2013

Predicting solar radiation based on available weather indicators

Frank Joseph Sauer

Department:

PREDICTING SOLAR RADIATION

BASED ON AVAILABLE WEATHER INDICATORS

by

FRANK JOSEPH SAUER

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN SYSTEMS ENGINEERING

2013

Approved by

Ivan Guardiola Ph.D., Advisor

Elizabeth Cudney Ph.D.

Donald Wunsch Ph.D.

ABSTRACT

Solar radiation prediction models are complex and require software that is not available for the household investor. The processing power within a normal desktop or laptop computer is sufficient to calculate similar models. This barrier to entry for the average consumer can be fixed by a model simple enough to be calculated by hand if necessary.

Solar radiation modeling has been historically difficult to predict and accurate models have significant assumptions and restrictions on their use. Previous methods have been limited to linear relationships, location restrictions, or input data limits to one atmospheric condition. This research takes a novel approach by combining two techniques within the computational limits of a household computer; Clustering and Hidden Markov Models (HMMs). Clustering helps limit the large observation space which restricts the use of HMMs. Instead of using continuous data, and requiring significantly increased computations, the cluster can be used as a qualitative descriptor of each observation. HMMs incorporate a level of uncertainty and take into account the indirect relationship between meteorological indicators and solar radiation. This reduces the complexity of the model enough to be simply understood and accessible to the average household investor.

The solar radiation is considered to be an unobservable state that each household will be unable to measure. The high temperature and the sky coverage are already available through the local or preferred source of weather information. By using the next day's prediction for high temperature and sky coverage, the model groups the data and then predicts the most likely range of radiation. This model uses simple techniques and calculations to give a broad estimate for the solar radiation when no other universal model exists for the average household.

ACKNOWLEDGMENTS

Thanks go out to Dr. Guardiola for his guidance throughout my college career. His perspective on analysis and course work enhanced my graduate studies. I would especially like to thank him for his time and patience on weekends and personal time.

I would like to thank Dr. Cudney and Dr. Wunsch for all their efforts in educating me throughout my university career.

I would like to thank Daniel Berc of the National Weather Service for his insight into the data collection and processing system s . His assistance was crucial to understanding the sensors.

Most of all I would like to thank my parents for their understanding and exceptional support during my entire academic career. You always helped me when I needed it and motivated me to continue when times were troubling. I would never have made it without both of you.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Page

LIST OF TABLES

Page

NOMENCLATURE

Symbol	Description
G	Monthly average solar irradiation
G_0	Monthly average extraterrestrial solar irradiation
L	Distance to sea
z	Altitude
ΔT	Monthly average minimum temperature
T_{ref}	Reference temperature
H_0	Extraterrestrial Radiation
H	Terrestrial Radiation
S_0	Extraterrestrial Sunshine Duration
S	Terrestrial Sunshine Duration
Δt	Difference in Maximum and Minimum Daily Temperature
$\bar{t_5}$	Five Day Average Temperature
M	Cluster Medoid List
m_i	Cluster Medoid for group *i*
C_l	List of Points in Cluster *l*
a(i)	Measure of Average Distance to Each Point in the Same Cluster
b(i)	Measure of Average Distance to Each Point in the Nearest Cluster

T	Length of the Observation Sequence
N	Number of States in the Model
M	Number of Observations in the Model
Q	Distinct Set of States of the Markov Process
V	Distinct Set of Possible Observations
O	Observation Sequence
λ	Model Description of the Hidden Markov Model $\lambda=(A, B, \pi)$
A	Matrix of Transition Probabilities
B	Matrix of Emission Probabilities
π	Initial Distribution of States
X	State of the Model (Hidden)
O_t	Observation at Time t
$\alpha_t(i)$	Partial Probability of Observations Before Time *t*
$\beta_t(i)$	Partial Probability of Observations After Time *t*
$\gamma_t(i)$	Most Likely State at Time *t*

1. INTRODUCTION

It is commonly known that there are benefits from renewable energy. The dwindling supply of fossil fuels and other non-renewable sources of power are a large influence on the development of other continuous sources for energy that do not rely on limited supplies of natural resources. These sources of power are also influenced by other factors such as “going green,” the minimization of environmental by-products from historic methods of power generation.

The most ancient source of power for the earth is the sun, and it is intuitively one of the best sources for a renewable source of continuous power. The main source of solar power is the photovoltaic (PV) cell. The PV cells capture radiation from the sun, and convert into Direct Current (DC) that can be stored directly to a battery.

1.1. PHOTOVOLTAIC CELLS

The main two inhibitors to large-scale solar power generation facilities are the inconsistent power generation and transmission of DC electricity.

The reliance on clear skies and consistent atmospheric conditions becomes problematic for a consistent supply of power. A reliable power system requires that there is a regular flow to be stored or directly feed the significant demand for electrical power. Solar power could supply sufficient energy given perfect conditions, but the weather hinders generation without a predictable pattern.

Solar photovoltaic circuits generate DC power. This DC power is used with most circuitry and batteries, but needs to be converted for most household appliances.

Household appliances have been designed to run on Alternating Current (AC) to be directly powered by the electricity supplied to all households. The transmission of DC power has remained a problem since the inception of the electricity supply system or "power grid." AC electricity was chosen instead of DC mainly for its ability to hold charge for longer distances when carried along power lines.

Because large-scale solar power generation facilities are impractical, the focus of many solar equipment suppliers turns to the independent household and self-sufficiency sector. The supply of residence scale and even handheld device-scale solar generation products has proliferated throughout the western world. However, the proliferation of PV cells is more influenced by the social status and how novel the product seems instead of the direct economic benefits.

Currently, the photovoltaic cells rely on social benefits for popularity and sales. Solar cells are not purchased for an economic benefit, but become a luxury good and status symbol. People adopt solar power when nearby people are environmentally minded, especially when their neighbors already use solar power. [Gillingham 2012]

Daily solar radiation is not available or known to the public in a manner comparable to temperature or other weather. Subsequently, the public is unclear on the returns from solar radiation. There is an implied relationship between a sunny day and high radiation, but there is not an understanding about how the radiation is measured or converted into electricity. The lack of predictable returns using solar power generation creates a barrier to new customers. The NOAA agency gathers information with the Surface Radiation (SURFRAD) Network. However, this information is not available for every city nor is the information definitive about solar power potential for the area.

New customers are much more likely to purchase a PV system when the uncertainty about the system is controlled. Some companies, such as Solar City and Sungevity, provide installation and calibration services with their PV systems in order to eliminate error and uncertainty. These companies lease the solar cells to the consumer and offer to pay the negative difference if the system does not operate as promised. [Gillingham 2012]

1.2. STOCHASTIC PROPERTIES OF SOLAR RADIATION

Solar radiation data has been gathered for extended periods of time. This data has been analyzed and shown to retain stochastic, or time based characteristics. These characteristics allow stochastic models to accurately and reliably represent the radiation data. A predictor model for the original stochastic data can be created when an understanding of the descriptors is developed.

Stochastic models of daily "insolation" or radiation data have existed for more than 40 years. A study in Solar Energy, found that when looking at 60 day periods throughout the year, that sequences can reproduce the sequential characteristics of the original data. This modeling technique verifies that period of time, and preceding day's value influences the radiation on the next day. [Brinkworth 1976]

Another Solar Energy study tested different techniques in modeling solar radiation, and found that the simple Markov Chain gave the most accurate representation of the radiation including "noise" or variance. [Mustacchi et al. 1979]

In 1988, a set of Markov Transition Matrices (MTM) was used to continue the Markov Model by creating separate prediction based on the average monthly clearness

index values. Each of the MTMs was used for a short range of clearness indexes. This created a large-scale model to be used throughout the year, but still contained specific representations of individual months. [Agular et al. 1988]

Markov Models will be discussed in depth in the Section 4. Methodology.

1.3. MONTHLY SIGNIFICANCE

As shown by [Brinkworth 1976] and [Agular et al. 1988], unique months during the year have significant impact on the daily radiation values. This is confirmed by [Skiba et al. 1997] using linear correlations of monthly mean daily sums. Their distribution based on the linear correlation equations has a maximum relative deviation less than 8% when compared to the actual values.

Monthly information can be significant to the model, but it was shown by [Olseth et al. 1984] that the time average values are unrelated to simultaneous input for solar driven processes such as PV electricity production. This study showed that using monthly clearness index can allow for unique distributions to accurately describe distributions throughout the year.

1.4. HORIZONTAL VERSUS DIFFUSE RADIATION

Studies have been done attempting to complete the total diffuse radiation based only on one directional component of the data. For this study it is significant that the horizontal data accurately relates to the total potential energy generation. The PV generation most commonly uses cells placed at an angle in order to directly face the sunlight. Using combinations of existing models, [Notton et al. 2005] found accurate

descriptions of the diffuse radiation with less than 11% Root Mean Squared Error as a percentage of the mean. Using a horizontal reading for this research can be an accurate description of the total radiation.

1.5. TEMPERATURE VERSUS RADIATION

Solar radiation affects many different factors on earth, many of which are complex and impossible to find direct relationships, linear or nonlinear. One of the many highly correlated measures, temperature is intuitively raised by the addition of radiated energy into the system. Temperature and light are the two outputs for additional energy in any system. In the past, illumination in the form of sunlight duration has been correlated with solar radiation, but recently temperature has been used as an indicator for solar radiation in many different models. Models with temperature include: temperature as a direct input for correlation analysis [Tiba et al. 2012][Prieto et al. 2009], neural networks [Alam et al. 2009] [Yacef et al. 2012], and as a replacement for sunshine duration in Ångström equations. Each of these models has additional inputs such as location or wind and often requires additional information that can't easily be generated or measured at every unique location (extraterrestrial radiation, solar duration with obstructions). This research uses temperature in a way that simplifies the relationships without loss of fidelity to the interaction of temperature and radiation.

[Tiba et al. 2012] uses temperature along with location, wind-speed and global solar radiation to show significant correlations between the module temperature and the electricity produced. This study proves that overheated PV modules lose efficiency and do not generate additional power over a certain temperature. Wind-speed was included in

the study as a cooling mechanism for the PV cells and not an additional source of power generation.

