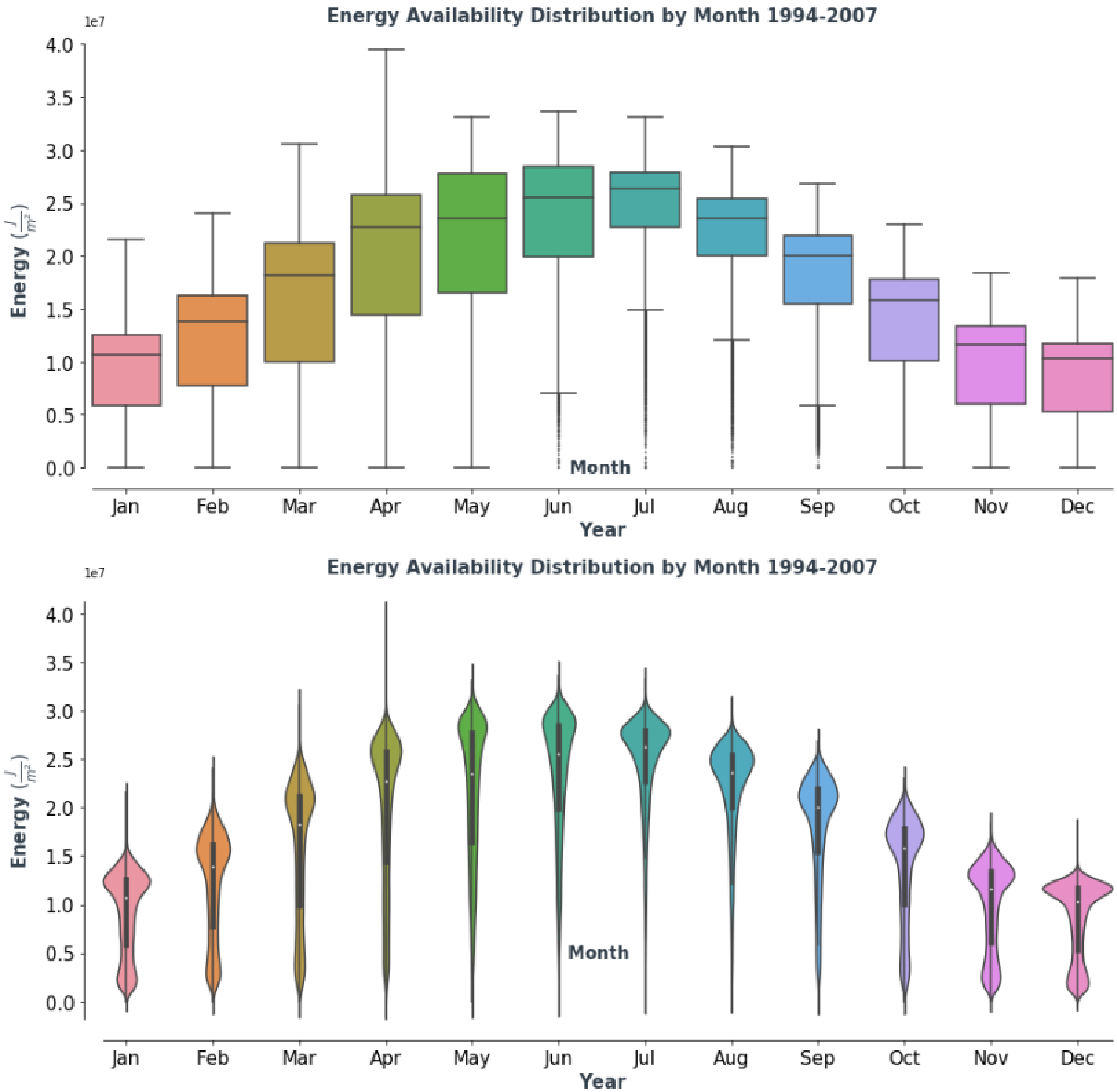

Figure 6: Daily energy availability distributions by month for all stations combined.

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.

