Contents

1 Introduction **3**

2 Data Sets **4**
2.1 Global Horizontal Irradiance Data 4
2.2 Global Forecast System Data 4

3 1-Hour Ahead Solar Forecasting via Linear Regression **5**
3.1 Adding Predictor Variables 7
3.2 Partitioning Data 8
3.2.1 Temporal Partitioning 8
3.2.2 Spatial Partitioning 9
3.3 Data Transformation 9
3.4 Results 10

4 Rare Weather Event Detection **13**
4.1 Discretizing Events 13
4.1.1 Principal Component Analysis 14
4.1.2 Discrete Wavelet Transform 14
4.2 Identifying Rare Events 15
4.3 Multi-Day Events 15
4.4 Results 16

5 Conclusion **17**

6 Future Work **17**

Short-Term Solar Irradiance Forecasting and Weather Analysis using Gridded Data

Todd K. Taomae
Advisor: Lipyeow Lim

Master of Science in Computer Science
Plan B Final Report
Fall 2015

Abstract

While solar energy has definite advantages over conventional power sources such as oil and coal, its unpredictable nature poses a number of potential problems for the electrical grid. With conventional generators, grid operators can relatively easily increase production to meet demand. However, the unpredictable nature of solar energy can make it difficult to properly balance production and demand which can lead to grid instability. Accurate solar irradiance forecasts can reduce the unpredictability and help to ensure a stable grid.

In this paper we present a method for short-term solar irradiance forecasting using gridded global horizontal irradiance (GHI) data, estimated from satellite images. We use this data to first create a simple linear regression model with a single predictor variable. We then discuss various methods to extend and improve the model. We found that adding predictor variables and partitioning the data to create multiple models both reduced prediction errors under certain circumstances. However, both these techniques were outperformed by applying a data transformation before training the linear regression model.

We also discuss a set of methods for identifying "rare weather events" by discretizing Global Forecast System (GFS) data by transforming the data using either principal component analysis or a discrete wavelet transform, then discarding and round the transformed data. While the current work only investigates identifying past events, future work could investigate predicting these events and using those predictions to improve solar irradiance forecasts.

1 Introduction

Using renewable energy sources such as solar and wind have obvious benefits for the environment. However, they can cause technical problems for the electrical grid and its operators. This is particularly true for Hawaii which is not only geographically isolated from the rest of the United States, but even within the state, each island's electrical grid is isolated from the others. The fact that each grid is isolated means that each grid must be self-sufficient and cannot borrow power from nearby interconnected grids.

It is the job of electrical grid operators to ensure that the amount of power being generated matches the needs of the consumers. Meanwhile, they also want to minimize the amount of surplus power being generated in order to minimize costs. Without factoring in renewable energy, the main concerns are the consumer demand and the amount of energy that can be produced by their generators as well as the amount of time it takes for those generators to start up and reach full capacity. The output of their generators is obviously known in advance and the general patterns of consumer demand are also well understood. There are of course exceptions which might cause spikes or drops in usage such as high heat and humidity leading to increased air conditioner usage.

When we also consider renewable energy, there is much more uncertainty involved since the amount of power generated is tied to the weather. This means that even if consumer usage was perfectly predictable, an unexpected drop in sunlight — and therefore power generated by solar farms and rooftop photovoltaic systems — would make it difficult for grid operators to meet the consumer demand and ensure grid stability. Accurate solar forecasts can help reduce uncertainty and ensure grid stability.

Solar irradiance forecasting can occur on timescales anywhere from 5 minutes to 14 or more days using a variety of methods. These methods can generally be broadly grouped into one of three different types. First there are statistical methods which are based on historical solar irradiance data. Another set of methods are based around cloud motion determined either from satellite images or from ground-based sky images. The last type of solar forecasting is based on a technique known as numerical weather prediction which uses observed weather variables as input into computer models which try to forecast the future state of the weather. A summary and comparison of a number of methods can be found in [1].

This paper presents a method which falls into the statistical category. Using linear regression models we make approximately 1-hour ahead solar irradiance forecasts from global horizontal irradiance data which has been estimated from satellite images. We start with a simple model using a single predictor variable and extend the model by including additional predictor variables, partitioning the data to creating multiple models, and transforming the data to account for differences at different times of day and different locations.

We also discuss identifying "rare weather events" from Global Forecast System data by discretizing the data. The current work only focuses on identifying past events, but the goal is to eventually incorporate rare events into the solar forecasts.

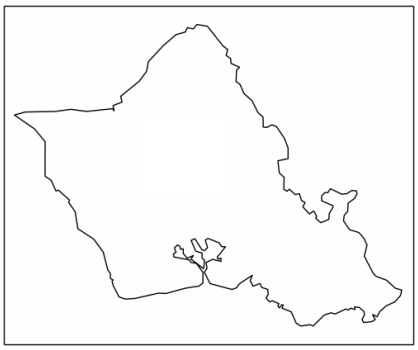


Figure 1: GHI map

2 Data Sets

2.1 Global Horizontal Irradiance Data

The first data set consists of Global Horizontal Irradiance (GHI) data for a region surrounding Hawaii for the years 2013 and 2014. GHI is the total solar radiation received by a surface horizontal to the ground and is measured in watts per square meter (W/m^2). The data was provided by AWS Truepower. While the exact methods used to obtain the data are not known, we do know that the GHI is estimated from satellite images using a method described by [2].