1.6. SKY COVERAGE VERSUS RADIATION

Intuitively, sky coverage and radiation are connected. When clouds come between the sun and the PV cell, the radiation is diminished and scattered. It would be beneficial to include sky coverage in the model in order to include the variability clouds introduce into the measurements. Previous models, such as [Skiba et al. 1997] and [Prieto et al. 2009] include sky coverage and note its significance to the model.

1.7. PV SYSTEMS SETUP AND ANALYSIS

Renewable energy (RE) systems can be accurately created, modeled, and optimized using computer simulation programs. There is no need for renewable energy installation optimization software as one already exists. One of the most popular applications is HOMER. This program can optimize parameters given the expected generation and load of the location. Parameters can be individual component type or model, and overall configuration of system, DC/AC power supply, power generation sources or storage components and capacity.

Load values can change just as drastically as power generation potentials. Because of this, the leveling of load is essentially impossible and power supply maximums are the goals of the generation system. The improvement of renewable energy systems becomes the supply side of power. Each system needs to be able to meet the load, by generating and storing power until it is needed. In order to create an efficient and

robust system, accurate predictions need to be made on the power generation sub-system. In order for this goal to be accomplished, the inputs (radiation, wind, hydro-electric) need to be modeled correctly. Accurate generation predictions are a requirement for HOMER software to optimize each power generation system.

Overall system configuration can be optimized when cost of resources and components is known. When all generation potentials are known and held constant, the most robust power supply system can be developed. Configuration comparisons are completed with stand-alone (SA) systems, those which are not supplied with electricity power from an existing power grid. This eliminates non-renewable energy supplied as one of the factors, and provides direct comparison between the different RE sources.

By using SA systems in rural locations, it has been found that there are systems which are consistently optimal. [Tzamalis et al. 2011] and [Bernal-Agustin et al. 2009] both found that the hydrogen energy storage systems are currently too expensive for their benefits and have an energy cost approximately 3 times as great. As stated by [Bernal-Agustin et al. 2009], “Energy storage in hydrogen, although technically viable, has a drawback in terms of its low efficiency in the electricity-hydrogen-electricity conversion process, besides the fact that, economically, it cannot compete with battery storage at the present time.” Diesel fuel is a more cost efficient alternative for most electrical generation situations. The environmental impacts of both diesel and batteries can easily be counteracted using the money saved by not using a hydrogen storage system.

When comparing SA energy systems to the grid-connected (GC) equivalent, the configuration can change significantly depending on the region RE potential and local grid-supplied energy cost. Using both wind turbine (WT) and PV power generation,

[Turkay et al. 2011] found that energy costs, including amortized system components, could be as high as $3.39/kWh for SA power systems, while GC systems with equivalent components were conservatively estimated to cost $1.2/kWh or as low as $0.307/kWh.

The GC system provides a lower cost of energy by utilizing a sufficient RE system to provide power supply for the constant load needed for a building while selling excess power to the grid provider and only buying the additional capacity needed at peak load hours. The SA system requires a power storage capacity high enough to obtain complete capacity of the load, but with a GC system can also reduce capacity significantly.

[Liu et al. 2012] completed analysis in Australia to again confirm the benefits to be obtained from GC power systems. Using only a 6kW PV system without WT generation, a household could produce 61% of the total electricity load, reduces electricity payments by 90%, and reduces carbon dioxide emissions by 95%. The study also showed that investment in a PV system has a 12-16.3% return on investment. The cost of energy is also reduced to at the highest energy costs, $0.092/kWh. The costs found by this study are heavily influenced by the tariffs and benefits for using PV systems in the region. This study also found that the slope of the PV modules should be facing a very regular slope between 20^{o} and 25^{o} depending on the city. This displays that the optimum installation is quite predictable when geographic location is considered.

It has been shown repeatedly that a PV system with a battery storage capacity can greatly reduce cost of energy and reduce carbon emissions. This is accomplished by using a GC configuration where generation ability can be optimized at each individual location without excessive energy storage capabilities.

2. LITERATURE REVIEW

2.1. LOCATION AND TEMPERATURE BASED

[Prieto et al. 2009] formulated a functional correlation methodology to use temperature, altitude, and distance from the sea to estimate horizontal radiation at different locations along the northern Spanish coast. The correlation equation was developed but required extraterrestrial radiation values and was reliant on experimental data to determine the function for altitude and distance to the sea. The correlation equation is used as follows from previous experiments:

$$\frac{G}{G_0} = f\left(\frac{z}{L}\right)\left(\frac{\Delta T}{T_{ref}}\right)^{0.5}$$

The model is particularly useful when related locations have recorded data, and that the location function can be used to predict values in new locations that are unknown. This model is unsatisfactory in predicting future radiation, and because of its dependence on the extraterrestrial radiation as an input is not an easily accessible method.

2.2. NEURAL NETWORKS

A Neural Network (NN) is a computing method that attempts to loosely follow the methods employed by the human brain. As described by [Haykin 2011] in Neural Networks and Learning Machines: "*A neural network is a massively parallel distributed processor made up of simple processing units that has a natural propensity for storing experiential knowledge and making it available for use*." It resembles the human brain in two respects:

1. Knowledge is acquired by the network from its environment through a learning process.
2. Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge.

The NN takes as inputs, many different configurations of variables and through a hidden function, correlates the inputs to outputs. The name is derived from the method's emulation of human neural processing and the lattice-like structure of the different input configurations. Figure 2.1 shows a simple description of the NN lattice.

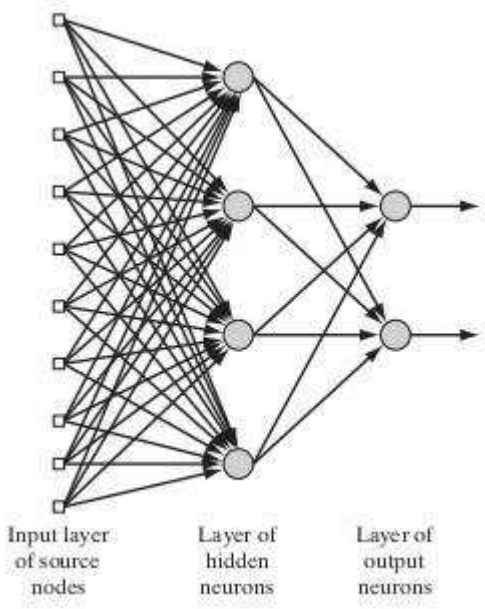

Figure 2.1 Neural Network Diagram

NNs require an immense dataset to train the model compared to the testing set. An often used rule of thumb is that the training set be ten times as large as the testing set. This requirement limits the application of the technique to sites when a large amount of

information is already known. Predictive models are extremely useful when the data is needed for a decision with a degree of certainty. NNs can imply or use an estimation based on the data shown, but they cannot be confident in the answer because of the structure of the model. Hidden nodes with uncertain parameters and reactions are used to foster experiential learning in a NN. This focus toward experiential learning then inputs an amount of uncertainty into the final model.

2.2.1. Assessment of Diffuse Solar Energy Under General Sky Condition. [Alam et al. 2009] This study uses a NN to model hourly clearness, measured as a ratio of terrestrial to extraterrestrial radiation. The study uses a set of sixteen different inputs to attempt to model the same output clearness. All of the measurements were gathered at ten Indian stations: Jodhpur, Kolkata, New Delhi, Pune, Chennai, Port Blair, Ahmedabad, Nagpur, Mumbai and Vishakhapatnam. The data was divided into seasonal sections for each of the models and the average or typical day of each season was modeled separately. The estimated hourly values for clearness are compared to the actual measurements taken and a mean square error (MSE) is taken for each model.

The benefits to this model are shown with the relative accurate predictions with some of the different models in different seasons. The three best performing models all contained air temperature, relative humidity and net long wavelength and returned smaller than 10^{-4} order MSE values. The NN application to daily radiation values out performs the comparable regression models.

The disadvantages are similar to overall NN: large training set and no prediction from the model. Unique to this method, a complication arises; the most influential factor

in the model is the net long wavelength. This data would be difficult to gather, and often is not gathered at every site where radiation values are needed. [Alam et al. 2009]

2.2.2. Artificial Neural Network Analysis of Moroccan Solar Potential. In Morocco, NN model techniques have been used to develop a method of interpolating between known data sites and estimating solar generation potential for all of Morocco. The study completed by [Ouammi et al. 2012], took 12 years of data from 41 sites and used it to generate "heat" maps of the Moroccan solar radiation potential.

This method, if proved accurate, could prove invaluable to anyone who desires estimation data on solar radiation on any of the land with unmeasured radiation currently. This method also works well for small regions where geographic patterns can be assumed essentially constant.

Unique disadvantages to this method are as follows; the inaccuracy of the model, the difficulty of mountains and land formations affecting weather patterns, and the overall location of Morocco in the African continent near the intersection of the Atlantic Ocean and Mediterranean Sea.

2.2.3. Prediction of Daily Global Solar Radiation Data Using Bayesian Neural Network: a Comparative Study. [Yacef et al. 2012] NNs have been shown to predict solar radiation. This study chooses to evaluate the inputs and improvements that might be done with an ordinary NN model. The NN model is improved by a Bayesian inference method; this adds probabilistic interpretation to the weights of the nodes. Input parameters to be evaluated with an automatic relevance determination are air temperature, relative humidity, sunshine duration, and extraterrestrial radiation. The study also determined the effects of the number of hidden nodes in the NN.

It was found that that Bayesian NN improves upon the regular NN. For the training set, Root Mean Square Error (RMSE), Mean Bias Error (MBE), and Mean Absolute Error (MAE) changed from 17.06 to 8.42, 4.70 to 3.07, and 7.10 to 5.91, respectively. The automatic relevance determination method found that the most important inputs in order from most to least important are: sunshine duration, air temperature, relative humidity and extraterrestrial radiation. For the Bayesian NN the optimum number of hidden nodes was two, with decreases in log evidence for any more hidden nodes.

This study is largely a comparison and verification of improvements in the NN technique as applies to solar radiation. The significance to this research is the improvement from a probabilistic technique and the limited number of hidden nodes. This shows that the statistics of the system can improve results and that the underlying model is not significantly complex that it can't be modeled simply. [Yacef et al. 2012]

2.3. ANGSTROM EQUATION SUMMARY

[Ångström 1924] proposed a basic relationship between the extra solar radiation and the actual observed radiation on the surface of the earth. It has since been developed into a more formalized linear expression for the estimation of solar radiation on the earth from extraterrestrial radiation values and sunshine durations, both extraterrestrial and on earth.

