Statistical Forecasting and Analysis of Solar Radiation in Arizona

Michael Olheiser, Joe McGuire, Page Ng, Yao Li
Winona State University, Dept. of Mathematics

(Dated: May 3, 2015)

The Arizona Meteorological Network (AZMET) has been taking measurements of various meteorological conditions including, but not limited to wind speed, evapotranspiration, and solar radiation since 1987. It is of particular interest to solar panel investors to know the future solar radiation outlook and exceedance probabilities for a given amount of solar radiation. This knowledge is useful to determine if a solar panel site would be profitable or useful. Our study aims to forecast and analyze solar radiation $[MJ/m^2]$ using OLS regression, ARIMA modeling, and exceedance probabilities. We also assess the variability of solar insolation for each site to determine which sites would have the highest, and most solar energy production.

I. INTRODUCTION

Due to strong increase of solar power generation, the predictions of incoming solar energy are acquiring more importance. Solar radiation varies nonlinearly due to atmospheric events such as cloudy weather, rain, humidity etc. Therefore estimation and prediction of solar radiation is an attractive issue for solar energy investment. Solar Insolation is a measure of solar radiation energy received on a given surface area and recorded during a given time expressed in MJ/m^2. To provide recommendations of locations to build solar power plants, forecasting of the mean daily solar irradiation by OLS Linear Regression with lagged variables and ARIMA were used. Exceedance probabilities and seasonality were also considered to assess variability of the solar insolation for each selected station of AZMET. These methods in conjunction allowed the identification of the top two primary recommendations with high and stable daily solar insolation.

II. DATA

With the intent of providing information for agricultural and horticultural interests, the Arizona Meteorological Network (AZMET) launched a full operation in 1986. January 1st, 1987, 9 automated weather stations started collecting meteorological data and weather-based information. Since the launch of collecting data in January of 1987, 29 more stations started collecting data and 11 of these 38 stations have since been uninstalled.

Provided on the AZMET website (ag.arizona.edu/azmet) are daily and hourly raw data collected by the weather stations by year. The weather stations took hourly readings of meteorological data such as air temp, relative humidity, vapor pressure deficit, total solar radiation, total precipitation, and max wind speed. Hourly readings are summarized in a daily raw file with summary statistics such as the maximum, minimum, mean of soil temperatures, total precipitation, and solar radiation.

As the offline sites are no longer collecting data, the remaining active weather stations were candidates for placing solar power stations. In order to have at least 10 years worth of data to forecast solar radiation and an extra year to check the prediction accuracy, sites active for at least 11 years were examined. Of the currently 29 active weather stations, 15 stations have been collecting data for at least 15 years.

Building predictive models and creating summary statistics first required data management of the daily data for these 15 stations. Separate yearly files had their rows merged by matching columns to create one file of data for each weather station. Data declared bad by AZMET that appears in the original data sets as 999 or 9999 were replaced with blanks. Observations that had the year mislabeled were deleted. Duplicate observations were also deleted. However, if two observations had the same day, but did not have the same entries for all of the columns, both observations were deleted.

III. METHODS

A. Variability

Solar radiation that reaches the surface is highly dependent on the number of clouds in the sky. The cloud fraction will likely vary from season to season, therefore it is important to determine when the seasons are changing in Arizona. We can find the seasons over Arizona by looking at which days of the year tend to me more variable than others. It should be noted that this method would not apply to weather that is variable year round such as the Midwestern United States, thus we are assuming Arizona has somewhat consistent and distinct seasons each year. Considering all daily values of solar insolation for all 38 sites, we can find an average standard deviation for days 1-365 by the following calculation for each day:

$$\text{Avg Std (day)} = \frac{1}{38}\sum_{i=1}^{38} \text{Std}_i(\text{day}) \tag{1}$$

Elevation may play a role in how variable the weather is for different stations due to processes such as orographic lift that creates more clouds on windward sides of mountains and drier weather on leeward sides. (Ahrens, 2013) To investigate this possibility, we used a similar approach to equation (1) by finding the standard deviation of each day, but in this case we consider the stations separately so the average takes place over the 365 day year.

B. Exceedance Probabilities

One measure of variability and risk are exceedance probabilities. An exceedance probability is a measure of the percentage of data points that occur above a reference value. In regards to the daily solar insolation, the calculation can be written as:

$$\text{E.P.} = \frac{\text{no. of days solar insolation} \geq \text{reference}}{\text{total days}} \tag{2}$$

We found exceedance probabilities for 15 sites at 50, 75, 90, and 95 percent and exceedance probabilities for a given reference value of solar insolation in each season.