The data set contains GHI information for a grid of 1,120 by 1,040 points at a 2 kilometer resolution, centered over Hawaii. However, for this paper we will focus on a 36 by 29 grid over Oahu, which is shown in Figure 1.

While the data set contains data for most of the day, we will focus on only the subset between 8AM and 5PM Hawaii Standard Time (HST). The reason for this choice is that we are guaranteed to have non-zero GHI throughout the entire year during this time interval. Most of the data is provided at 15 minute intervals; however, there are some gaps of 30 minute intervals. Table 1 shows the times at which GHI data is available, within the window in which we are interested. The rows identify the hour in HST and the columns identify the minute within the hour. There is also some data missing throughout both years resulting occasional gaps greater than 30 minutes and up to several hours.

We will use the following notation to represent the GHI at the grid coordinate (x, y) at time t.

$$S(x, y, t)$$

While specific values are not of interest to us, this notation allows us to easily discuss relative times and location. For example, $S(x + 1, y - 1, t - 60)$ refers to the GHI at the grid point one unit to the east and one unit to the south of (x, y) at the time 60 minutes before t.

2.2 Global Forecast System Data

The second data set comes from the Global Forecast System (GFS), which is a weather forecast model produced by the National Centers for Environmental Prediction (NCEP). The data consists of dozens of weather variables, some of which are available at multiple

	00	15	30	45
08	x		x	x
09	x	x	x	x
10	x	x		x
11	x		x	x
12	x	x	x	x
13	x	x	x	x
14	x		x	
15	x	x	x	x
16	x	x	x	x

Table 1: Times available for GHI data

Name	**Description**	**Units**
PWAT	Precipitable water	kg/m^2
RH	Relative humidity	%
TMP	Temperature	K
UGRD	U-component of wind	m/s
VGRD	V-component of wind	m/s
VVEL	Vertical velocity (Pressure)	Pa/s

Table 2: GFS variables

altitudes. The variables that we are interested in are listed in Table 2. Throughout this paper we will refer to these variables by the names listed in the table.

With the exception of PWAT, all the variables that we are interested in are available at multiple altitudes. The altitude is measured by pressure altitude and in particular we are interested in the data at 850 millibars, which is approximately 1500 meters. PWAT represents the amount of water present in a column of the atmosphere, which is why there is no data for different altitudes.

GFS data is available at several different resolutions across the entire globe. Historical data is available at 1 degree or 0.5 degree resolution and more recent data is available at 0.25 degree resolution. We are only interested in the 1 degree resolution data for the years from 2010 to 2014. The region that we are interested in is shown in Figure 2.

The data is available at 6 hour intervals. Specifically it is available for 00:00Z, 06:00Z, 12:00Z, and 18:00Z of each day.

3 1-Hour Ahead Solar Forecasting via Linear Regression

In this section we describe a method for making approximately 1-hour ahead predictions of the solar irradiance at a given location. We will call the time for which we wish to predict the GHI, time t. We will use information available to us at time $t-n$, where n is the number

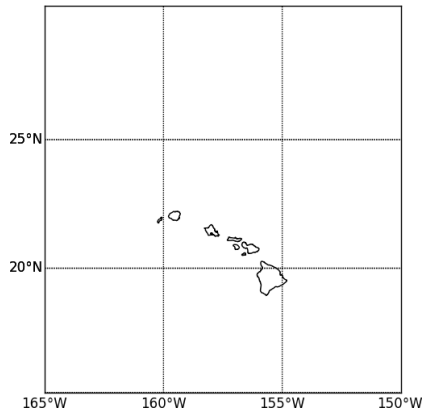


Figure 2: GFS map

of minutes prior to t at which we are making the prediction. Intuitively, one might expect that the solar irradiance at $t - n$ at the same location would be a good predictor for our target. This is the basis on which we build linear regression models for forecasting GHI.

Linear regression is a method for modeling the relationship between a dependent variable and one or more predictor variables. In our case, the dependent variable is $S(x, y, t)$ and, in the simplest example, the predictor variable is $S(x, y, t - n)$. In this example with a single predictor variable, we wish to create a model as shown in Equation 1 with constants c_1 and c_0 that will minimize the sum of the squared residuals, by way of ordinary least squares. The residual is the difference between the observed $S(x, y, t)$ and the value estimated by the linear regression model. This model will be based on all possible $S(x, y, t)$ and $S(x, y, t - n)$ for the region surrounding Oahu (shown in Figure 1) during 2013.

$$S(x, y, t) = c_1 S(x, y, t - n) + c_0 \tag{1}$$

This can also described visually. Figure 3 is a scatter plot of the GHI for a given point at time t versus the GHI at that same point at $t-60$ for a portion of the data set. The residual is the vertical distance between a point and the line given by Equation 1. So the goal of a linear regression is to choose c_1 and c_0 such that we minimize the sum of the squared vertical distances.

The model described above relies on a few assumptions. We are assuming that the GHI patterns are the same for all points on the grid, for all times of the day, and for all days of the year. However, this is obviously not true. For example, during the mornings GHI will generally trend upward while in the afternoons it will trend downward. While these assumptions allowed us to make a simple model, we will show later that it results in relatively poor performance. Throughout the remainder of this section, we will discuss various techniques to potentially increase the accuracy of our model. These techniques all work independently of each other and can be used individually or in combination with each other. First we will discuss including additional predictor variables, followed by partitioning the data and creating separate models for each partition of data, and lastly we will discuss a technique where we transform the data and use the transformed data to create the models.