In 1940, J. A. Prescott derived Ångström's original postulation into the form most commonly known as today's Ångström correlation. This derivation formalized the

equation into set values. The formulation considered to be the "classical Ångström-Prescott correlation" is as follows:

$$\frac{H_O}{H} = a + b\left(\frac{S_O}{S}\right)$$

Where H_E is extraterrestrial radiation, H_T is terrestrial radiation S_T is terrestrial sunshine duration, and S_E is extraterrestrial sunshine duration. The coefficients *a* and *b* must both be found through experimental data. The coefficient *a* corresponds to the relative diffuse radiation and *b* loosely relates to cloudiness.

This derived version of the Ångström-Prescott correlation equation is commonly used as a benchmark because of its simple formulation and a relative accuracy given from so few inputs. This same simplicity limits the model and the results it gives. It relies on extraterrestrial data, either measured or estimated and assumes it to be true. The correlation is also heavily reliant on experimental data to find the coefficient values. This limits any predictions by requiring the assumption that the data falls within the dataset used to derive the model.

2.3.1. Models for Obtaining Daily Global Solar Radiation from Air Temperature Data.

A study in Romania [Paulescu et al. 2005] replaced the ratio of sunshine duration in the classical Ångström-Prescott correlation with a function of the difference in maximum and minimum daily temperatures and the 5-day average temperature.

$$H = H_0 \cdot f(\Delta t, \bar{t_5})$$

Where H is terrestrial radiation, H_0 is extraterrestrial radiation, Δt is the difference in max/min daily temperatures and $\bar{t_5}$ is the 5-day average temperature.

The function contains the coefficients to be determined from experimental data and is shown:

$$f(\Delta t, \bar{t_5}) = \sum_{i=1}^{n} c_i \cdot (\Delta t)^{p_i} (\bar{t_5})^{q_i}$$

The coefficients c_i p_i and q_i all are determined from experimental data as i corresponds to the location of the dataset. There were 6 testing datasets from 6 locations; in this study *n* equals 6.

This new temperature based correlation had RMSE of less than .15, but the datasets were limited to days with clear, cloudless conditions. This method has been shown as feasible, but with significant restrictions to the application. However, the strong results from using short term temperature as the only input are impressive.

2.3.2. Simple Nonlinear Solar Radiation Estimation Model. [Şen. 2007]

Turkish researcher Zekai Şen, extended the classical Ångström-Prescott equation with a non-linearity coefficient, making the model:

$$\frac{H}{H_0} = a + b\left(\frac{S}{S_0}\right)^c$$

The variables represent the same as the classic correlation, with the addition of c as an effective measure of system dynamics.

This study does show improvements on classic Ångström-Prescott estimations, but the additional benefits are minimal. The average improvement from the nonlinear correlation is only 6.32%. None of the location datasets were estimated to have a nonlinear coefficient, c, greater than 1.9. So while there is evidence that some locations have significantly dynamic radiation potential, not every location has a nonlinear model. By adding nonlinearity to the estimation, the complexity of the equation greatly increases

and the “strength” of the model declines. Strength of the model is used to gauge its robust nature and applicability for all different data. This nonlinear adaptation is not as strong as the classic formulation. [Şen. 2007]

2.3.3. A New Formulation for Solar Radiation and Sunshine Duration Estimation. [Şahin. 2007] This method proposed doesn't actually change the classic Ångström-Prescott correlation equation, but uses a novel restructuring to create estimations and prediction values. When rewritten, the classic Ångström-Prescott correlation can be shown as:

$$\frac{S_0}{H_0}(1 - R_e) = \frac{S}{H}$$

This description uses R_e to represent the reduction in extraterrestrial radiation before it is measured on Earth. R_e then can be rewritten as:

$$R_e = \frac{\left((S_0/H_0) - (S/H)\right)}{(S_0/H_0)}$$

Given that this value can be found from a testing set, the estimation for both terrestrial sunshine duration and radiation become simple equations:

$$S = HS_0\frac{(1 - R_e)}{H_0}$$

and

$$H = \frac{SH_0}{\left(S_0(1 - R_e)\right)}$$

These equations are beneficial due to their simplicity and that the value can be used to estimate an average or daily value by only knowing the overall ratio for the location.

This method retains simplicity and improves on the implementation of the classical Ångström-Prescott correlation. This variation is superior because it provides a reduction in parameters without an increase in inputs, and it doesn't require coefficients for the equation to be estimated with a least-squares method. However, the method is still restrained to need extraterrestrial values and must have all knowledge specific to the location. [Şahin. 2007]

2.4. SOLAR RADIATION ESTIMATED BY MONTHLY PRINCIPLE COMPONENT ANALYSIS

Using the inputs from the Ångström-Prescott correlation, a study completed by [Şen et al. 2008] replaced the linear estimation equation with a Principle Component Analysis (PCA). PCA is a technique used to reduce parameters of the data by twisting the axes in the direction of the most variation. This technique gives a good estimate about the distribution of the data. When PCA is used to determine the distribution of sunshine duration as compares to solar radiation, the results are equivalent to the classical Ångström-Prescott correlation.

Despite very similar estimations and accuracy, Ångström-Prescott is restricted in many ways that PCA is not.

- Nonlinearity- PCA can describe data which does not fit on a straight linear correlation.
- Normality- It is an assumption of regression models, that the data is normally distributed. PCA does not require normality.

- Conditional Distributions- It is assumed that the distribution at each value of sunshine duration in the Ångström-Prescott will be equally distributed on each side of the estimation.
- Homoscedasticity- It is assumed that all the conditional distributions have equal variances
- Independence- It is assumed that all the variables in the regression models are independently distributed.

By removing these restrictions from the estimation and retaining accuracy, PCA is an improvement over classical Ångström-Prescott regression estimation. [Şen et al. 2008]

2.5. SUMMARY OF METHODS

Most of the estimation models to estimate solar radiation have recognized the non-linear relationship, and the restriction of the models that were originally used. Direct correlation has proved to limit success, and special difficulty with highly variable weather in unique locations.

Modern techniques have embraced the nonlinearity, the immeasurable interactions of the environment and the necessity for a model that can be easily customized for unique locations without a large number of parameters.

However, none of the models use stochastic properties inherent in solar radiation to accurately make predictions into the future. This void can be filled with a simple model, based on stochastic assumptions, which utilizes few parameters to predict the day's radiation.

3. BACKGROUND/PROBLEM DESCRIPTION

This thesis seeks to connect the known weather data with the unknown radiation data in a simple manner. Weather data and radiation data are measured separately, and most locations do not even have radiation measurements. Using Hidden Markov Models, it is possible to estimate the unknown radiation values based on the available weather data. Weather data is widely available at any location. This estimation will be based on a model simple enough to be recreated and used at any household. Clustering will be used to simplify the data from a series of multiple measurements into a single list of input values.

As shown in the Literature Review, modeling techniques are unable to provide a simple and universal model to apply at any location. Most all models rely on the radiation to be measured, and a significant portion of them also require extraterrestrial measurements as well. All of the reviewed methods are reliant on precise quantitative data, and are unable to accept discrete or qualitative measures. These restrictions limit prediction models to those organizations with the ability for high-end processing and modeling capabilities. By creating a method capable of accepting qualitative indicators, predictions become more accessible.

Without relying on measurable radiation data for a prediction model, a measure of uncertainty and hidden behavior must be taken into account in a predictive model. The models which have hidden processes rely on the ability to make complex calculations and simulate a multitude of variables simultaneously. Hidden Markov Models develop this uncertainty with analysis at a low-level complexity while still maintaining accuracy.

Clustering techniques are a mathematically justified method to find commonality amongst large data sets and to simplify information into fewer dimensions. Logically changing a large, multiple dimension dataset and simplifying the indicators by joining those with common qualities could retain much of the information while creating a succinct summary of the information. This in turn can improve simple models for solar radiation prediction by limiting the loss of information. However, there are currently no models available which use clustering in any relation to solar radiation prediction..

4. METHODOLOGY

Previous methods to predict solar radiation allow us to understand what is significant in solar radiation modeling. Hidden Markov Models (HMM) are appropriate because of the inherent probability, time sequence information and the use of hidden states/interactions. This research seeks to create an HMM to describe the solar radiation patterns and their emission of temperature data. In order for a new HMM to be an improvement, it must use a small observation set and a small state set. These small sets are found by the creation of clusters within the data. These clusters have similar properties and simplify the data without loss of information. To use data for clustering, the data must be segmented, uniformly spaced, and exist in the same time-ordered as it was recorded.

This section will first show an example of a potential HMM model of solar radiation with known parameters. The total development process will be outlined. Then the rest of the section will be organized in the chronological order of the processes which start with raw information from the source, prepare the data, cluster the data, then use cluster data to create the HMM models, then HMM models will be trained and models will make predictions for radiation.

4.1. PROCESS OUTLINE

1) Data Processing [see Methodology section 4.2]
    a) Inputs: time coded measurements
    b) Outputs: daily summary vectors
2) Clustering [see Methodology section 4.3]
    a) Inputs: daily summary vectors
    b) Outputs:
        i) list of daily group (qualitative)
        ii) daily temperature estimation (Low, Medium, High)
3) Data Segmentation [see Methodology section 4.4]
    a) Inputs: list of daily group
    b) Outputs: sections of days with the same group
4) Prediction by Hidden Markov Modeling [see Methodology section 4.5]
    a) Inputs:
        i) section of days with the same group
        ii) list of daily temperature estimation
    b) Outputs: prediction of daily radiation level (Good, Bad)

This thesis connects the available weather data with the unknown radiation data. This problem has been decomposed into four main sub-problems: Data Processing, Clustering, Segmenting the data, and Predicting the unknown radiation by HMM. Data processing takes the data from raw inputs and creates a daily vector. Clustering simplifies that daily vector into a single variable, the day's cluster, which qualitatively describes the day. The Hidden Markov Model creates a predicted amount of radiation based on the daily input cluster.

Data, on radiation and weather respectively, is gathered from two separate datasets is combined and aligned into daily values. There is a limitation of which

indicators are available, our choices for the weather indicators are temperature and sky clarity. All other data is problematic because of corruption, static input, or inaccurate measurements. Data processing outputs a daily vector for each of the days in the year consisting of aggregate radiation (W/m2 per three minutes), high temperature (degrees Celcius), and most frequent sky clarity (Clear, Scattered, Broken, Overcast). For a more detailed description of data see subsection 4.2 of the Methodology.