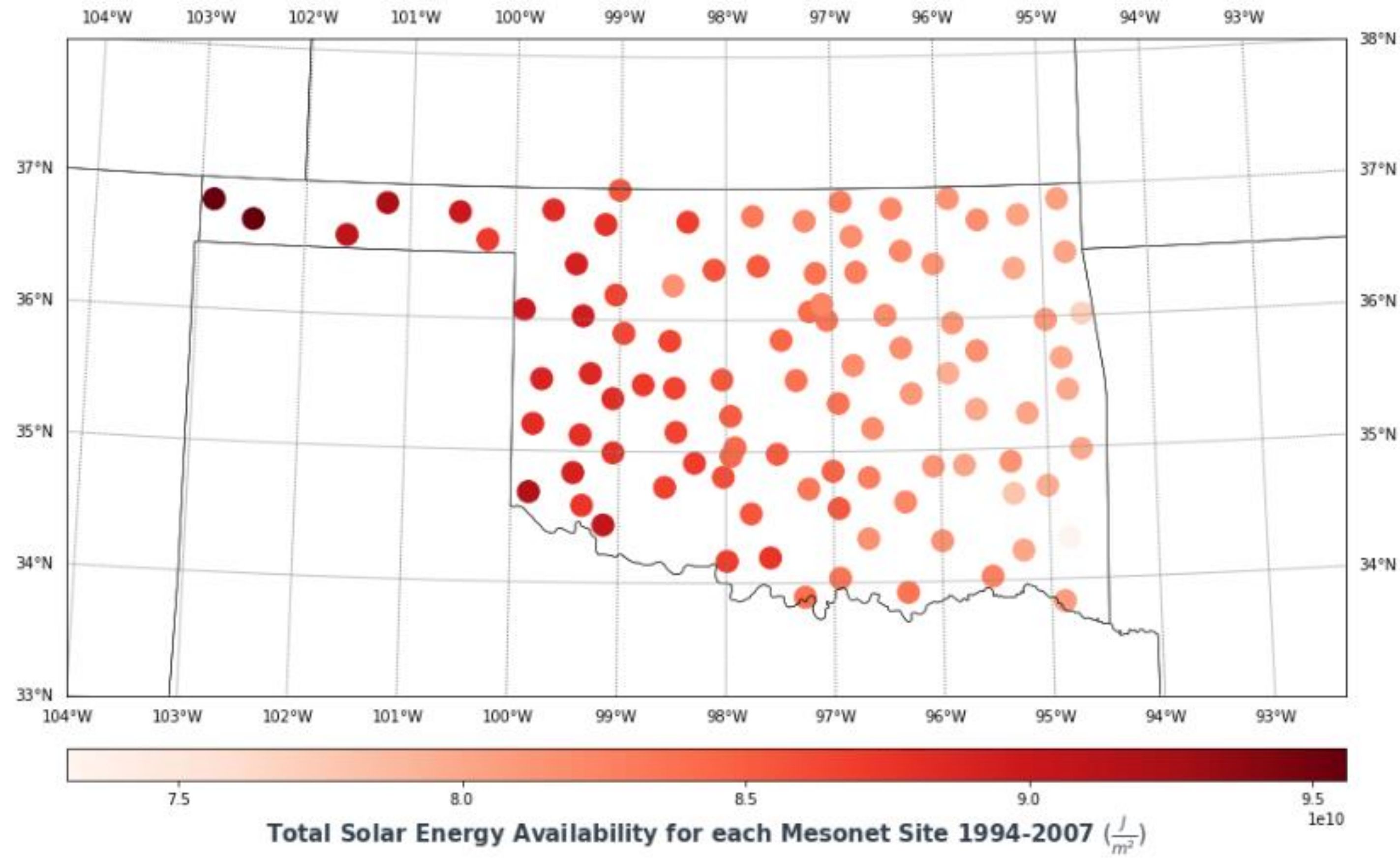

Figure 7: Total solar energy availability for each station.

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.