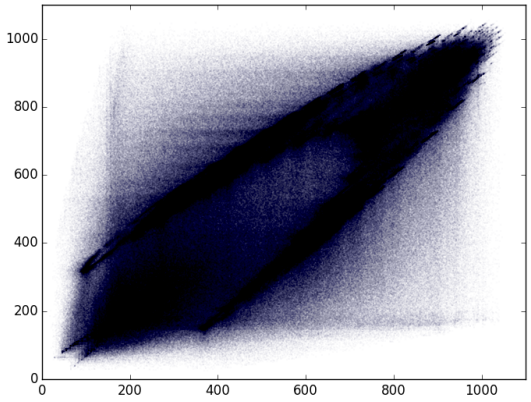


Figure 3: $S(x, y, t)$ vs. $S(x, y, t - 60)$

3.1 Adding Predictor Variables

When using the simple model described above, there may be important information that is not captured by using only a single predictor variable. We can add predictor variables by either expanding temporally or spatially.

If we expand temporally, we would include more past data. Instead of using only data at $t - n$, we can, for example, use $[t - n_0, t - n_1, \ldots, t - n_m]$. Since we are including additional predictor variable, obviously our original model will no longer work. Instead, we will have a model that looks like Equation 2 which uses data from k past times and will have constants c_k and c_0.

$$S(x, y, t) = \left(\sum_{k=1}^{m} c_k S(x, y, t - n_k) \right) + c_0 \tag{2}$$

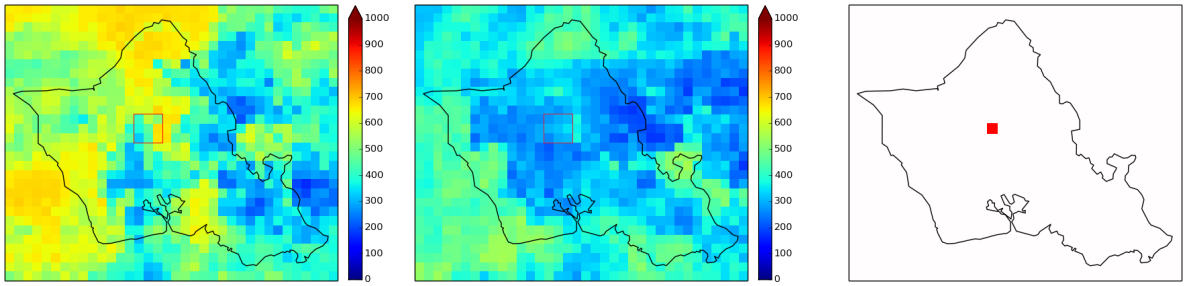
If we expand spatially, we include data from neighboring grid points. We define r as the "radius" of the surrounding region, which will be centered on (x, y). The model is now defined by Equation 3. Constants are not indexed by i and j.

$$S(x, y, t) = \left(\sum_{i=-r}^{r} \sum_{j=-r}^{r} c_{i,j} S(x + i, y + j, t - n) \right) + c_0 \tag{3}$$

We can also combine both of these techniques and expand both temporally and spatially. The resulting model is given by Equation 4. We now index constants by i, j, and k.

$$S(x, y, t) = \left(\sum_{i=-r}^{r} \sum_{j=-r}^{r} \sum_{k=1}^{m} c_{i,j,k} S(x + i, y + j, t - n_k) \right) + c_0 \tag{4}$$

Figure 4 shows an example which uses both of these techniques. The red square in Figure 4c represents $S(x, y, t)$ and the red boxes in Figures 4b and 4a are our predictor variables.



(a) $t - n_1$: Predictors (b) $t - n_0$: Predictors (c) t: Target

Figure 4: Predictors and Target

3.2 Partitioning Data

In the original model, we made several assumptions which led to a single model for all data. However, it is unlikely that GHI patterns will be the same for all points on the grid, for all times of the day, and for all days of the year. In order to account for such differences, we can partition the data based on these differences and create separate models for each partition. This partitioning can be either temporal or spatial.

3.2.1 Temporal Partitioning

There are two main ways that we might partition the data temporally. We can partition on either a daily or yearly scale. On a daily scale, we might, for example, want to create a different model for each hour of the day, while on a yearly scale, we might create a different model for each month of the year. It is also possible to combine both of these which would leave us with a different model for each hour of each month.

While partitioning the data at this granularity could potentially provide more accurate models, it also means that we will have less data to train each model and we will have many more models. If we were to combine both hourly and monthly partitioning, we would end up with 108 different models, each with approximately 108 times less data (9 hours between 8AM and 5PM; 12 months in a year).

Rather than partitioning by each hour, a more reasonable approach might be to simply partition the data into morning and afternoon. This also has some intuitive justification since there is a clear distinction in general GHI patterns in the morning and afternoon. In the morning, we expect the GHI to generally be increasing, while in the afternoon it will tend to decrease. The hope is that these patterns can be more accurately captured by the linear regression models if they separated.