Clustering takes the daily values and groups them into similar clusters. A cluster contains similar days and provides a single variable description for the type of day. This research uses K- Medoids clustering; also known as Partitioning Around Medoids (PAM) clustering. Clustering is completed with three input lists: radiation alone, temperature alone, and temperature with sky coverage. The radiation clusters are used to justify the number of states. The only-temperature clusters are a simple description of the day and are input into the HMM as an observation list. The sky coverage and temperature clusters are used to segment the year into ranges of similar days.

Clustering radiation provides us with two separate radiation clusters. Clustering temperature provides us with two separate temperature levels. Clustering sky coverage and temperature provides us with three separate clusters; {{Cloudy, Any Temp},{No Clouds, Low Temp},{No Clouds, High Temp}}. The extra cluster when using sky coverage with temperature reveals that sky coverage displays additional information about the data. However, when temperature is clustered into three groups, the results are

very similar. The three temperature clusters can be used to describe the data without the added complexity of using sky coverage. By using a single variable, we reduce the complexity of the input, but we don't have significant loss of information. For more details on the clustering process see Section 4.3 of the Methodology.

Data segmentation uses the simplified description of the days provided by clustering the sky coverage and temperature. This description is used to section the data into ranges of days with the same cluster. The daily values are displayed chronologically. Figure 5.4 is visually inspected, and then justified by statistical testing. The following are the ranges of similar data found: {January, February}, {March, April}, {May, June, July}, {August, September}, and {October, November, December}. The visual depiction of the data is shown in Figure 5.4.

Hidden Markov Models (HMM) can predict the unknown state of a system without direct measurement. The HMM requires an estimated number of states in the system. The two radiation clusters directly translate to two separate states to describe radiation. The three temperature clusters are used as a qualitative observation of each day. Each range of days, as found in data segmentation, is modeled separately. Each model predicts the {Bad, Good} radiation level relative to other days within the model's range. This means that the Good days in Model #1 are potentially similar to the Bad days in Model #3.

4.2. DATA SELECTION AND PROCESSING

4.2.1. Data Choice. Data needed for the HMM should be of a finite set of daily values. These values should include, an accumulated solar radiation value, a high

temperature value, and an average sky coverage (cloudiness) value. This Section describes the process from raw data, through source selection, year selection and through processing to prepare the data for clustering.

The following Section will describe the method of data selection for this research; covering the selection of the data source, the requirements for the databases, and the selection of the appropriate years.

4.2.2. Data Source. Datasets were obtained from two separate databases. They are monitored and distributed by the National Oceanic and Atmospheric Association (NOAA) and its subsidiary organization the National Climatic Data Center (NCDC), respectively. The measurements of radiation, temperature and sky coverage for this research were not found in one dataset, therefore it was required to obtain separate datasets and combine them. Both databases contained different measurements; the NOAA set contained radiation and temperature information, while the NCDC set contained sky coverage information. In order for the two datasets to be compiled and used in conjunction, the data and measurements needed to come from the same geographic location. The NOAA dataset and the NCDC dataset coincided at the same geographic coordinates. The data measured at the "Mercury" station on the airfield in Desert Rock, Nevada sent data to both databases. The two datasets (NOAA and NCDC) from Desert Rock was selected for this reason.

4.2.2.1 Database requirements. The model also requires that the data is input as daily measurements. For the NOAA dataset, measurements were taken at 3 minute intervals. For the NCDC dataset, the measurements were taken sporadically most commonly once an hour. Despite the difference in data collection technique, each dataset

was complete enough for conversion to a compatible daily value. Daily values provide summary insight into the data collected without losing much information. For the data collected, the additional challenge is to provide measures that can easily cope with the difference in collection intervals and make a singular value for each day.

4.2.2.2 Data collection. The data was collected by the SURFRAD researchers. All data was collected with sensors in the same proximity and relative sensor locations can be shown in Appendix A

The radiation data was measured by a standardized platform of sensors including: Multi-Filter Rotating Shadowband Radiometer, UVB-1 Ultraviolet Pyranometer, LI-COR Quantum Sensor, ventilated Eppley pyrgeometer, and ventilated Spectrosun pyranometer. These sensors are used to gather direct, diffuse and global radiation data. For this experiment we are inspecting direct downwelling global radiation.

Temperature data was measured by a Vaisala air temperature and relative humidity probe at a height of ten meters.

Sky coverage data was measured by a ceilometer. This instrument measures clouds to 12,000 feet directly above the sensor. The measurement is returned in oktas which are also used in meteorology. This sensor is automated and algorithmic, making it sensitive to low clouds directly above the sensor and inaccurate after the range of its detection.

The radiation, temperature and sky coverage data are measured in Watt meter squared, degrees Celsius, and oktas, respectively. Sky coverage, although algorithmically derived is a qualitative measurement relating to a range for each returned measurement. Sky coverage of CLR (clear) was shown at 0 oktas, SCT (scattered clouds) was shown

from 1-5 oktas, BKN (broken clouds) was shown from 6-9 oktas, and OVC (overcast) was shown at 10 oktas.

4.2.2.3 Year selection. In order to select the year for training and testing, several factors were taken into account. The completeness of the information, the quality of the data and the relevancy of the information obtained all influenced the year range to be modeled. For example, measurements included in the NOAA database contained values only for the first half of the day. This would remove data for lack of completeness. The NOAA database data for some years also contained coding within the extracted text files. It is assumed that there were technical difficulties influencing the data collection and the system logged errors instead of correct data. The NCDC data pertaining to sky coverage was often returned as a null value and is not relevant or useful to our model. The system is automated and it is assumed that the algorithm was inconclusive at these points in the data.

For this research, data was selected from 2004 for model training and from 2005 for testing of the prediction model. The years were chosen to be sequential to simulate the application of such a model using the previous year to predict the values of the current year. The data from 2004 was complete and relevant with only two days excluded; one because of corruption and another because the data was incomplete. The data from 2005 was much more problematic, but enough data remained to test the model based on 2004. In 2005 there were nine days missing entirely and of the remaining, 25 days were missing sky coverage data; 225 out of the 331 available were cloudless.

4.2.3. Data Formatting and Preprocessing. From both databases the measurements were formatted as time-stamped rows in a large matrix. In order to create a daily string of values from each of the datasets with different intervals, different measures were taken for each of the values needed in the model. The measures chosen for radiation, temperature and sky coverage were, respectively: aggregate, high and mode.

4.2.3.1 Daily indicators. Radiation and temperature were originally measured in three minute intervals. In order to convert these measures to daily measures, the aggregate of all the 3 minute measures was taken for radiation, and the highest value was taken from all the measured temperatures. For sky coverage, the data was measured sporadically. This presented a challenge to convert into daily values.

Sky coverage data is automatically generated, but the values were not measured on regular intervals. In order to regulate the number of measurements in a day, the coverage data was limited to hourly values, and then the mode was taken. This provided the daily value of the sky coverage from the inconsistent data set.

When using the developed daily values for sky coverage, substitutions must be made for the qualitative values in order to allow the clustering algorithm to adjust. Fortunately, the qualitative values are determined using a numeric algorithm. The qualitative data corresponds directly to a number value of average coverage during the measurement period. For Example, a value of SCT (scattered clouds) contains the range of 1-5 oktas. When converting back to the quantitative value, the estimated value would be 3 oktas for all SCT values.

4.2.3.2 Normalizing values. In order to create even weighting between temperature and sky coverage, each range of data must be normalized to a value between one and zero in order to use clustering based on the Euclidean distance. This is done by dividing all recorded daily values by the maximum observed data in the year. For the data from 2005, the maximum in 2004 will be used.

4.2.4. Summary. Data was chosen from the NOAA databases for downwelling global radiation, air temperature and sky coverage (cloudiness). These measures were converted to daily measures of accumulated radiation, maximum air temperature, and most frequent sky coverage.

4.3. CLUSTERING

For this research, hard partitional clustering will be used. Clustering procedures all follow the same general outline as described in [Xu, Wunsch 2009]:

1. Data is sampled
2. Features of the data are selected
3. The clustering algorithm is selected, then used on the data
4. The clusters are validated and the results are interpreted.

4.3.1. Feature Selection Using information researched in the Introduction Section Literature review, the features for this clustering algorithm are accumulated radiation, maximum air temperature and most frequent sky coverage. These features have previously been correlated with many other studies, and are sufficient in the datasets chosen.

Data from all sources is prone to corruption and inconsistency due to the automated measurement facilities. A large portion of the data is corrupted or non-existent leaving limited options to use as features.

4.3.2. Clustering Algorithm Selection and Application Clustering will be completed with the k-medoids algorithm. This algorithm successfully groups data with similar information while being resistant to the effects of outliers. This method adjusts each cluster after the addition of a new point, this provides robust measure of membership when the order of data points is fixed.

Application of the k-medoids clustering algorithm is described as follows:

1. Randomly choose k points in the data to serve as medoids or "centers" for the clusters. Remove these points from the unassigned data and store in the cluster medoid list

$$M = [m_{1,} m_{2,} \ldots m_k]$$

2. Assign each data point to a cluster based on the nearest medoid.

$$x_j \in C_l, if \left\| x_j - m_l \right\| < \left\| x_j - m_i \right\|$$
$$for j = 1,2..n \quad \text{and} \quad for i \neq l \; i,l = 1,2..k$$

3. Recalculate the cluster medoid list

$$m_i = min \sum_{(x_m \in C_i)} x_j - x_m$$
$$for i = 1,2, ...k \text{ and } m \neq j \quad m \forall x_m \in C_i$$

4. Repeat Steps 2 and 3 until there is no change for the clusters

4.3.3. Cluster Validation In order to optimize the number of clusters, k, the algorithm was repeated for k= 1, 2 … 6, and the optimum number of clusters was found. The optimum number of clusters is determined by silhouetting. The Silhouetting method is a comparison between a cluster and the next nearest cluster. With it, the validity of the clusters returned can be compared to other methods. The formula for the "Silhouette Weight" described in the method is:

$$s(i) = \frac{b(i) - a(i)}{max \;\; a(i), b(i)}$$

Or alternatively:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} if a(i) < b(i) \\ 0, if a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 if a(i) > b(i) \end{cases}$$

Where a(i) is a measure of average distance to each point in the same cluster, and b(i) is the average distance to each point in the nearest cluster. The average Silhouette Weight is used as a validity matrix for the entire group of clusters.