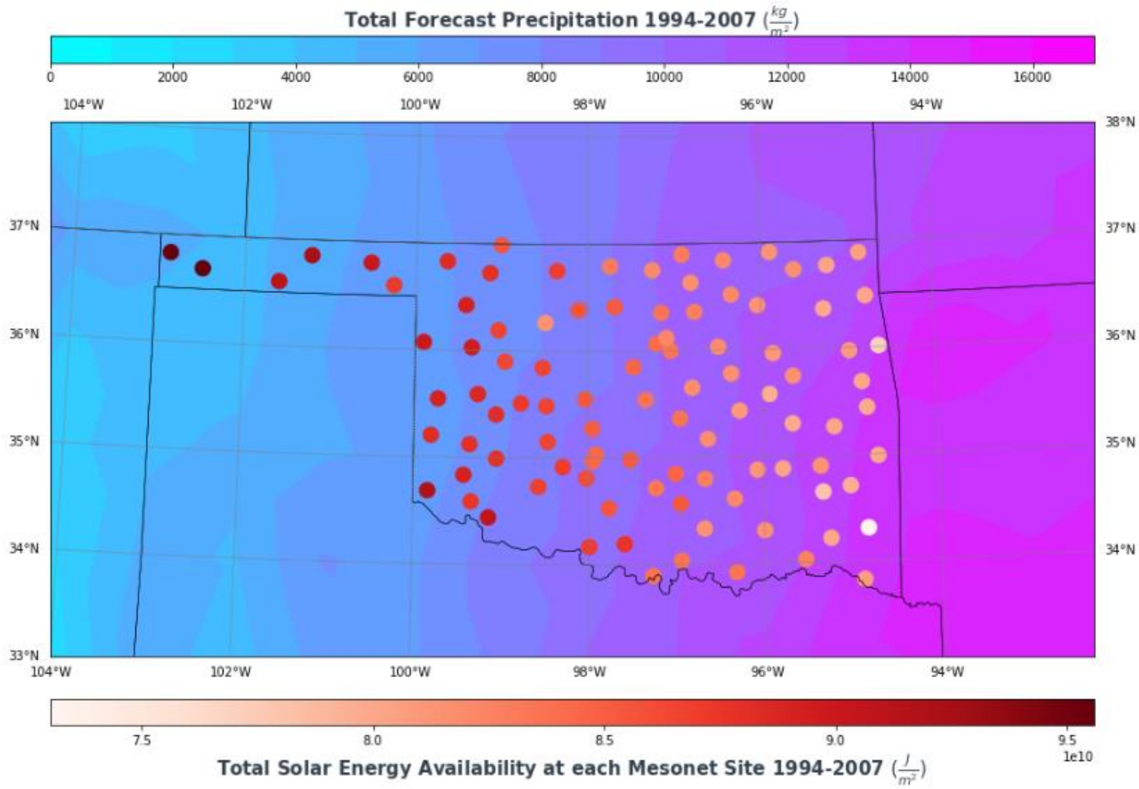

Figure 8: Total forecast precipitation and total solar energy availability for each station.

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.

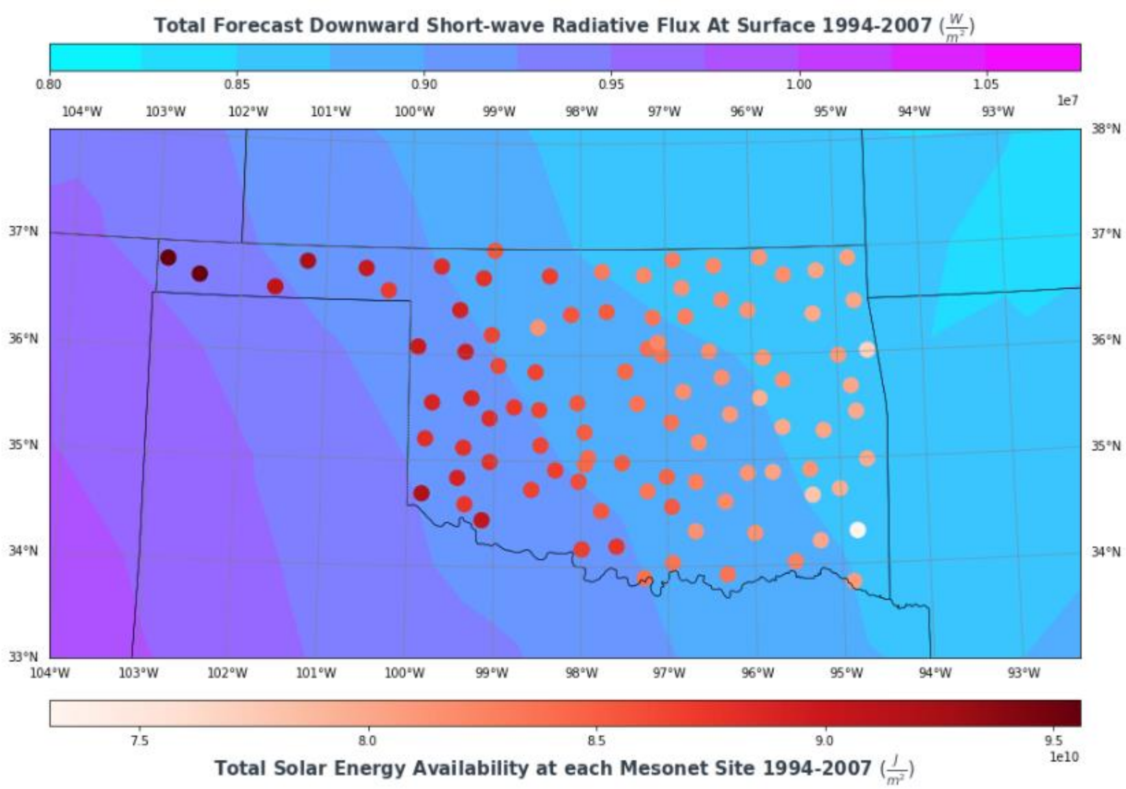

Figure 9: Total forecast short-wave radiative flux and total solar energy availability for each station.