On the yearly scale, it is less obvious how to partition the data. Seasonal partitioning might be a good candidate. However, since the seasonal patterns are not as distinct in Hawaii as many other places we will focus on a bi-annual partition. We will refer to the combination of winter and spring as "winter" and the combination of summer and autumn as "summer." While using equinoxes and solstices as the seasonal boundaries might provide slightly more accurate models, for convenience, we will divide seasons on monthly boundaries. Each season will consist of three months, starting with winter consisting of December, January, and February and the rest of the seasons following in three month chunks.

3.2.2 Spatial Partitioning

In addition to partitioning temporally, we can also perform spatial partitioning. On a global scale, it makes sense that, for example, the GHI patterns in Hawaii will be very different from those in Alaska. However, since we are only considering Oahu, it is less obvious how we might partition the data.

One way that we might partition the data is by land and ocean. This particularly makes sense for our application since there are obviously no rooftop solar panels or solar farms in the ocean. Another, perhaps less obvious, way to partition the data is by elevation. There are two main motivations behind partitioning by elevation. First, more of the population lives at lower elevation. Therefore, there will likely be more rooftop solar panels in that region. The second motivation has a meteorological basis; due to warm air rising along the mountains, cloud formation is more likely at higher elevations as the air cools and condenses.

We obtained elevation data from the Google Maps Elevation API. From this we can easily partition each grid point by elevation, but we can also use this to approximate which grid points are on the land or ocean. We consider any point with an elevation of 0 meters or less to be part of the ocean and any point with greater than 0 meters elevation to be part of the land.

3.3 Data Transformation

In this section we describe a method for processing the GHI data which provides an alternative method to account for differences in time of day and location that does not reduce the amount of data that we have to create the model as was the case with partitioning. The transformed value is represented by the following notation.

$$\hat{S}(x, y, t)$$

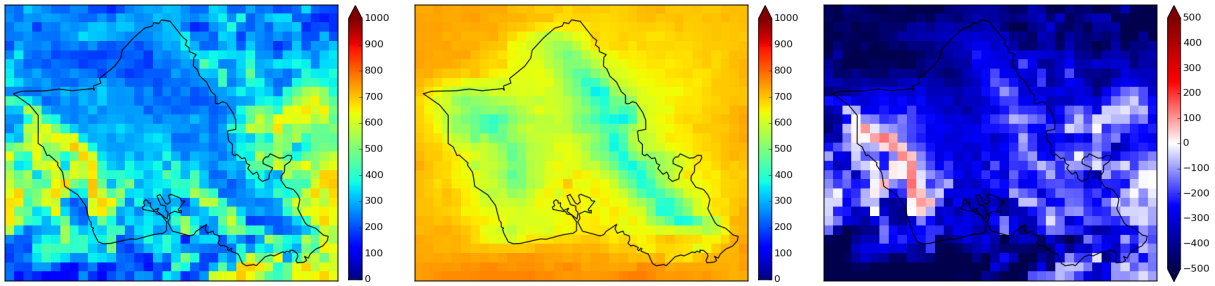
After we apply the transformation, the data can be used as before. For example, Equation 5 defines the linear regression model equivalent to the one give by Equation 1.

$$\hat{S}(x, y, t) = c_1 \hat{S}(x, y, t - n) + c_0 \tag{5}$$

Our transformation will be based on the deviation from some average value. Consider the following scenario. Suppose that the GHI at some location is 500 W/m^2. If this is in the afternoon, then this might be a normal value for that time of day, so we might expect that it will follow typical patterns. However, if this is in the middle of the day, then it might be lower than we expect, so we probably would not expect it to follow typical patterns. This is the motivation behind this model. We also take into account the fact that the typical value at a given time of day might vary from one location to the next.

Following the reasoning from the example above, our averaging should take into account both temporal and spatial information. This is accomplished by computing the average for each grid point at each time of day. We use the following notation to represent the average GHI for a given grid point at a time of day specified by $hh{:}mm$.

$$\bar{S}(x, y, hh{:}mm)$$



(a) Actual data for 2013-01-01T12:00HST

(b) Average for 12:00HST

(c) Deviation from average for 2013-01-01T12:00HST

Figure 5: Process for computing deviation

The data transformation simply subtracts the corresponding average value from the actual average. For example, $\hat{S}(x, y, \text{2013-01-01T12:00}) = S(x, y, \text{2013-01-01T12:00}) - \bar{S}(x, y, \text{12:00})$.

Figure 5 shows the process visually. Figure 5a shows the data for one specific day at noon and Figure 5b shows the average that was computed from all the data at noon. Lastly, Figure 5c shows the deviation from the average for one specific day at noon, obtained by subtracting the values in 5b from 5a. This is data that will be used for creating the linear regression model.

3.4 Results

Due do the number of combinations that are possible as well as the number of parameters that can be adjusted, it would be unreasonable to test all possible models. Also, due to our limited data set, it would not make sense to try all possible combinations to find one that is the best as it would likely be specific to our data set. Rather, it is more important to find which techniques seem to reduce errors in general.

This section will describe the performance of a number of different models. Each model is created using 2013 data only and predictions are made for 2014. It is important to note that when applying the data transformation, the average is based only on 2013 and when computing the deviation from the average, even for 2014 data, we will use the 2013 average. The performance will be measured by mean absolute error (MAE). In other words, the average of the absolute difference between our prediction and the actual irradiance. In the tables listing the results, we will also report the standard deviation in parentheses.