The optimum number of clusters is found for radiation, temperature and a two dimensional temperature and sky coverage dataset. After the optimum number of clusters is found the results are compared. This method is referred to as relative clustering criteria. The clusters allow us to differentiate between month segments of the daily data. This

allows us to cluster the total year in similar segments by inspecting the distribution of temperature and sky coverage clusters within the ranges for the radiation clusters.

4.4. DATA SEGMENTATION

Data is segmented by plotting the clustered value of radiation {Bad, Good}, and the cluster describing the sky coverage and temperature {{Cloudy, Any Temp},{Clear, Low Temp}, {Clear, High Temp}}. The inspection first separates sections with a fixed radiation cluster. Within each range of static radiation, ranges of constant temperature are found. This gives us segments among the similar radiation levels. Then, segments of like radiation are compared to find similarities of weather cluster. For Example: In the summer, radiation is continuously {Good}. Inspecting summer as one segment, there are obvious ranges of cluster {Clear, Low Temp} and other ranges of cluster {Clear, High Temp}.

In order to retain model simplicity, the ranges were limited to monthly sections. This means that a new range can only start on the first of the month and not on a day in the middle of the month. By restricting the model in this fashion, we keep the time to change models predictable and easy to follow for the common household.

The segments are statistically tested for difference of means for both temperature and radiation. The ranges of data with a statistically significant difference in either temperature or radiation are then separated into unique segments. The similar ranges are combined into one segment. The daily temperature clusters are then separated into corresponding segments and used to train individual HMMs for use in predicting the next year’s radiation for the respective time segment.

4.5. HIDDEN MARKOV MODELING

Similar ranges of months have been segmented from relative cluster validation in the training data. For each of the segments an HMM is created and trained to estimate the unknown radiation as {Good} or {Bad}. The states are immeasurable to the models. The number of radiation states can be justified using clustering on the known radiation from the training dataset. The number of radiation clusters is directly related to the number of unknown radiation states.

Each model is created based on the total observation space. The observation space contains the different types of days which we can observe. These observations contain the number of sky coverage and temperature clusters as found previously in the clustering procedure: Methodology Section 4.2. The number of clusters may change for other locations, subsequently changing the observation space. For this data and location, the cluster from temperature alone provides equivalent estimation to the clusters found from using both sky coverage and temperature.

HMM models are probabilistic representations of the actual environment. They require the parameters to be initially randomized before training. Each HMM is initialized with random, but equivalent, probabilities for the transitions between states, the emission of observations from each state, and the initial distribution of states. Figure 4.1 shows a graphical representation of an HMM.

Training the HMM is the process of improving the accuracy of the model by changing the initial parameters. Since the initial parameters are approximately equally likely, the model will not be representative of the actual behavior. Training the model finds the probability of the observation sequence in the training set, the most likely state

at each time, and the probability for each state at every different time. Using the entire observation sequence, the parameters can be re-estimated. The probability of the observation sequence measures the accuracy of the parameters. The model is iteratively retrained until the probability of the known observation sequence is no longer improved. Daily temperature clusters are used for observations, and the parameters are found for each subset of data respectively. For an example HMM see Appendix B

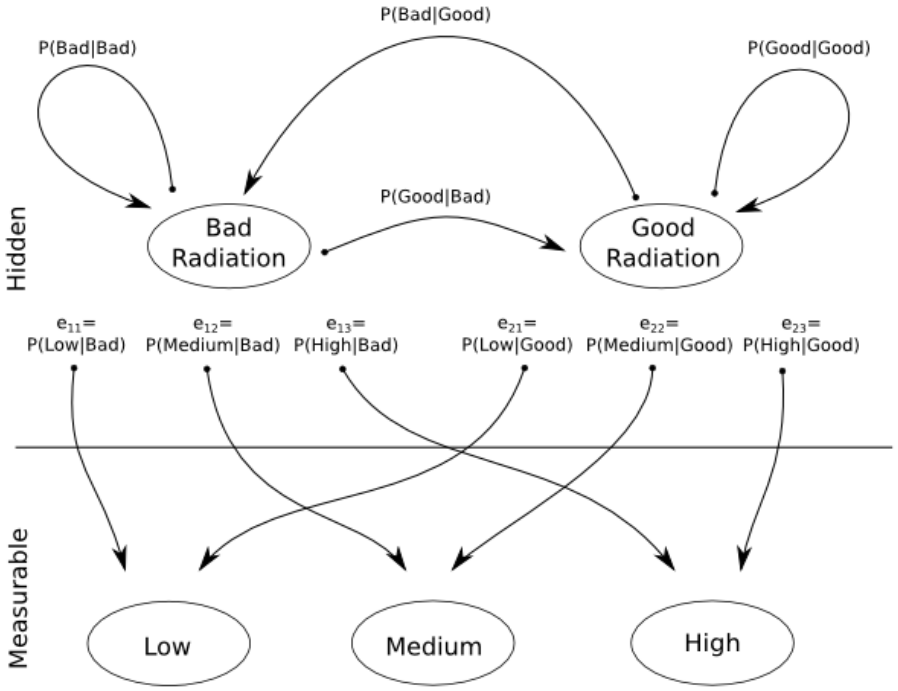

Figure 4.1 Hidden Markov Model Two States with Measureable Temperature

To predict the radiation state during the testing set in the year 2005, the trained model from the respective time during 2004 is used to find the most likely state of the system at the time. These models have parameters estimated by the observations in 2004. The models have not used the measured radiation values from 2004. The state estimated by the HMM is completely based on the temperature-cluster observation. When predicting, a smaller sequence of observations is used to predict the state at the next day after the final observation. This thesis uses testing observation sets of three to predict the future state.

5. RESULTS

This Results Section is split into four subsections, each describing the information found by their respective processes. The Data processing Section describes the raw data, and the subjective nature of the location. The Clustering Section shows the description, separations and simplification of the data. The Data Segmentation Section interprets the clustering results and finds groups of similar data. The HMM Section shows the models as trained by the data, the predictions of those models, and the accuracy of those predictions.

5.1. DATA PROCESSING

Data from Desert Rock Nevada was processed for two years: 2004 and 2005. These years were chosen because 2004 was the most complete year available. In order to retain simple models, the following year was selected to test the models created based on the 2004 data. The following will describe some attributes of the data after it was converted to daily values. Descriptive measures of the data can be found in Table 5.1.

This data was selected because it was the only location for both the radiation and the sky coverage to be measured. All other recording stations had miles of distance between the instruments to measure radiation and sky coverage. The location is in an arid country, and has a less diverse weather pattern than other locations with data for radiation or sky coverage respectively.

Table 5.1: Data Description

		2004	2005
Radiation W/m^2 (3 min)	Minimum	5279.1	9015.5
	Maximum	186053	187200
	Mean	115279	111496
Temperature (^{o}C)	Minimum	5.3	6.4
	Maximum	40.6	43
	Mean	24.49	24.44
Sky Coverage	Overcast	32	12
	Broken Clouds	4	91
	Scattered Clouds	19	3
	Clear, No Clouds	310	225
	*** (Corrupted)	1	25

The fact remains, that 310 of the 366 day in the training set were CLR (no clouds). This lack of variation clustered most of the model by temperature regardless of sky coverage.

5.2. CLUSTERING

Clustering with the PAM algorithm was completed with 3 sets of data, single variable: radiation, single variable: temperature and a two dimensional: sky coverage and temperature data set. The clustering results can be shown in the following graphs. Figure 5.1, Figure 5.2, Figure 5.3, Figure 5.4 all display different combinations of clustering compared with relative criteria to actual and other clusters.

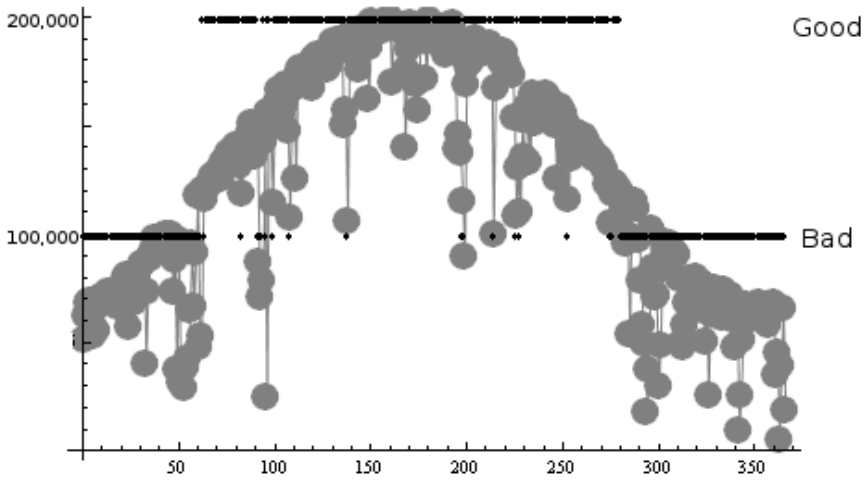

Figure 5.1 Radiation Clusters versus Radiation Actuals

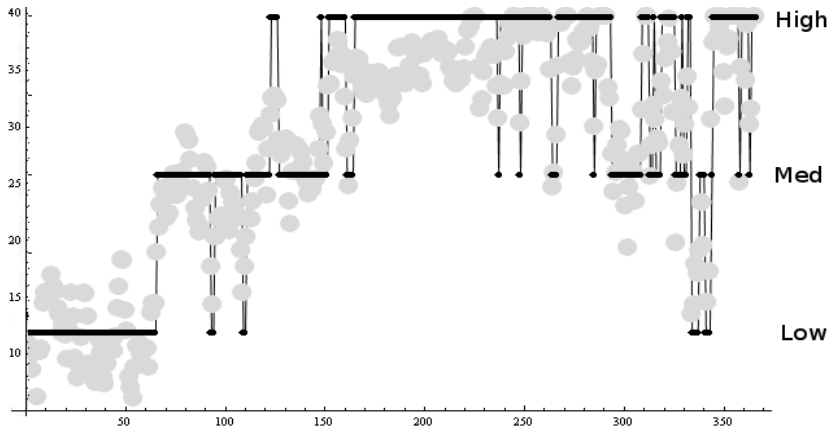

Figure 5.2 Temperature Clusters versus Temp. Actuals

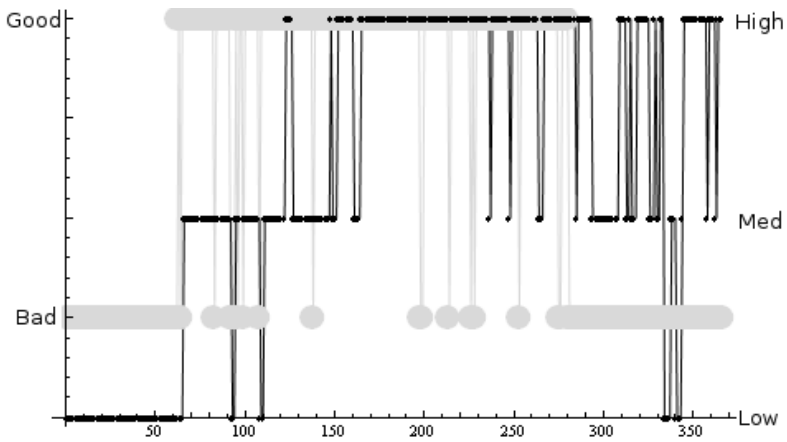

Figure 5.3 Temp. Clusters versus Radiation Clusters

These graphs show the clusters developed in the time sequence of the original data set. This allows for the description of the data to retain the sequential properties necessary to the HMM.