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

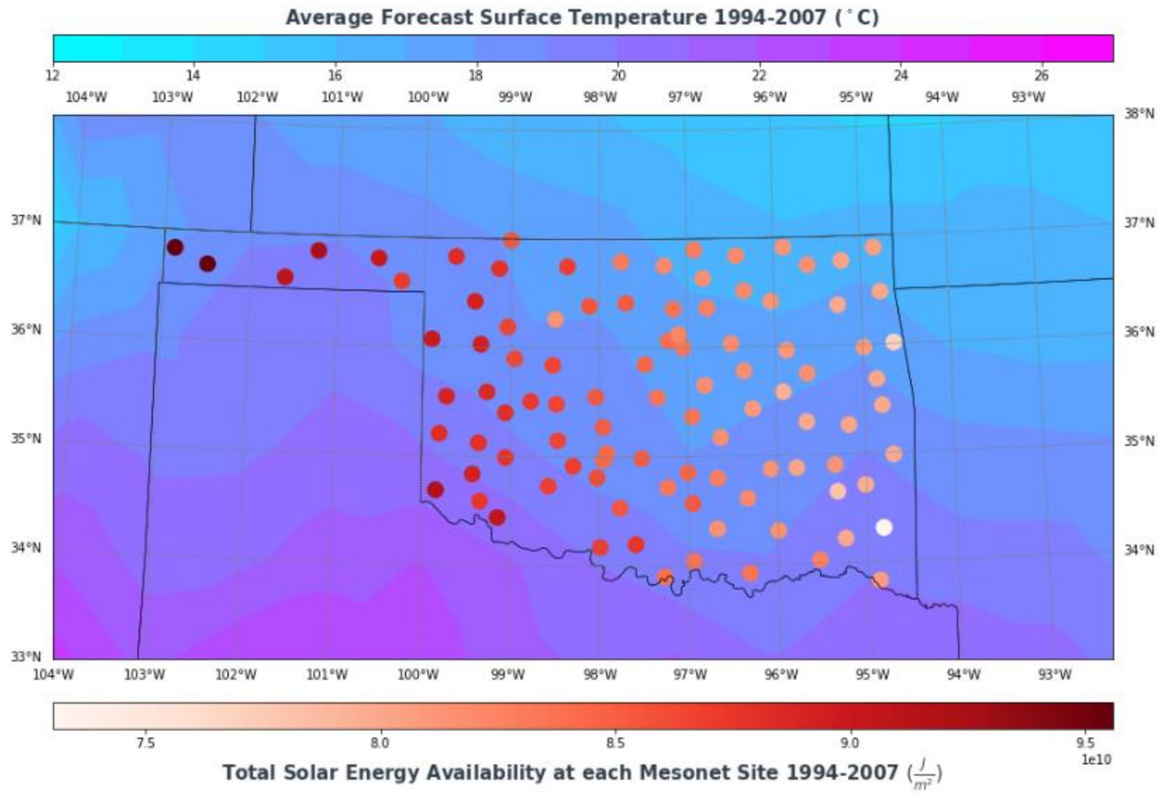

Figure 10: Average forecast temperature and total solar energy availability for each station.

3.2 Statistical data analysis

3.2.1 Correlation matrix

The Pearson correlation coefficient has been calculated for between each weather variable and energy availability (the eventual machine learning target variable) and between each weather variable and all other weather variables. The resulting values can be visualized in a correlation matrix (Figure 11).