Tables 3, 4, and 5 show the performance of our method as we increase the amount of spatial or temporal data used. The labels on the top and left describe how many predictor variables are used in each model. The spatial data is labeled as $n \times n$ which describes the size of the surrounding region that we will use. The temporal data is labeled as a list of times used, relative to the target time. "(DT)" indicates that our data transformation was also used for that particular model.

From Table 3 we can see that the error of our model decreases consistently up to a 13×13 region. However, the improvement is largest when moving from 1×1 to 3×3 and anything beyond 5×5 seems to have very minimal impact. We have not tested using regions larger than 13×13 so it is no clear what will happen beyond that point. However, we suspect that

	[t-30]	[t-30] (DT)	[t-60]	[t-60] (DT)
1×1	109.16 (140.89)	89.83 (127.52)	149.66 (178.17)	109.18 (146.64)
3×3	103.15 (133.08)	83.15 (119.87)	146.11 (173.45)	103.55 (140.32)
5×5	101.96 (131.35)	81.65 (117.89)	145.38 (172.36)	101.99 (138.34)
7×7	101.39 (130.57)	80.91 (116.89)	145.00 (171.82)	101.07 (137.20)
9×9	101.13 (130.20)	80.53 (116.36)	144.82 (171.54)	100.90 (136.48)
11×11	101.01 (130.03)	80.33 (116.06)	144.72 (171.40)	100.08 (135.98)
13×13	100.95 (129.93)	80.21 (115.90)	144.67 (171.31)	99.78 (135.64)

Table 3: Increasing Spatial Data

	1×1	1×1 (DT)
[t-60]	149.66 (178.17)	109.18 (146.64)
[t-60, t-90]	150.33 (180.29)	109.93 (146.96)
[t-60, t-75, t-90]	149.05 (180.93)	110.15 (147.89)
[t-60, t-90, t-120]	145.70 (177.52)	113.16 (150.09)
[t-60, t-75, t-90, t-105, t-120]	136.94 (172.04)	118.32 (156.83)

Table 4: Increasing Temporal Data

at some point it will include too much information from too far away and it will begin to degrade the performance.

Table 3 also shows the impact of increasing from 30 to 60-minute predictions. The result is not very surprising. As we try to forecast farther into the future, the error increases.

Table 4 shows the impact of including more temporal data. What is interesting here is that when we do not apply the data transformation, using up to 90 minutes of temporal data does not have much impact, but using up to 120 minutes of temporal data reduces the prediction error by a significant margin. In contrast, if we do apply the data transformation, adding temporal data does not help and in fact increases error.

Table 5 shows the effect of including both temporal and spatial data. This table suggests that the two factors work mostly independently of each other. Adding spatial data helps in all cases, as it did in Table 3, and adding temporal data helps in the same way as it did in Table 4. Specifically, if the data transformation was not applied, then the additional temporal data seems to help

The results of applying temporal and spatial partitioning are shown in Tables 6 and 7.

	3×3	3×3 (DT)
[t-60]	146.11 (173.45)	103.55 (140.32)
[t-60, t-90]	141.27 (172.41)	105.69 (142.49)
[t-60, t-75, t-90]	140.50 (173.17)	106.45 (143.93)
[t-60, t-90, t-120]	134.95 (167.95)	108.82 (145.81)
[t-60, t-75, t-90, t-105, t-120]	130.94 (166.13)	114.53 (153.03)

Table 5: Increasing Spatial and Temporal Data

	Full Day	Morning	Afternoon
Full Year	149.66 (178.17)	109.65 (141.52) *[144.17]*	131.33 (163.59) *[152.53]*
"Winter"	144.38 (172.74) *[145.06]*	106.56 (138.74) *[134.36]*	127.13 (158.05) *[150.55]*
"Summer"	155.31 (183.12) *[153.40]*	112.64 (143.15) *[154.05]*	136.08 (169.39) *[154.58]*

Table 6: Temporal Partitioning

"Ocean" (elevation $\leq$ 0m)	152.31 (180.76) *[152.78]*
"Land" (elevation $>$ 0m)	144.37 (172.98) *[144.75]*
0m $<$ elevation $\leq$ 50m	146.95 (174.96) *[145.81]*
50m $<$ elevation $\leq$ 100m	146.19 (174.15) *[145.65]*
100m $<$ elevation $\leq$ 150m	146.68 (175.56) *[146.78]*
150m $<$ elevation $\leq$ 200m	147.53 (176.12) *[147.01]*
elevation $>$ 200m	140.90 (170.34) *[143.32]*
elevation $>$ 500m	129.06 (159.73) *[137.28]*

Table 7: Spatial Partitioning

These results are based on a model which uses a single predictor variable for 60-minute forecasts. Table 6 shows the performance of various combinations of daily and yearly partitioning and Table 7 shows the performance for various spatial partitions. In addition to the mean absolute error and standard deviation, we also report — in square brackets and italics — the mean absolute error of the equivalent non-partitioned model, for the partition in question. In particular, we use the results from the non-partitioned, non-transformed, single predictor variable model and rather than looking at the error across all predictions, we look at the error specifically for only morning hours, or only afternoon hours, etc. This will allow us to see if the performance is truly improving or if it simply due to the fact that it is naturally easier to predict for a particular partition.