Using these visual representations of the data, the clusters were compared using relative criteria to validate the clusters found. In this case, we are looking for both correlations between the dimensions of the data and for segments of time containing similar data. With these two goals in mind, the data appears to be segmented into 5 distinct and unique ranges: Low temperature and low radiation, middle temperature and middle radiation, high temperature and high radiation, high temperature and middle radiation, and variable temperature and low radiation.

This inspection confirms the inherent logic that the seasons are highly correlated with radiation behavior, but additionally provides confirmation of when the model should be changed and also the difference between the end-of-year and the beginning-of-year during the expected winter season.

It was found that radiation could be clustered into two levels of output, temperature could be clustered into two levels, and the temperature/coverage data set could also be clustered into three groups.

The addition of sky coverage into the clustering algorithm did not significantly influence the cluster selection. Therefore it will not be used as an observation vector. The influence of sky coverage was not significant for this data set, but should not be discounted for other geographic locations. The minimal impact can be attributed to the fact that 310 of the 366 day in the training set were CLR (no clouds). This lack of variation clustered most of the model by temperature regardless of sky coverage.

This research aims to provide a simplified method for an average household. Restricting the time of model change to coincide with a month change allows for a simpler and more accessible model. Model changes can be approximated by the nearest month change to the shown cluster changes. These segments were confirmed by grouping the data and comparing monthly radiation values to show the statistical difference between the groups.

5.3. DATA SEGMENTATION

Figure 5.4 shows the comparisons of the differently segmented days. These days are limited to change when the month changes. This provides an additional degree of simplicity, and therefore accessibility, to the model.

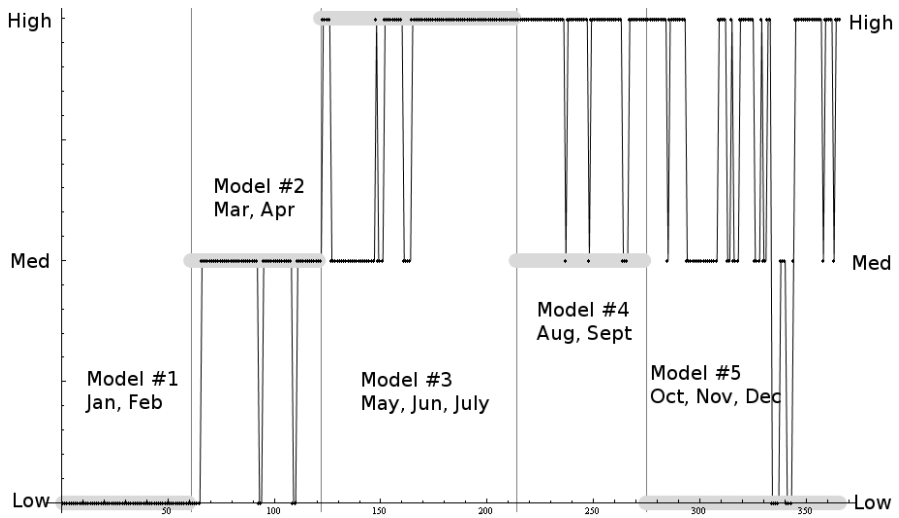

Figure 5.4 Monthly radiation clusters (Grey) versus temperature clusters (Black)

Using the clusters found from sky coverage and temperature, the segments to be separately modeled are {January, February}, {March, April}, {May, June, July}, {August, September}, and {October, November, December}. This shows the need for five separate HMM models to accurately describe each segment of the year.

5.4. HMM MODELING

The models were found from the training method described in the Methodology Section 4.5. These trained models require some interpretation before they can be used for prediction of radiation. Each segment of similar time has a unique model to predict radiation.

The training method does not imply intensity of the states it models, because it doesn't actually predict a value, but a more qualitative description of the hidden process

by a “state.” This means that in order for the original state estimations to apply toward a directional measure, information is needed to imply the order of the states as “bad” or “good” radiation states.

For each of the unknown radiation state, there are known ranges of temperature associated. Using the proven correlation between temperature and radiation, the state which is more likely to emit high temperature can be implied to have a higher radiation. This is a large assumption in the model as the qualitative descriptor is used to imply of an intensity of radiation. Original and adjusted model parameters can be found in Appendix C.

5.5. HMM PREDICTION

Looking at the confusion matrices, it is apparent that there are some models which are better predictors than others. Model 1 has issues because there is only one observation ever emitted in the data set (Low Temp). Model 3 has issues because it is homogeneous with all High Temp, High Radiation observations and states, respectively. The models with significant fluctuation greatly improve when using HMMs. The confusion matrices for the HMM prediction can be found in Table 5.2.

Model 1 corresponds to {January, February}, Model 2 corresponds to {March, April}, Model 3 corresponds to {May, June, July}, Model 4 corresponds to {August, September}, and Model 5 corresponds to {October, November, December}.

Table 5.2 Confusion Matrices found from HMM Prediction

All Year	Actual	
Predicted	Bad	Good
Bad	104	42
Good	72	122
Accuracy	0.664705882	
Adjusted	0.792982456	
Model 1	Actual	
Predicted	Bad	Good
Bad	29	23
Good	0	0
Accuracy	0.557692308	
Model 2	Actual	
Predicted	Bad	Good
Bad	14	10
Good	16	19
Accuracy	0.559322034	
Model 3	Actual	
Predicted	Bad	Good
Bad	0	0
Good	32	55
Accuracy	0.632183908	
Model 4	Actual	
Predicted	Bad	Good
Bad	17	1
Good	7	26
Accuracy	0.843137255	
Model 5	Actual	
Predicted	Bad	Good
Bad	44	8
Good	17	22
Accuracy	0.725274725	

The first confusion matrix in Table 5.2 contains the results from all the models for the entire testing year, 2005. The accuracy for each model is taken as the number of correctly predicted values divided by the total number of predictions. The adjusted accuracy for the total model removes the inaccurate predictions from Model 1 and Model 3 in order to provide a better estimate of the accuracy of the technique despite the subjective influence of the dataset.

The individual models were tested on 52, 59, 87, 51 and 91 days, respectively. The confusion matrices for each model are shown and can be interpreted individually. The two most notable models are Model 1 and Model 3.

Model 1, {January February} contains only {Low} temperature observations and {Bad} radiation days. The Hidden Markov Model is not accurate when given a fixed input. As seen in Appendix C, the parameters of Model 1 are approximately equal for transition probabilities and initial distribution. This suggests that no real training has taken place, and that the model is not suitable for use.

Model 3 receives input observations {Medium} and {High}. This model however initially returned a fixed state {Bad}. This is inherent in the nature of the HMM training. The model uses unknown states, and therefore cannot compare between states. The model returned {Bad} as the state only because it was the first position for state. By using logic and the known correlation between radiation and temperature, we can correct the state order for the models. This means reordering the states so that the state more likely to emit a higher temperature is the {Good} state. The original and adjusted model parameters can be seen in Appendix C.

6. CONCLUSION

Using Clustering and HMM can provide a crude estimate for the prediction and estimation of solar radiation. The assumptions associated with the HMM improve upon the probabilistic models but are not as accurate as the complex analyses which can be done. This methodology is suitable for a rough estimate when precision is not needed, but a general trend of the radiation is important. Using logic and some inherent properties of correlation make this modeling technique applicable and surprisingly accurate with fluctuations in the observed data.

Using clustering methods provided justified distinctions between the seasonal ranges during the year. These ranges are also not restricted to even distributions in the year. The unique segments found adhere to the general idea of the seasons but are bounded at better dates. By changing the ranges based on the data, a more accurate set of models is created.

Using Hidden Markov Models only approximates a state of the sun or environment, not the actual radiation output. Using logic and known correlations, the predicted state does provide a good estimate of radiation on a daily basis. The model trained can predict overall distribution of good and bad days, predict the state and observation into the future, and estimate the most likely observation sequence in any week of the year.

6.1. DISCUSSION

Considering the data set, the modeling results shown from this technique stand as a proof of concept. The technique provides more accurate results under periods of higher variation, such as during the fall when temperatures were highly variable, and radiation is decreasing. See Table 5.2 Model 4. This implies that the technique used will provide accurate results when used in a location that is not an arid desert.

This data is also very subjective because the only input observation is temperature. Sky coverage, or cloudiness, was used during the data segmentation, but 85% of the training data was clear skies. The majority of the clustering was based only on temperature; the addition of another variable added complexity without adding accuracy. For other datasets and other locations, more than one weather indicator may be used without additional complexity.

6.2. FUTURE WORK

Future work includes re-applying this method to other geographic locations. Model accuracy shows its ability to predict variable and fluctuating patterns; application in a less regular climate is suggestible.