C. ARIMA

One method of forecasting used here is an Autoregressive Integrated Moving Average (ARIMA). The shorthand notation for a seasonal ARIMA model is written as (p,d,q)(P,D,Q) where the parameters are orders of autoregression, differencing, and moving average respectively while the uppercase refers to the seasonal portion and lowercase as the non-seasonal portion.

Our time series consists of 52-week years, where each week has a daily average reading of solar radiation. In a seasonal oscillation, the first step is to difference the data by 52 week years. An autocorrelation plot would show that the data is now stationary and contains large negative spikes at seasonal lags. This is evidence for a seasonal and non-seasonal moving average model (Duke University). The forecasting equation can be derived for the (0,0,1)(0,1,1)52 ARIMA model as,

$$X_t = X_{t-52} + e_t - \theta e_{t-52} - \theta e_{t-1} - \theta\phi e_{t-53} \tag{3}$$

where e is the stochastic error term, θ is the non seasonal moving average weight, and ϕ the seasonal moving average weight. X_t is the solar radiation at some time t in weeks. The weights are found computationally through a least squares fit.

Using a non-seasonal differencing order as (0,1,1)(0,1,1) is a very common model. In our 52-week years, this model actually over differenced the data and assumes a time varying trend. Solar radiation will not continue to increase or decrease indefinitely, therefore we would assume a flat trend over decades of time, thus providing support for only a seasonal differencing. We also explored the possibility of an autoregression term that may act as a partial differencing since it weights previous values. A seasonal autoregression was insignificant, so a non-seasonal autoregression was tested. After predicting back 1 year of solar radiation, it is apparent that there is not much distinction between the models with and without autoregression. (Figure 3), that is comparing (1,0,1)(0,1,1) to (0,0,1)(0,0,1) respectively. Furthermore, attempting to predict back 5 years caused the model with autoregression to not converge, so the model was not useable. The best fitting model was a moving average model with 52-week seasonal differencing as (0,0,1)(0,1,1)52.

D. Lag Regression

The solar radiation data exhibits stability from year to year (Figure 6). Therefore, the dominance of year long cycle makes it straightforward to build predictive models. The ARIMA models took the approach of 52 week cycles to provide solar radiation specific to seasons. The lag regression procedure is tested on monthly solar radiation data obtained from the 15 sensor stations in Arizona. The following formula was used to calculate the percent error of each prediction:

$$\text{Percent Error} = \frac{\text{Actual - Predicted}}{\text{Predicted}} * 100 \tag{4}$$

A negative percent error indicates the prediction over predicts the mean total solar radiation while a positive positive percent error indicates the prediction under predicts the mean total solar radiation. The actual regression equation and its significance is discussed in the results section for lag regression.

IV. RESULTS

A. Variability and Seasons

In the plot of Average Standard Deviation versus days of the year (Figure 1), notice the spike that occurs around Days 160-250. This peak suggests there is high variability around Arizona at this time of the year which aligns closely with the known monsoon season (Fiero, 2014). This seems reasonable as the monsoon season produces more clouds and precipitation over Arizona that would block solar insolation. Moving to the right of the peak, there is a dip in variability that extends to the beginning of the next year. This low variability corresponds to the winter season where it tends to be drier with less clouds. After the winter seasons, there is an increase in variability that includes spring and summer up to the monsoon. This third regime will be referred to as the warm season. If we find a daily average solar insolation for each week of the year, we can allow for the change in seasons to occur in a more realistic range than an exact day Therefore the monsoon season corresponds to weeks 25-37, winter is 38-52 and 1-8, and the warm season is weeks 9-24. In regards to solar panel investment, it is important to ensure that adequate solar insolation is reaching the surface in each season to meet energy demands that also change seasonally.

B. Variability and Elevation

To investigate the effect of elevation on variability, we plotted the average standard deviation for the 15 stations versus elevation that produced a correlation value of 0.88 (Figure 2). This provides evidence that elevation does have an affect on the variability of the weather. Furthermore, the lower elevation stations have lower variability, thus measuring more consistent solar insolation at the surface.

C. Exceedance Probabilities

The following table ranks the top three and lowest three sites for various excellence probabilities and the corresponding value of solar insolation.

Rank/EP	**50%**	**75%**	**90%**	**95%**
1.	Yuma Valley (21.48)	Bonita (14.71)	Yuma North (12.13)	Yuma North (10.21)
2.	Harquahala (21.37)	Yuma North (14.61)	Yuma Valley (11.86)	Yuma Valley (9.95)
3.	Maricopa (21.29)	Yuma Valley (14.55)	Harquahala (11.63)	Parker (9.55)
13.	Queen Creek (20.64)	Phx. Greenway (13.79)	Phx. Greenway (11.01)	Phx. Greenway (8.41)
14.	Tucson (20.55)	Mohave (13.58)	Phx. Encanto (10.80)	Mohave (8.39)
15.	Safford (20.45)	Phx. Encanto (13.43)	Mohave (10.77)	Phx. Encanto (8.30)

Yuma Valley and Yuma North tend to be the best performing sites in terms of exceedance probabilities, while Phoenix sites tend to do poorly. These probabilities are really going to be concerned mostly with the summer months where solar radiation is reaching its peak, therefore it would be appropriate to determine how each site performs in each season. In the results section we explore the dependence on seasons as well as elevation.