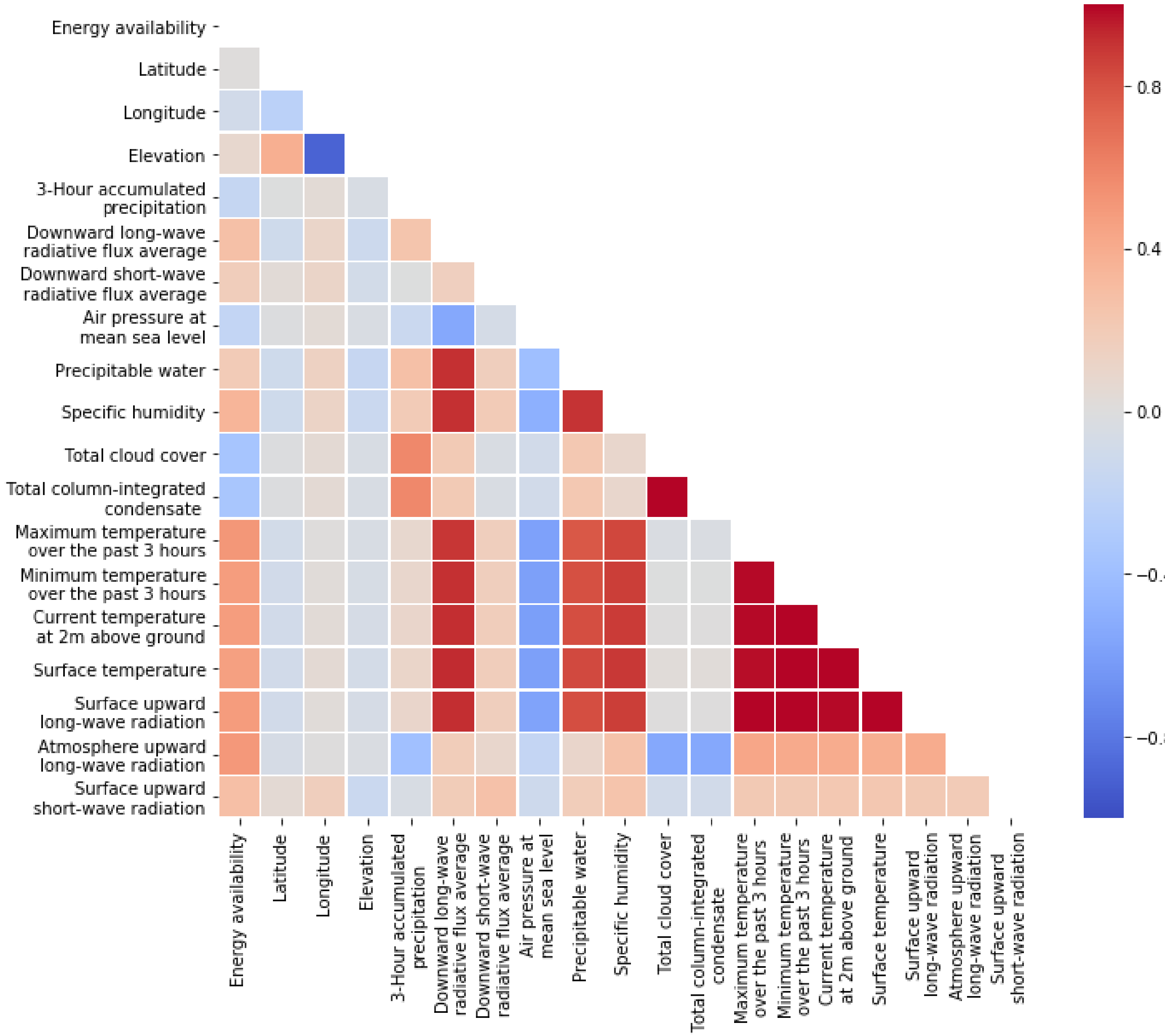

Figure 11: Correlation matrix (Pearson correlation coefficient) for energy availability and weather variables at forecast hour 0.

Energy availability is correlated with almost all variables. The weakest correlations are with latitude and elevation. The strong correlations indicate that the variables can be used in a machine learning model to predict energy availability. Between weather variables there are many strong correlations. These correlations will be considered during feature selection.

3.2.2 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 12.

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.

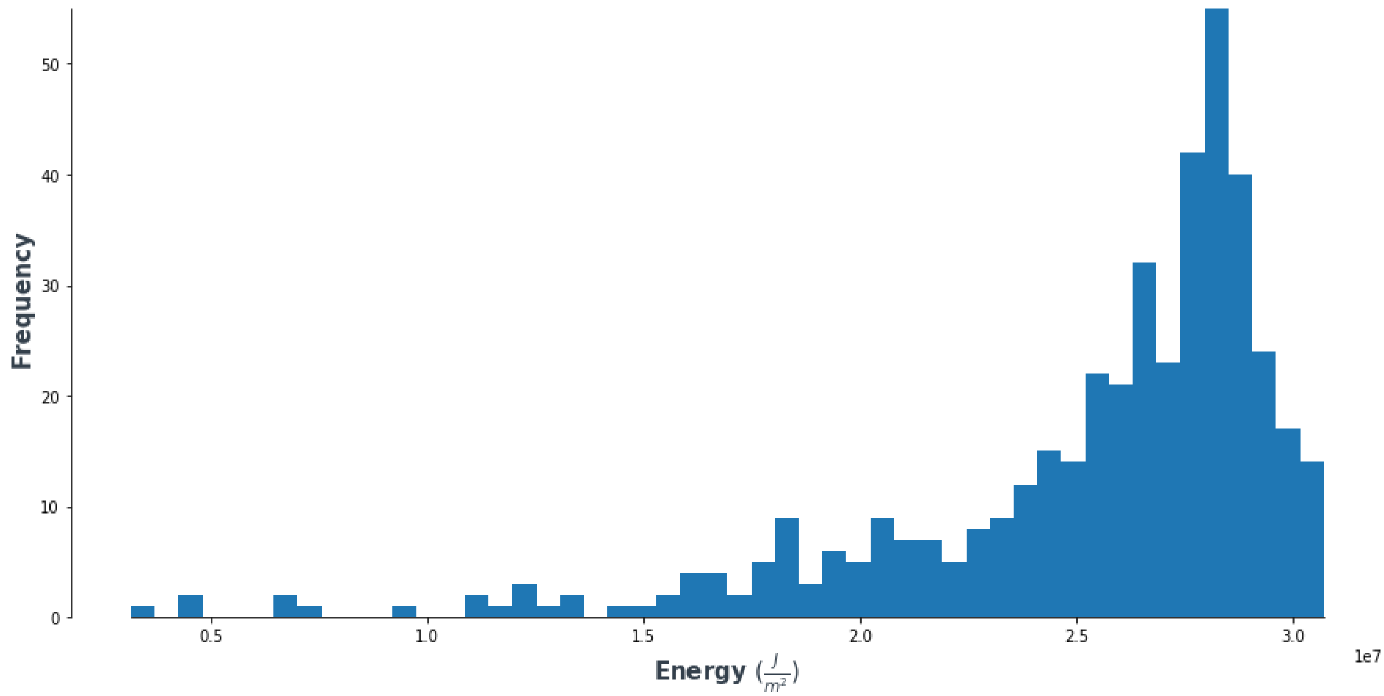

Figure 12: CHER station July daily solar energy availability 1994-2007

The distribution of bootstrap replicates of the mean is shown in Figure 13. 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.