Additional accuracy could be found by recording the sky coverage measurements in oktas (a one to eight value of cloudiness) before the automated recording system logs only the qualitative measure {BKN, OVC, etc.}. These qualitative measures do not correspond to actual numeric values, but to either a range or a singular value. This is not accurate when converting back to a number in a range.

APPENDIX A.

SURFRAD SENSOR ARRAY

Sensor Location at Desert Rock, Nevada

Latitude: 36.63 degrees North

Longitude: 116.02 degrees West

Elevation: 1007 meters

APPENDIX B.

EXAMPLE HIDDEN MARKOV MODEL

Example HMM:

Let's say we are stuck in a room. We are working on research and are unable to leave or see a window to outside. We are able to see a thermometer that reads the outside temperature. We can guess the solar radiation outside based on what the temperature it is shown on the thermometer.

We create a two state HMM to describe if it is bad or good radiation. This is in relation to a photovoltaic electricity system. The good radiation is a high level, and the bad radiation is a low level. Notation as good or bad insures difference between temperature levels. We are only able to make a general observation of the temperature; whether it is low, medium or high. Figure A1 displays such an HMM.

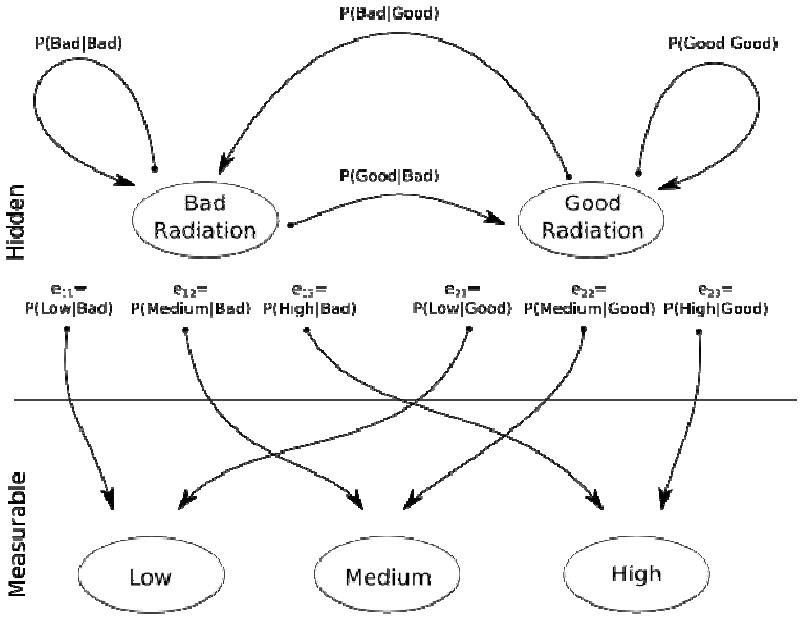

Figure A1: Bad or Good Radiation HMM

If we know the probabilities associated with this model, we are able to make predictions based on the observations of the thermometer. Let's assume we have already trained a model to have the values as follows:

Transitions (A):

		To	
		Bad	Good
From	Bad	0.60	0.40
	Good	0.30	0.70

Emissions (B):

		Observation		
		Low	Medium	High
State	Bad	0.20	0.10	0.70
	Good	0.35	0.60	0.05

Initial Distribution (π):

State	Bad	Good
Initial	0.20	0.80

We also see the following series of observations from the temperature: {Low, Medium, High}.

With the parameters of the model we can solve two of the most common problems associated with HMMs. The likelihood of this observation sequence at all, and the most likely state sequence to produce this series of observations. Let's call the likelihood of this observation series, Problem #1, and the most likely state series to emit these observations, Problem #2. Both techniques to solve Problems #1 and #2 can be extended to predict the next observation or state in the sequence, respectively.

Problem #1 Observation Likelihood:

The likelihood of each series of observations can be approximated by the sum of the emission probabilities from every possible state sequence. For our example, the state can be either Bad or Good, giving us a set of eight potential sequences. Take one state sequence, {Good, Good, Good}, as an example. The probability this state sequence emitted the sequence of variables can be defined by:

$$P(Good,Good,Good \mid Low,Medium,High) =$$

$$= \pi_{Sun} b_{Good}(Low) a(Good_2 \mid Good_1) b_{Good}(Medium) a(Good_3 \mid Good_2) b_{Good}(High)$$

$$P(Good,Good,Good \mid Low,Medium,High) = (0.80)(0.60)(0.70)(0.35)(0.70)(0.05) = 0.004116$$

The other state sequence possibilities can be calculated the same way and are found in Table A1

Table A1: State sequence likelihood to emit {Low, Medium, High}

State Sequences	Probabilities
Good, Good, Good	0.004116
Good, Good, Bad	0.024696
Good, Rain, Good	0.000576
Good, Bad, Bad	0.012096
Bad, Good, Good	0.000098
Bad, Good, Bad	0.000588
Bad, Bad, Good	0.000048
Bad, Bad, Bad	0.001008

Totaling the likelihood of every possible state sequence emitting the series of observations, you can find that the probability of observing the series {Low, Medium, High} is 0.043226.

Problem #1 and its solution can be algebraically described by the process:

Let the given HMM model be λ=(A, B, π) and let the series of observations be *O*= {O_1, O_2, O_3... O_{T-1}, O_T}. Problem #1 wants to find the P(*O*| λ).

Let *X*= {x_1, x_2, x_3... x_{T-1}, x_T}. Using the emission probabilities from B, we can describe the problem again as a series of emission probabilities:

$$P(O \mid X,\lambda) = b_{x1}(O_1) * b_{x2}(O_2)...b_{(xT-1)}(O_{(T-1)}) * b_{xT}(O_T)$$

Define the likelihood of a state sequence as:

$$P(X \mid \lambda) = \pi(x1) * a(x2 \mid x1) * a(x3 \mid x2)...a(xT-1 \mid xT-2) * a(xT-1 \mid xT-2)$$

Using the Conditional Probability, we know that

$$P(O,X \mid \lambda) = P\frac{(O \cap X \cap \lambda)}{(P(\lambda))}$$

Which allows us to find:

$$P(O \mid X,\lambda)P(X \mid \lambda) = P\frac{(O \cap X \cap \lambda)}{(P(X \cap \lambda))} \times P\frac{(X \cap \lambda)}{(P(\lambda))} = P\frac{(O \cap X \cap \lambda)}{(P(\lambda))}$$

Solved for the needed probability:

$$P(O,X \mid \lambda) = P(O \mid X,\lambda)P(X \mid \lambda)$$

Summing over all state sequences we find that:

$$P(O,X \mid \lambda) = \sum_X P(O,X \mid \lambda)$$

$$= \sum_X P(O \mid X,\lambda)P(X \mid \lambda)$$

$$= \sum_X \pi_1 * b_{x1}(O_1) * a(x2 \mid x1) * b_{x2}(O_2) ... * a(xT-1 \mid xT-2) * b_{(xT-1)}(O_{(T-1)})$$

The computation of this sum is largely inefficient and for complex applications has been evaluated with algorithms so that higher order models are feasible. This research uses the “forward pass” algorithm to reduce the number of multiplications from STN^2 down to N^2T.

Forward Pass:

Define the probability of the partial series of observations at time *t* :

$$\alpha_t(i) = P(O_1, O_{2,} \ldots O_t, x_t = q_i \mid \lambda)$$

The initial values of α are defined by

$$\alpha_1(i) = \pi_{xi} b(O_1 \mid x_i)$$

For *t*=1,2,3...T-1 and i=1,2...N compute:

$$\alpha_t(i) = \sum_{j=1}^{N} \left[\alpha_{(t-1)}(j) a(x_i \mid x_j)\right] b(O_t \mid x_i)$$

Which simplifies to:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_{(T-1)}(i)$$

Problem #2 Most Likely State Sequence:

As you can see in Table A1, the most likely probability is the state sequence: {Good, Good, Bad}. However this chained probability is limited to system without independent probabilities for each time, *t*. Using the table finds the dynamic programming solution, but it is not necessarily the most likely solution from the HMM. The answer must be confirmed by taking the total of each sequence with each state in each position. For example, the probability of the first state as Good, would be the first four sequence probabilities over the total probability of the series of observations. The state probabilities at each time are shown in Table A2.

Table A2: State Probabilities to emit {Low, Medium, High}

	t=1	t=2	t=3
Good	0.95970018	0.68241336	0.11192338
Bad	0.04029982	0.31758664	0.88807662

This leaves us with confirmation from the HMM probabilities that the most likely state sequence is {Good, Good, Bad}.

Problem #2 can be described algebraically as follows.

Using the results of the forward pass algorithm, we additionally need a "backward" pass which iterates through the time series of data in the opposite direction, end to beginning.

Let us define β, the probability of the partial observation sequence after time *t*

$$\beta_t(i) = P(O_{(t+1)}, O_{(t+2)}, \ldots O_{(T-1)} \mid x_t = q_i, \lambda)$$

For *t*=t, t+1...T-2,T-1 and *i*= 1, 2... N

$\beta_t(i)$ can be computed recursively the same way that $\alpha_t(i)$ was previously.

$$\beta_{(T-1)}(i) = 1$$

For *t*=t, t+1...T-2 and *i*= 1, 2... N

$$\beta_{(t)}(i) = \sum_{j=0}^{N} a(x_i \mid x_j)\, b_j(O_{(t+1)})\, \beta_{(t+1)}(j)$$

Using both the forward and backward pass probabilities together, the most likely state at time *t* can be defined by $\gamma_t(i)$:

$$\gamma_t(i) = P(x_t = q_t \mid O, \lambda)$$

This is evaluated as:

$$\gamma_t(i) = \frac{(\alpha_t(i)\beta_t(i))}{(P(O \mid \lambda))}$$

Training:

When parameters of the system are unknown, the accurate model for the system becomes a more complicated problem. The HMM prediction model is reliant on an accurate number of states and estimated probabilities.

HMMs parameter estimations are improved by two distinct methods; supervised and unsupervised training. Using supervised training, the data is separated into two parts: the training set and the testing set. For unsupervised training, data can be either separated or can be left with a continuously updated model current to a moving segment of the entire data. The model we will be using attempts to estimate data with supervised training using the previous year as a training set of data.