Using equation (2), we can instead fix a reference value for each season and then find the associated exceedance probability. A reference value of solar insolation in MJ/m^2 per day chosen for the monsoon season is 21, winter is 10, and the warm season is 19. The following table ranks the top three sites in exceeding those reference values for each season. Following that table are the correlation values for excellence probability versus elevation.

Monsoon EP of 21	**Warm EP of 19**	**Winter EP of 10**
1. Yuma Valley (94.0%)	1. Yuma North (93.1%)	1. Yuma Valley (94.3%)
2. Maricopa (94.0%)	2. Yuma Valley (92.9%)	2. Bonita (94.2%)
3. Harquahala (94.0%)	3. Paloma (92.6%)	3. Harquahala (94.2%)

correlations	**Monsoon**	**Warm**	**Winter**
Elevation	-0.82	-0.26	0.01

Yuma Valley tends to exceed the reference value the most and is at the lowest elevation. Also notice that the Monsoon season has a strong negative correlation due the higher variability of this season. Winter having the lowest variability, has the lowest correlation here for the exceedance probability.

D. ARIMA

Our ARIMA model was tested by predicting back 1, 5, and 10 years of data (Figure 4). It is apparent that a smaller data set increases the error in its prediction. It is also important to note that the increase in error is dependent on how well the model fits. A site like Tucson actually has decreasing error as more years are predicted back, but it also has the worst fit measured by the R-square value at 0.875. On the other hand, Yuma Valley has the lowest percent error for predicting back one and five years while having a stronger R-square fit of 0.927. The models were then used to predict solar insolation a decade into the future. The daily average over the next decade (2015-2024) is plotted for each station. (Figure 5). Notice the Yuma Valley site is predicted to have the highest daily average solar radiation, which is the site that had the lowest variability, lowest percent error, and lowest variability. It has been found that higher elevation seems to produce more variable weather, but elevation does not correlate strongly with the daily average of solar insolation.

correlations	**Monsoon**	**Warm**	**Winter**	**All Seasons**
Elevation	-0.60	-0.58	0.48	0.29

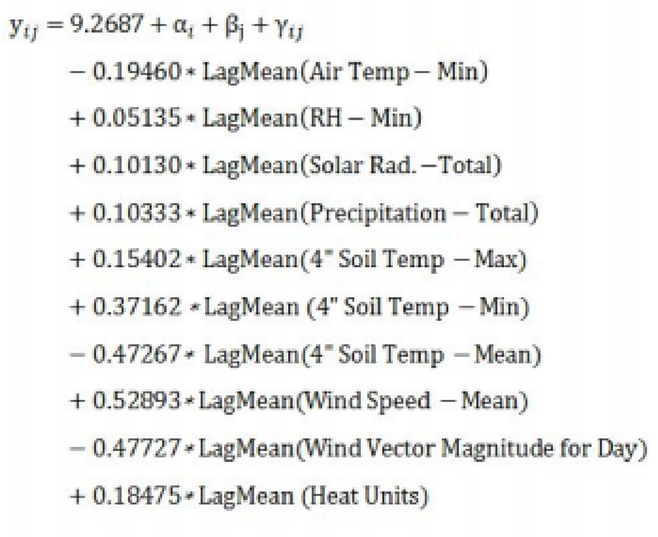
The negative correlation for the monsoon season is expected, as higher elevation produces more variable weather usually by means of more cloudy days that block solar radiation from reaching the surface. In the other seasons there is only a slight positive correlation, so with the 15 sites analyzed, elevation does not necessarily influence the total amount of solar radiation over all seasons.

E. Lag Regression

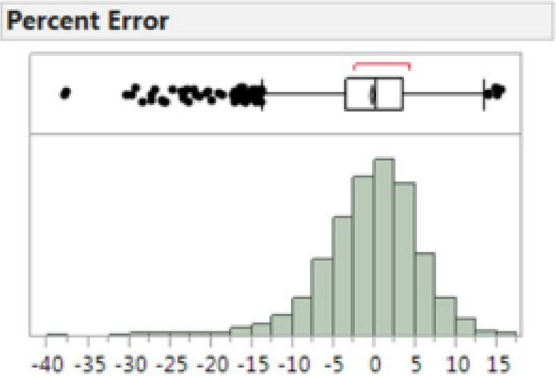
The lag regression model had an R-Square of 0.971801 and the significant variables left by backward elimination with a 0.05 p-value threshold are shown in the table below. Since the Month*Station interaction is significant (p-value < 0.001), there is evidence of a difference in mean total solar radiation between the stations during different months. The percent error ranges from -38.04% to 15.36% while a majority of the percent errors fall between -15% and 15% (as seen in the box-and-whiskers plot below).