We can see from Table 6 that temporal partitioning is useful in all cases except for the "summer" partition. We can also see that partitioning by morning and afternoon reduces errors much more than partitioning by seasons. This could be due to the fact that seasonal differences are not very significant in Hawaii.

Table 7 shows us that spatial partitioning is not very useful for our data set. While, in most cases, the partitioning does improve the error relative to the average error of the original model, when we compare it to the error for only that partition (the number in square brackets), the improvement is much less significant and in fact often performs slightly worse. The one exception is when elevation is greater than 500 meters. However, that only accounts for less than 10% of the total land and likely an even smaller portion of the population and rooftop solar panels.

Table 8 contains the results of applying temporal partitioning as well as data transformation. Again we see that the "summer" partition performs worse while the others perform better. However, the improvements are much less significant than applying those seen in Table 6. This is likely due to the fact that the data transformation already captures the same patterns that temporal partitioning tries to account for.

	Full Day	Morning	Afternoon
Full Year	109.18 (146.64)	104.31 (140.41) *[110.90]*	101.16 (137.18) *[108.27]*
"Winter"	107.14 (145.49) *[108.74]*	101.10 (137.32) *[109.91]*	100.62 (137.57) *[108.14]*
"Summer"	111.58 (147.40) *[109.62]*	107.80 (142.51) *[111.91]*	101.95 (138.02) *[108.41]*

Table 8: Temporal Partitioning with Data Transformation

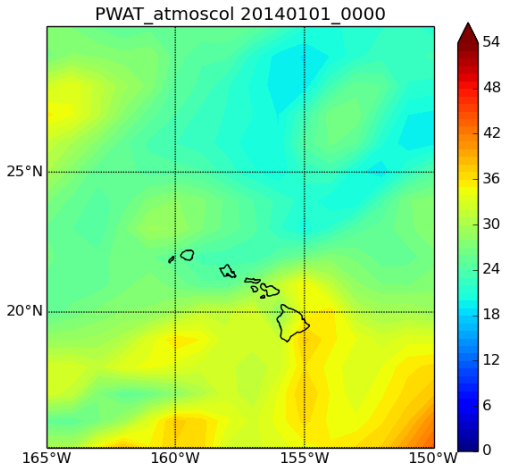


Figure 6: PWAT event for 2014-01-01T00:00Z

4 Rare Weather Event Detection

In this section we describe methods for identifying "rare weather events" from GFS data. We define an *event* simply as a grid of GFS data for a particular variable at a specific date and time. For our purposes, we will focus on a 15° × 15° region, approximately centered on Hawaii. Specifically, the northwest corner of the region will be at 30°N and 165°W and the southeast corner will be at 15°N and 150°W. Figure 6 is an example of an event for the region that we are interested in. It is important to keep in mind that while the region is 15° × 15°, this leaves us with a grid of 16 × 16 points.

Since each event consists of many elements of continuous data, it is very unlikely that two events will be the same. If all events are distinct, then we have no way to distinguish between rare and common events. In a sense, they are all equally rare. In order to have a more meaningful distinctions of rare and common events, we must group similar events together by discretizing the events. Once we have discretized the events, we identify a rare event as one which has few occurrences.

4.1 Discretizing Events

Discretization is accomplished by first transforming the data into a more useful feature space using either principal component analysis (PCA) or a discrete wavelet transform (DWT), followed by rounding the transformed data until the events are no longer all unique.

The reason we perform the transformation first is that when looking at the unprocessed data, we do not know which features are most important. Both PCA and DWT transform the data in a way that allow us to hopefully perform the rounding without losing important aspects of the event.

4.1.1 Principal Component Analysis

Principal component analysis is a method for transforming the data into linearly uncorrelated variables called *principal components*. The transformation is defined such that the first principal component will have the greatest variance, and each subsequent principal component will have less variance than the previous. The hope is that if we discard later principal components we will still keep the important features of the events.

PCA takes as input a $n \times p$ matrix $\mathbf{X}$, where n is the number of samples in the data set and p is the number of features per sample, and each row is a feature vector. In our case each feature is the value for a GFS variable at a particular grid point. Since our data is a 2-dimensional grid, we must transform it into a 1-dimensional feature vector. This can be accomplished by simply appending all rows of the grid into a 1-dimensional vector. The exact method for turning the grid into a 1-dimensional vector is not important as long as one method is used consistently for all data.

As output, we are left with the transformed data as a $n \times p$ matrix $\mathbf{T}$, where each row is now principal components of the original vector. If we keep all principal components then the rows will still be unique since we have not lost any information. By discarding principal components we are essentially projecting the original data onto a lower dimensional space. In our case, even after dropping all except the first principal component, all samples are still unique.

Since we still cannot distinguish rare events, we can further reduce the amount of information in the data by performing some kind of rounding or truncation of the first principal component. The approach we took was to round down to the nearest multiple of some rounding factor. For example, if our rounding factor is 7, anything in the range $[7, 14)$ would be rounded to 7, anything in the range $[14, 21)$ would be rounded to 14, and so on.

This approach is limited by the fact that an appropriate rounding factor will depend on the data. Depending on the input data, the range of possible values for the transformed data could be arbitrarily small or large. If, for example, the transformed data is all in the range $[0, 5]$, then a rounding factor of 7 would round everything to 0. On the other hand, if the range is $[0, 5000]$, then a rounding factor of 7 may leave use with too many unique values. In section 4.2 we will introduce a method for choosing a rounding factor.