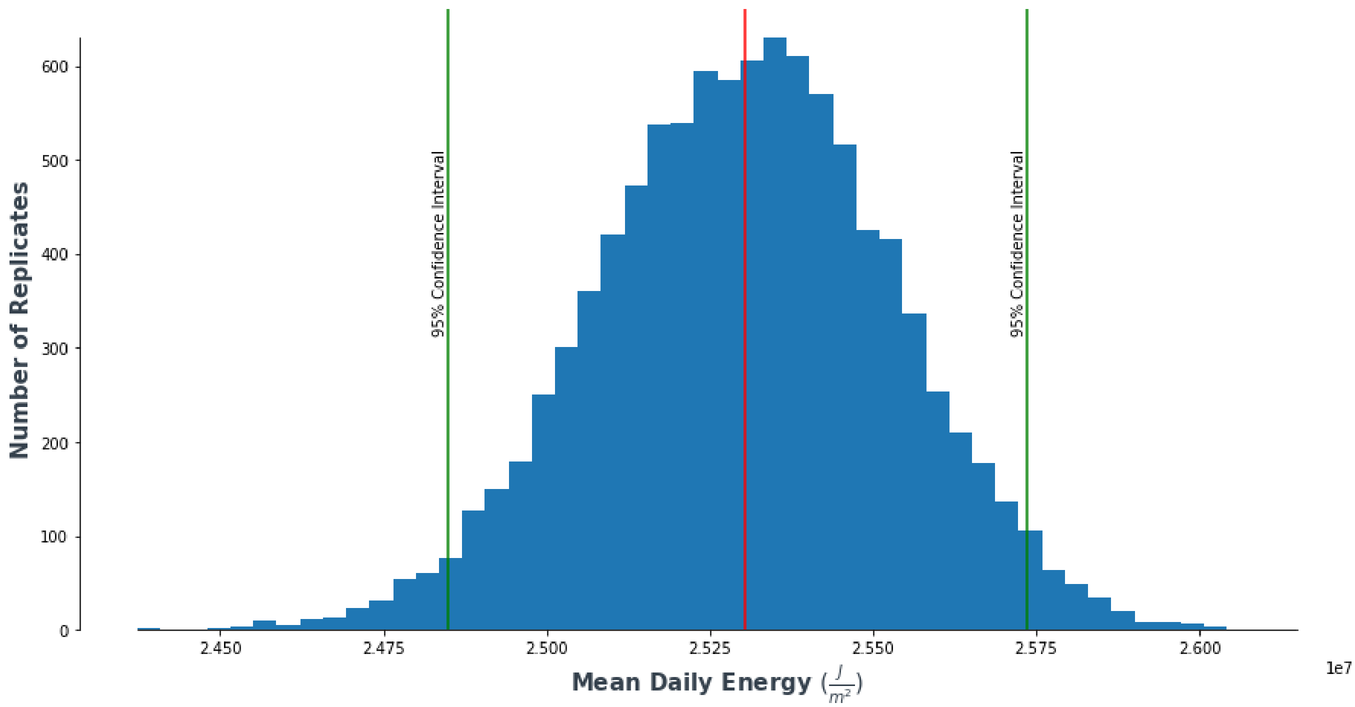

Figure 13: CHER station July daily solar energy availability bootstrap replicates of mean.

3.2.3 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 14).