To adjust or "train" a model's parameters, use the following process:

1. Initialize $\lambda = (A, B, \pi)$
2. Compute $\alpha_t(i)$, $\beta_t(i)$, $\gamma_t(i)$, and $\gamma_t(i, j)$
3. Re-Estimate the model $\lambda = (A, B, \pi)$
4. If the $P(O \mid \lambda)$ increases, go back to 2; otherwise, quit

The estimation process relies on random initialization values for all the parameters. In addition to the values found from the forward and backward passes, there must be an additional measure $\gamma_t(i, j)$ to estimate the likelihood for being in state q_i and transitioning to state q_j.

$$\gamma_t(i, j) = P(x_t = q_i, x_{(t+1)} = q_j \mid O, \lambda)$$

Written in terms of α, β, A, and B:

$$\gamma_t(i, j) = \frac{(\alpha_t(i)a(x_i \mid x_j)b_j(O_{(t+1)})\beta_{(t+1)}(j))}{(P(O \mid \lambda))}$$

Using the measures that we have found using the initial model, the parameters can be re-estimated by the following formulas:

For the initial probabilities, π, while *i* = 1,2... N :

$$\pi_i = \gamma_1(i)$$

For the transition probabilities, A, while *i* = 1,2... N and *j* = 1,2... N :

$$a(xj \mid xi) = \frac{\left(\sum_{t=1}^{T-1} \gamma_t(i,j)\right)}{\left(\sum_{t=1}^{T-1} \gamma_t(i)\right)}$$

For the emission probabilities, B, *j* = 1,2... N and k = 1, 2 … M :

$$b(O_k \mid x_j) = \frac{\left(when\ \ O_k \quad \sum_{t=1}^{T-1} \gamma_t(j)\right)}{\left(\sum_{t=1}^{T-1} \gamma_t(j)\right)}$$

APPENDIX C.

MODEL PARAMETERS

Original Parameters:

		Transition Matrix				Emission Probabilities				Initial Probabilities	
		To				Observation					
Model 1		Bad	Good			Low	Medium	High			
From	Bad	0.556423	0.443577		State Bad	1	0	0		State Bad	0.468193021
	Good	0.487779	0.512221		Good	1	0	0		State Good	0.531806979
		To				Observation					
Model 2		Bad	Good			Low	Medium	High			
From	Bad	0.998279	0.001721		State Bad	0.120691	0.827585	0.051724		State Bad	3.83007E-05
	Good	0.000172	0.999828		Good	1	0	0		State Good	0.999961699
		To				Observation					
Model 3		Bad	Good			Low	Medium	High			
From	Bad	1	0		State Bad	0	0.288889	0.711111		State Bad	1
	Good	0	0		Good	0	0	0		State Good	0
		To				Observation					
Model 4		Bad	Good			Low	Medium	High			
From	Bad	0.999839	0.000161		State Bad	0	0.155108	0.844892		State Bad	0.012095368
	Good	0.023402	0.976598		Good	0	0	1		State Good	0.987904632
		To				Observation					
Model 5		Bad	Good			Low	Medium	High			
From	Bad	0.989782	0.010218		State Bad	0.797753	0.191011	0.011236		State Bad	3.59197375391527*^-7
	Good	3.63E-05	0.999964		Good	0	1	0		State Good	0.999999641

Adjusted Parameters:

		Transition Matrix				Emission Probabilities				Initial Probabilities	
		To				Observation					
Model 1		Bad	Good			Low	Medium	High			
From	Bad	0.556423207	0.443576793		State Bad	1	0	0		State Bad	0.468193021
	Good	0.487778734	0.512221266		Good	1	0	0		State Good	0.531806979
		To				Observation					
Model 2		Bad	Good			Low	Medium	High			
From	Bad	0.99982813	0.00017187		State Bad	1	0	0		State Bad	0.999961699
	Good	0.00172098	0.99827902		Good	0.120690816	0.827585114	0.05172407		State Good	3.83007E-05
		To				Observation					
Model 3		Bad	Good			Low	Medium	High			
From	Bad	0	0		State Bad	0	0	0		State Bad	0
	Good	0	1		Good	0	0.288888889	0.711111111		State Good	1
		To				Observation					
Model 4		Bad	Good			Low	Medium	High			
From	Bad	0.999838914	0.000161086		State Bad	0	0.155108494	0.844891506		State Bad	0.012095368
	Good	0.023401832	0.976598168		Good	0	0	1		State Good	0.987904632
		To				Observation					
Model 5		Bad	Good			Low	Medium	High			
From	Bad	0.989781864	0.010218136		State Bad	0.797752803	0.191011242	0.011235955		State Bad	3.59197375391527*^-7
	Good	3.63208E-05	0.999963679		Good	0	1	0		State Good	0.999999641

BIBLIOGRAPHY

Aguiar, R. J., M. Collares-Pereira, and J. P. Conde. 1988. "Simple Procedure for Generating Sequences of Daily Radiation Values using a Library of Markov Transition Matrices." Solar Energy 40 (3): 269-279.

Alam, S., S. C. Kaushik, and S. N. Garg. 2009. "Assessment of Diffuse Solar Energy Under General Sky Condition using Artificial Neural Network." Applied Energy 86 (4): 554-564.

Ångström, Anders. 1924. “Report to the International Commision for Solar Research on Actinometric Investigations of Solar and Atmospheric Radiation.” Quarterly Journal of the Royal Meteorological Society 50 (210): 121-126.

Bernal-Agustín, J. L. and R. Dufo-López. 2009. "Simulation and Optimization of Stand-Alone Hybrid Renewable Energy Systems." Renewable and Sustainable Energy Reviews 13 (8): 2111-2118.

Bollinger, B. and K. Gillingham. 2012. "Peer Effects in the Diffusion of Solar Photovoltaic Panels." Marketing Science 31 (6): 900-912.

Brinkworth, B. J. 1977. "Autocorrelation and Stochastic Modelling of Insolation Sequences." Solar Energy 19 (4): 343-347.

Haykin, S. O., 2011. “Neural Networks and Learning Machines.” McMaster University: Prentice Hall.

Kaundinya, D. P., P. Balachandra, and N. H. Ravindranath. 2009. "Grid-Connected versus Stand-Alone Energy Systems for Decentralized Power-A Review of Literature." Renewable and Sustainable Energy Reviews 13 (8): 2041-2050.

Liu, G., M. G. Rasul, M. T. O. Amanullah, and M. M. K. Khan. 2012. "Techno-Economic Simulation and Optimization of Residential Grid-Connected PV System for the Queensland Climate." Renewable Energy 45: 146-155.

Mustacchi, C., V. Cena, and M. Rocchi. 1979. "Stochastic Simulation of Hourly Global Radiation Sequences." Solar Energy 23 (1): 47-51.

Notton, G., C. Cristofari, M. Muselli, and P. Poggi. 2004. "Calculation on an Hourly Basis of Solar Diffuse Radiations from Global Data for Horizontal Surfaces in Ajaccio." Energy Conversion and Management 45 (18-19): 2849-2866.

Notton, G., P. Poggi, and C. Cristofari. 2006. "Predicting Hourly Solar Radiations on Inclined Surfaces Based on the Horizontal Measurements: Performances of the Association of Well-Known Mathematical Models." Energy Conversion and Management 47 (13-14): 1816-1829.

Olseth, J. A. and A. Skartveit. 1984. "A Probability Density Function for Daily Insolation within the Temperate Storm Belts." Solar Energy 33 (6): 533-542.

Ouammi, A., D. Zejli, H. Dagdougui, and R. Benchrifa. 2012. "Artificial Neural Network Analysis of Moroccan Solar Potential." Renewable and Sustainable Energy Reviews 16 (7): 4876-4889.

Panahandeh, B., J. Bard, A. Outzourhit, and D. Zejli. 2011. "Simulation of PV-Wind-Hybrid Systems Combined with Hydrogen Storage for Rural Electrification." International Journal of Hydrogen Energy 36 (6): 4185-4197.

Paulescu, M., L. Fara, and E. Tulcan-Paulescu. 2006. "Models for Obtaining Daily Global Solar Radiation from Air Temperature Data." Atmospheric Research 79 (3-4): 227-240.

Prieto, J. I., J. C. Martínez-García, and D. García. 2009. "Correlation between Global Solar Radiation and Air Temperature in Asturias, Spain." Solar Energy 83 (7): 1076-1085.

Ross, S. M., 2009. “Introduction to Probability Models.” Massachusetts: Academic Press.

Şahin, A. D. 2007. "A New Formulation for Solar Radiation and Sunshine Duration Estimation." International Journal of Energy Research 31 (2): 109-118.

Şen, Z. 2007. "Simple Nonlinear Solar Radiation Estimation Model." Renewable Energy 32 (2): 342-350.

Şen, Z. and S. M. Cebeci. 2008. "Solar Radiation Estimation by Monthly Principal Component Analysis." Energy Conversion and Management 49 (11): 3129-3134.

Skiba, M., M. Mohr, and H. Unger. 1997. "A Simple Model for Estimating Monthly Mean Daily Sums of Solar Radiation and its Local Distribution." International Journal of Energy Research 21 (12): 1145-1155.

Tiba, C. and R. E. A. Beltrão. 2012. "Siting PV Plant Focusing on the Effect of Local Climate Variables on Electric Energy Production - Case Study for Araripina and Recife." Renewable Energy 48: 309-317.

Türkay, B. E. and A. Y. Telli. 2011. "Economic Analysis of Standalone and Grid Connected Hybrid Energy Systems." Renewable Energy 36 (7): 1931-1943.

Tzamalis, G., E. I. Zoulias, E. Stamatakis, E. Varkaraki, E. Lois, and F. Zannikos. 2011. "Techno-Economic Analysis of an Autonomous Power System Integrating Hydrogen Technology as Energy Storage Medium." Renewable Energy 36 (1): 118-124.

Xu, R., Wunsch, D. C.. 2009. “Clustering.” Hoboken: IEEE Press.

Yacef, R., M. Benghanem, and A. Mellit. 2012. "Prediction of Daily Global Solar Radiation Data using Bayesian Neural Network: A Comparative Study." Renewable Energy 48: 146-154.

VITA

Frank Sauer. He graduated high school from Granbury High June 2006. He graduated from University of Missouri-Rolla, December 2010 with a Bachelors of Science in Engineering Management and a Bachelor of Arts in Economics. He continued to pursue a Master of Science Degree in Systems Engineering under the guidance of Dr. Ivan Guardiola. The focus of his studies was the statistical representation and prediction of energy data.