4.1.2 Discrete Wavelet Transform

Discrete wavelet transform is another method for transforming a set of data. A major difference between PCA and DWT is that DWT can be performed on a single sample in a data set independently of any other samples. In contrast, PCA is dependent on all samples within a data set and the results will vary if a subset of the data is used or if additional samples are added to the data set. Another difference is that DWT can be applied to either 1-dimensional or 2-dimensional data.

In order to use a 1-dimensional DWT, we must convert the gridded data into a 1-dimensional vector. In this case, the method we use to convert the data is important because wavelet transforms operate on adjacent elements in the feature vector. So, as much as possible, we want to keep adjacent grid points adjacent within the resulting vector. This can be accomplished by using the method described previously. That is, simply append each row of the grid, in order, into a single vector.

The result of a DWT is a set of *approximation coefficients* and a set of *detail coefficients* where each set is half the size of the original. The process can be recursively applied to the approximation coefficients until there is a single approximation coefficient and many detail coefficients. This is a lossless process which means that all information is retained and we can also reverse the process to get back our original data. This also means that the events will still be unique. Unlike with our PCA method, we will not discard any data, but we will perform the same form of rounding.

There are different wavelets that could be used. So far, we have only used the Haar wavelet, which is the simplest possible wavelet.

4.2 Identifying Rare Events

In order to identify rare events we must choose a rounding factor f that we will use for rounding the transformed data as well as a "rarity threshold" r which tells us how uncommon an event must be in order to determine if it is rare. If a discretized event has fewer than r occurrences, then those occurrences will be considered.

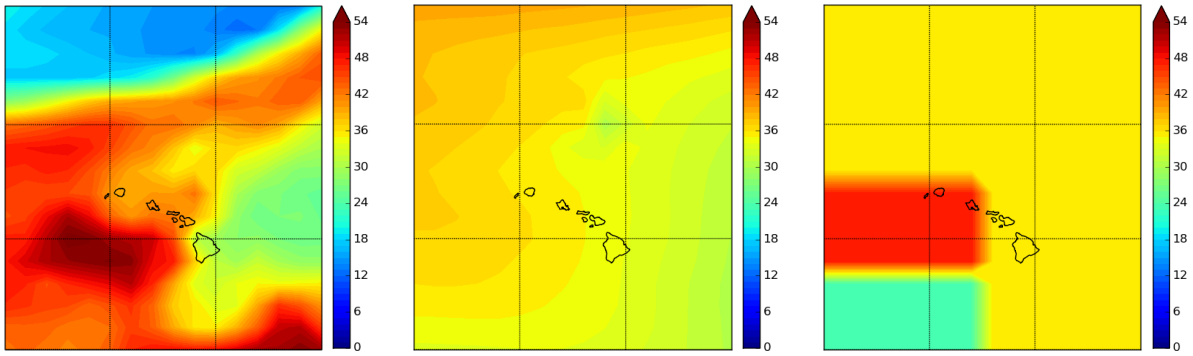
As we mentioned previously, the choice for f will depend on the input events and the resulting transformed data. If the data has a large range of possible values, then a larger rounding factor will be needed. The choice for r is somewhat arbitrary but should generally be small. One way to choose r is to use a percentage of the total number of events. For example, if we have 5000 total events, we can set r to 5, which is of 0.1% of the total number of events.

Since f will vary depending on the input, we propose a method for automatically selecting a value. We start by choosing a value for r as well as the approximate number of desired rare events. As we increase f, more events will be grouped together since we are performing more rounding. Since events are not more likely to be the same as others, there will be fewer events that are below the rarity threshold, meaning that we will have fewer rare events. Inversely, if we decrease f, fewer events will be grouped together and we will have more rare events. In order to quickly find a value for f we can use a binary search until we have our desired number of events.

Rather than simply using all events for a single variable as input, we may be interested in only events at a specific time. For example, the events that occur at 06:00Z (20:00HST) may not be of particular interest to us for solar forecasting. Instead we might use only events at 18:00Z (8:00HST).

4.3 Multi-Day Events

Rather than only considering events as a grid of GFS data for a single point in time, we can also consider an event to be a sequence of grids spanning multiple days. In order to reduce



(a) Original (b) Principal Component Analysis (c) Discrete Wavelet Transform

Figure 7: PWAT event at 2014-11-05T12:00Z

the amount of data being processed, we have chosen to use grids at 24 hour intervals rather than the 6 hour intervals provided. For example, a 2-day event might consist of data for 2014-01-01T00:00Z and 2014-01-02T00:00Z. This means that we can have a different set of rare events for each time of day provided.

Since we are now operating on multiple 2-dimensional grids, we perform the DWT on each grid individually but perform the rounding on the entire set of coefficients from multiple grids. We can do this for either 2-dimensional or 1-dimensional DWT (if we first convert the grid into a 1-dimensional vector, as before).

4.4 Results

We found that transforming the data using PCA was not particularly useful as it would not preserve interesting features. Instead it would tend to essentially group events based on the average value across the grid. For example, Figure 7 shows an example event before transformation and after applying either PCA or 2-D DWT and performing rounding. Although Figure 7c clearly does not look anything like an actual weather pattern because of the constant values and sharp changes, it was able to capture the fact there is a region of high PWAT in the southwest. In contrast, Figure 7b looks more like a real weather pattern but does not capture any of the interesting features of the original data. For this reason we have focused mainly on DWT as our transformation method.