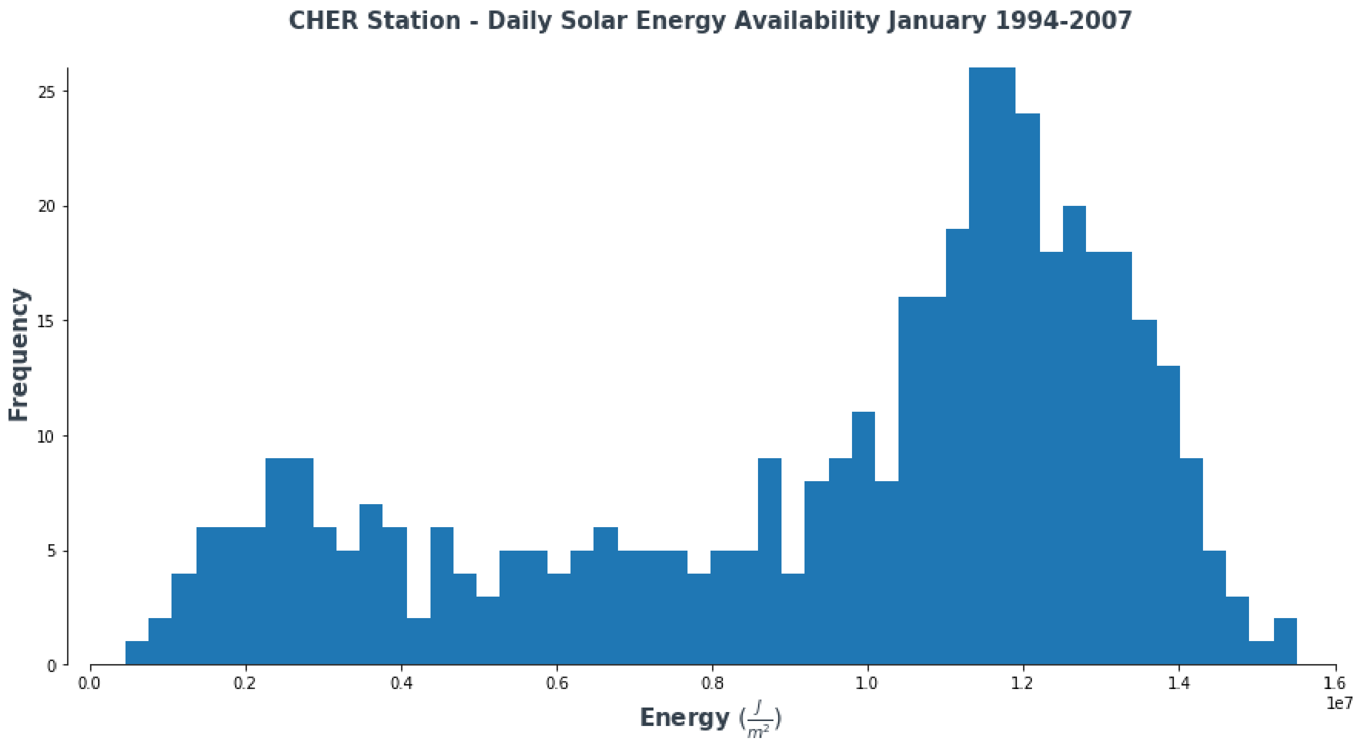

Figure 14: CHER station January daily solar energy availability 1994-2007

Figure 15 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.

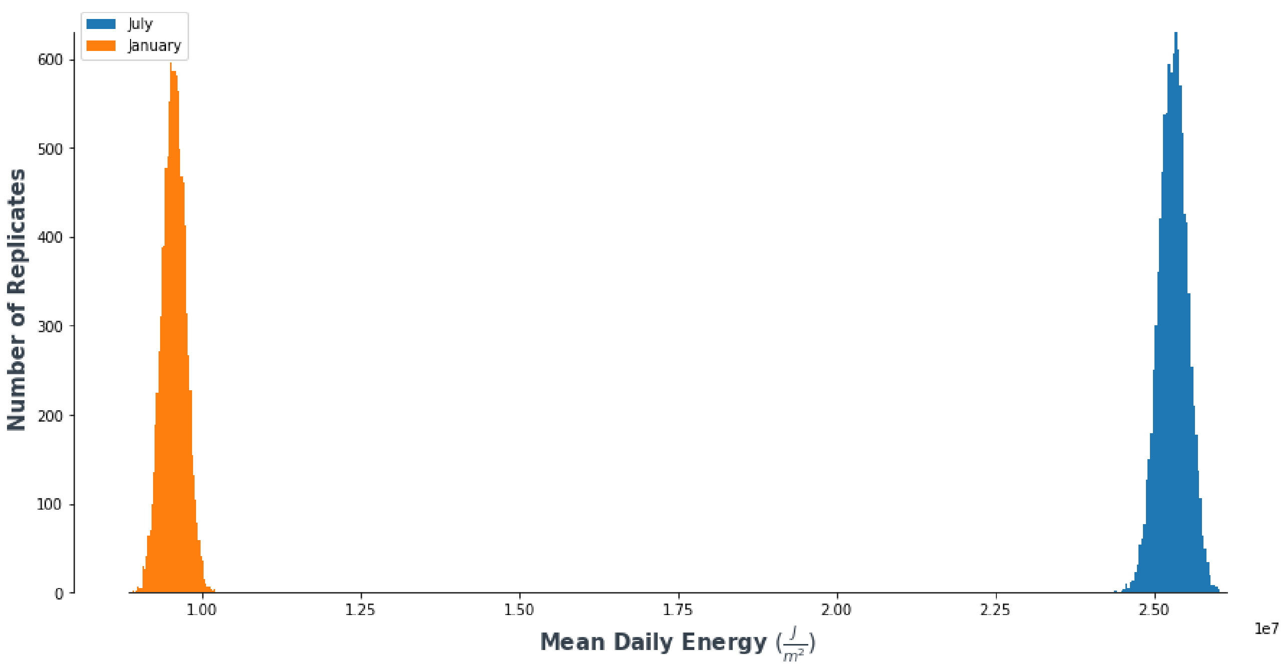

Figure 15: CHER station January and July daily solar energy availability bootstrap replicates of mean.