Variable	p-value
Month	<0.001
Station	0.0018
Month*Station	<0.001
Lag Mean (Min Air Temp.)	<0.001
Lag Mean (Min RH)	<0.001
Lag Mean (Total Solar Rad.)	<0.001
Lag Mean (Total Precip.)	0.0054
Lag Mean (Max 4" Soil Temp)	0.0495
Lag Mean (Min 4" Soil Temp)	0.0006
Lag Mean (Mean 4" Soil Temp)	0.0090
Lag Mean (Mean Wind Speed)	0.0038
Lag Mean (Wind Vector Magnitude)	0.0197
Lag Mean (Heat Units)	0.0076

The prediction formula for the mean daily irradiation per month from the OLS linear regression model is pictured below. In the prediction formula, i is an index for the month while j is an index for the station where y_{ij} is the predicted mean daily irradiation per constant for the month, α is the constant for the i^{th} month, and β is the constant for the j^{th} station.



$$\begin{aligned}
y_{ij} = 9.2687 &+ \alpha_i + \beta_j + \gamma_{ij} \\
&- 0.19460 * \text{LagMean(Air Temp} - \text{Min)} \\
&+ 0.05135 * \text{LagMean(RH} - \text{Min)} \\
&+ 0.10130 * \text{LagMean(Solar Rad.} - \text{Total)} \\
&+ 0.10333 * \text{LagMean(Precipitation} - \text{Total)} \\
&+ 0.15402 * \text{LagMean(4" Soil Temp} - \text{Max)} \\
&+ 0.37162 * \text{LagMean (4" Soil Temp} - \text{Min)} \\
&- 0.47267 * \text{LagMean(4" Soil Temp} - \text{Mean)} \\
&+ 0.52893 * \text{LagMean(Wind Speed} - \text{Mean)} \\
&- 0.47727 * \text{LagMean(Wind Vector Magnitude for Day)} \\
&+ 0.18475 * \text{LagMean (Heat Units)}
\end{aligned}$$



In the LS Means Plot (Figure 7) the differences of mean daily irradiation per month between the months and stations is seen. The predicted mean daily irradiation per month for station 19 (Paloma) is below the values for the rest of the stations by month. From the months of June to October there is a drop in mean daily irradiation per month compared to the rest of the stations for stations 1, 4, and 9. (Tucson, Safford, and Bonita respectively). The remaining stations? mean daily irradiation per month appear to have a similar monthly trend.

V. CONCLUSIONS

The uncertainty in weather ultimately determines the amount of risk involved for solar plants in Arizona. Elevation seems to create a higher risk environment by way of a more variable monsoon season. Yuma Valley and Yuma North were the two lowest elevation stations and both have the least variablility. The ARIMA model predicted Yuma Valley, to have the highest average solar insolation per day while the OLS regression suggested there to not be a significant difference in solar insolation other than Tucson, Safford, and Bonita that were ranked the lowest. The best stations seem to be Yuma North and Yuma Valley which both lie at low elevation, have more stable solar insolation, higher insolation for most exceedance probabilities, and are not predicted to have low levels of solar insolation by either ARIMA or OLS regression in comparison to other sites. On the other hand, Tucson and Safford seem to be the highest risk by having the lowest rank of 50% exceedance, lowest insolation prediction by OLS regression, and having higher variability.

VI. FURTHER RESEARCH

In our study, we assumed that the location of a solar plant would be located at one of the AZMET sites. Geostatistical methods such as kriging can be used to identify an arbitrary site location in Arizona. This method could be ideal to limit the uncertainty of weather has on solar insolation while gaining the highest amount of solar insolation. Satellite data of cloud fraction could be used to create a very reliable model to forecast solar radiation if cloud fraction largely determines the amount of solar radiation that reaches the surface.

VII. REFERENCES

Ahrens, D. 2013 Meteorology Today: An Introduction to Weather, Climate, and the Environment Tenth Edition. Ch. 6

Arizona State University. The Arizona Meterological Network (April 21, 2015) Retrieved from: http://ag.arizona.edu/azmet/

Duke University. Introduction to Arima Models (April 21, 2015) Retrieved from: http://people.duke.edu/ rnau/411arim.htm

Fiero, Brad. Tucson's Five Seasons (April 21, 2015) Retrieved from: http://wc.pima.edu/Bfiero/tucsonecology/climate/seasons.ht