Without deeper investigation into the details of each event and whether or not they would be considered "rare" in a meteorological context, the actual rare events identified are generally not of particular interest. One notable exception are events which correspond to hurricanes or tropical storms passing near Hawaii. These are obvious candidates for rare events and our methods were often able to identify them as such.

Since we are not focusing on the specific events, another way to look at the results is to look at patterns within the results and to compare different methods. One thing that we found was that rare events often occurred consecutively or in close succession. For example, the following is a partial list of the time stamps for rare events identified using 2-D DWT on PWAT data, which demonstrates this point: 2014-03-03T00:00Z, 2014-03-03T18:00Z, 2013-

12-23T06:00Z, 2013-12-23T12:00Z, 2013-12-23T18:00Z, 2013-12-24T00:00Z. We also found that there were many similarities between these sequences of events and the events identified by the multi-day approach.

Another interesting result is that there were similarities between rare events identified using different variables with the same method. This could mean either the exact same time stamp was identified as rare in both cases or there was a separation of only a few days between events identified using the two different variables. For example, using 2-D DWT with PWAT identified 2014-02-07T12:00Z as rare and using 2-D DWT with VVEL identified 2014-02-10T06:00Z as rare.

We found that there were not many similarities in the rare events identified by different methods. For example, using either 1-D or 2-D DWT with TMP would not identify many of the same rare events with some exceptions such as hurricanes. Further work is needed to identify which method is best.

5 Conclusion

In this paper we discussed a method for short-term solar irradiance forecasting using linear regression as well as several techniques for improving the linear regression model.

We found that adding predictor variables was useful in reducing errors. In particular including additional spatial data improved predictions in all cases that we tested, while adding temporal data only improved predictions under certain circumstances. We used regions up to 13×13 grid points, but the rate of improvement was greatly reduced beyond a 5×5 region.

Partitioning the data and creating separate models for each partition was also helpful in certain cases. Spatial partitioning turned out not to be very helpful except at high elevations which accounts for only a small portion of Oahu. Temporal partition was more useful with dividing the data into morning and afternoon partitions reducing the errors the most.

All of the above methods were out-performed by first applying a data transformation which takes into account different patterns based on the time of day and location. In addition this can be further improved by including additional spatial data.

We also discussed a method for identifying "rare weather events." Although much work is needed in this area, we have some initial justification for our method as it was able to identify hurricanes as rare events.

6 Future Work

There are many potential directions for future work. First, we can investigate more ways to improve and refine our model. We can also apply our methods to different data sets to see if it can be used in more general cases rather than specifically for Oahu. Lastly we can extend our method for rare event detection and apply potentially apply it to our solar irradiance forecasting.

One way that we might improve our linear regression model is to look into different ways to transform the data. For example, we might want to focus on how the solar irradiance has

changed in the last few hours. Suppose we want to use $S(x, y, t - 60)$ and $S(x, y, t - 90)$ to predict $S(x, y, t)$; rather than using the values directly like we would currently, we can use $S(x, y, t - 60)$ and $(S(x, y, t - 60) - S(x, y, t - 90))$ to help capture how the solar irradiance has changed over time.

We could also incorporate GFS data to improve predictions. One way to do this is to use GFS variables directly as additional predictor variables for our linear regression model. We could also use the GFS data as a criteria for partitioning. For example, we might create partitions based on the wind direction.

In terms of applying our methods to different data sets, one way to do this is to expand to regions other than Oahu. The simplest approach would be to simply create a separate model for each region. However, if a different region is similar, perhaps we could create a single large model consisting of multiple regions. The most obvious next step is to try other islands in Hawaii, but perhaps our methods will work well for other locations as well. It might make sense to create a separate model for each location since the weather and climate will be different. However, we found that our method of transforming the data was able to account for both spatial and temporal patterns, so it may be possible to apply our data transformation to data from all regions and create a single model.

Rather than scaling up, we could also scale down. For example, we can see if we can use similar methods to create a model based on data from a ground based weather station. Alternatively, rather than creating the model on the data from the weather station, we could investigate the performance of our existing models using the weather station data as input. Of course, we would not be able to use any model which incorporates spatial data. This would be useful because there is much more data to train the model if we use the gridded data rather than only the data from a single station. This would also potentially allow us to use a single model to make predictions for different weather stations.

Lastly, our rare event detection methods needs to be studied further. First, we must identify a way to determine which of our methods, if any, are able to accurately identify rare or unusual weather patterns. We can also work towards predicting rare events and using those predictions as part of our solar forecasting model. Rare events could be used as a criteria for partitioning. For example, we could create separate models for days on which a rare event occurred.

References

[1] Maimouna Diagne, Mathieu David, Philippe Lauret, John Boland, and Nicolas Schmutz. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, 27:65–76, 2013.

[2] Richard Perez, Kathy Moore, Marek Kmiecik, Cyril Chain, Pierre Ineichen, Ray George, and Frank Vignola. A new operational satellite-to-irradiance model–description and validation. In *PROCEEDINGS OF THE SOLAR CONFERENCE*, pages 315–322. AMERICAN SOLAR ENERGY SOCIETY; AMERICAN INSTITUTE OF ARCHITECTS, 2002.