3.2.4 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 14) 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 16) and also simulate data (Figure 17). 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.

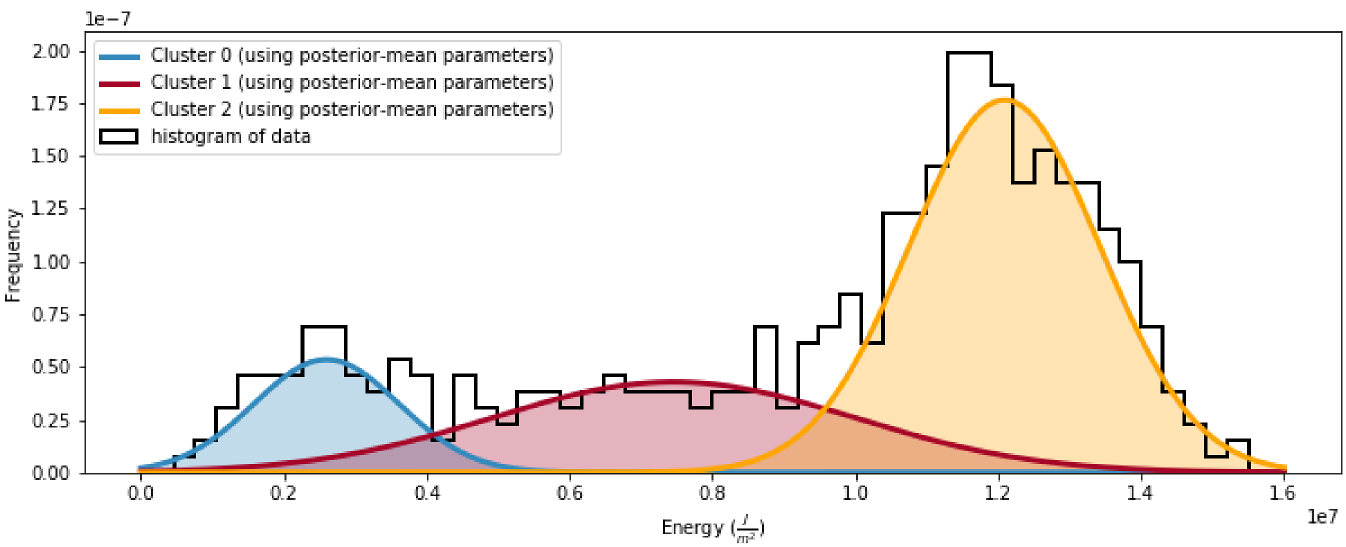

Figure 16: Posterior-mean parameter clusters visualized alongside a histogram of the data.

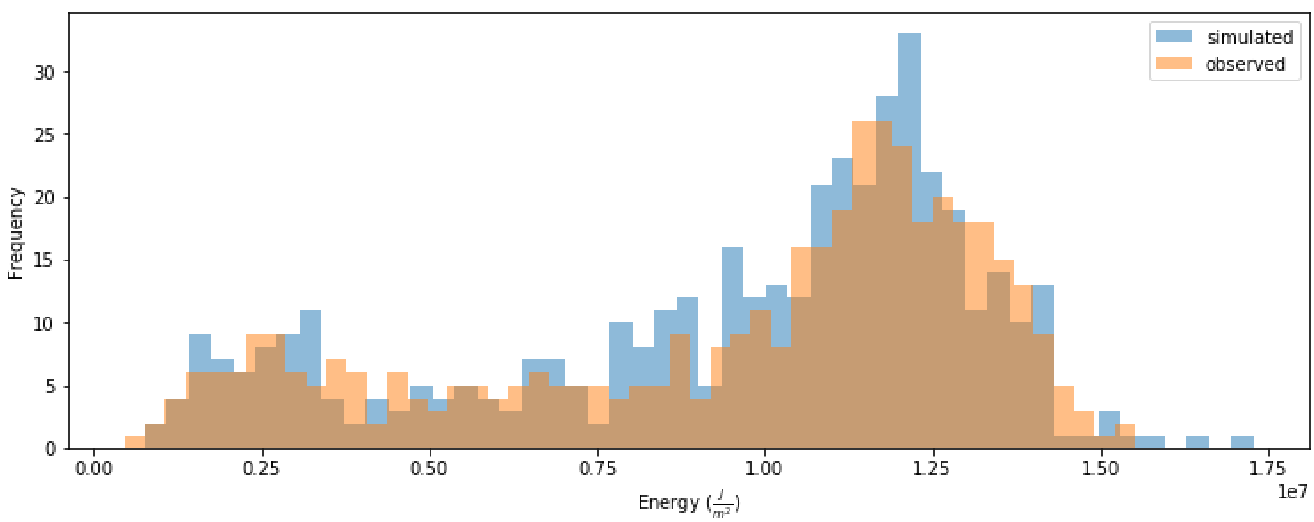

Figure 17: Data simulated using posterior-mean parameters compared with observed data.

4. 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.