AFIT-ENV-MS-19-M-192

MODELING POWER OUTPUT OF HORIZONTAL SOLAR PANELS USING MULTIVARIATE LINEAR REGRESSION AND RANDOM FOREST MACHINE LEARNING

THESIS

Christil K. Pasion, 2d Lt, USAF

AFIT-ENV-MS-19-M-192

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the United States Government and is not subject to copyright protection in the United States.

AFIT-ENV-MS-19-M-192

MODELING POWER OUTPUT OF HORIZONTAL SOLAR PANELS USING MULTIVARIATE LINEAR REGRESSION AND RANDOM FOREST MACHINE LEARNING

THESIS

Presented to the Faculty

Department of Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Engineering Management

Christil K. Pasion

2d Lt, USAF

March 2019

AFIT-ENV-MS-19-M-192

MODELING POWER OUTPUT OF HORIZONTAL SOLAR PANELS USING MULTIVARIATE LINEAR REGRESSION AND RANDOM FOREST MACHINE LEARNING

Christil K. Pasion
2d Lt, USAF

Committee Membership:

Lt Col Torrey Wagner, PhD
Chair

Lt Col Clay Koschnick, PhD
Member

Maj Steven Schuldt, PhD
Member

Kevin Hallinan, PhD
Member

AFIT-ENV-MS-19-M-192

Abstract

United States Air Force energy resiliency goals are aimed to increase renewable energy implementation among its facilities. Researchers at the Air Force Institute of Technology designed, manufactured, and distributed 37 photovoltaic test systems to Air Force installations around the world. This research uses two types of modeling techniques, multivariate linear regression and random forest machine learning, to determine which technique will better predict power output for horizontal solar panels. Many previous solar panel prediction studies use solar irradiation data as an input. This study does not use irradiation as an input and aims to predict power output with input variables that are more readily available. If power output of a horizontal solar panel can be predicted using available weather data, then assessing the possibility of utilizing horizontal panels in any global location becomes possible. Input variables used for each model was latitude, month, hour, ambient temperature, humidity, wind speed, cloud ceiling, and altitude. The variance each model accounted was used as a comparison measure. The multivariate linear regression model accounted for 56.2% of the variance in a sample validation dataset. The random forest machine learning model accounted for 65.8% variance. The random forest model outperformed the multivariate linear regression model by accounting for 9.6% more variance. The most important variable in reducing the random forest model mean squared error was the month of the year, closely followed by cloud ceiling. Wind speed was the least important variable in reducing model error. More predictor variables are needed to increase predictability of horizontal solar panel power output if irradiation is not present as an input.

Acknowledgments

Initial funding and sponsorship were provided by the Air Force Civil Engineer Center and the Air National Guard Bureau Operations Division.

Data collection was facilitated by numerous personnel located at each site who received and set-up the equipment. These site monitors also checked the equipment for damage and functionality periodically and sent monthly data from their respective site(s) back to AFIT researchers.

Tec^Edge Works provided indoor workspace for the researchers to conduct manufacturing, testing, and distribution of the test systems. Roof-space at the location was also utilized for one of the test systems. Tec^Edge Works is part of the Wright Brothers Institute whose foundational stakeholders include the Air Force Research Laboratory.

Christil K. Pasion

Table of Contents

Page

List of Figures

Page

List of Tables

Page

List of Equations

Page

MODELING POWER OUTPUT OF HORIZONTAL SOLAR PANELS USING MULTIVARIATE LINEAR REGRESSION AND RANDOM FOREST MACHINE LEARNING

I. INTRODUCTION

Clean and renewable energy as a means to supplement power is currently being utilized on Air Force installations in order to align with the Air Force's strategic energy goal to improve resiliency [1]. Solar photovoltaic panels, also known as solar panels, are one form of renewable energy. This technology converts solar irradiance into electricity. Solar irradiance defines the output of energy from the sun onto a given area on the surface of the Earth [2]. The research presented in this thesis builds upon prior work aimed at determining how temperature and humidity affect the power output of horizontal solar panels in multiple different climate regions [3-4]. This study analysis encompasses a longer timeframe, incorporates more climatic factors into the analysis, and utilizes two separate modeling techniques. This study aims to broaden the knowledge of the effects of weather variables on horizontal solar panels and explore available modeling techniques. The remainder of this chapter discusses the general issue and background information behind this research, the problem statement driving the study, research questions this study is aiming to answer, assumptions regarding items outside of the researcher's control, research scope, a brief methodology overview, and materials and equipment utilized for data collection.

General Issues and Background

The Air Force accounts for 48% of all energy consumed by the Department of Defense (DoD) [1]. In the 2017 Air Force Energy Flight Plan, the Department emphasizes three energy goals: improve resiliency, optimize demand, and assure supply [1]. Under the goal to assure the Air Force energy supply, a strategic objective is to increase facility use of clean energy by 25% by fiscal year 2025. To achieve this goal, a variety of clean energy options must be considered. According to the Flight Plan, clean energy includes renewable sources. Multiple sources of renewable energy options, such as wind and hydropower, exist and are heavily used. Although these options may help the Air Force achieve its energy goals in its own ways, this study focuses on solar energy because of its availability across a wide range of locations.

One form of technology used to capture solar energy is solar photovoltaic (PV) panels. Capturing solar energy via solar PV panels is a clean and renewable source of energy production. PV panels are becoming more available, less expensive, and more efficient [5]. The panels can be arranged as an array in a field and oriented horizontally, vertically, or tilted toward the sun’s path at a specific location to capture solar irradiation and convert it to usable energy. Several Air Force installations began using solar PV panels to produce renewable energy for their power needs. As an example, Nellis Air Force Base and the United States Air Force Academy currently utilize large PV arrays. Each of these arrays has hundreds of PV panels spread over 140 acres and 43 acres [6-7]. This method takes up acres of space and not all Air Force installations have the available land to install a solar array on their base. Horizontal solar panels used in ways other than field arrays can broaden the possibilities of where and how solar energy can be converted to electricity on Air Force installations. Solar roadways are one such possible use for horizontal solar panels. This option utilizes previously allocated space and does not have to use any additional space. The work currently conducted in the field of solar roadways and solar pavements is further explored in Chapter II.

Problem Statement

Numerous PV system performance prediction models sponsored by the Department of Energy (DoE) have been developed, including the Solar Advisory Model, PVWATTS, and PVFORM [8]. The models were developed for tilted panel array applications and were validated using tilted panels at Sandia Laboratory in Albuquerque, NM. These models are effective in predicting the power output of PV arrays because the behavior of latitude-tilted panels that are elevated on racks is generally well known to modelers [8]. External factors affecting the behavior of horizontal solar panels are seldom studied across numerous locations and climate regions.

As mentioned previously, tilted solar PV panel arrays often take up acres of land space. As a result, not all Air Force installations have the space for a PV array that will yield a high enough return to meet their energy requirements. Horizontal solar panels may be used in novel ways, such as solar PV pavements or atop certain types of flat roofs. However, it is not fully known whether horizontal solar PV panels will produce a sufficient amount of power generation to justify their investment. Therefore, factors that have the most influence on the horizontal panel’s ability to convert solar energy into electricity must be explored to apply that information in future modeling efforts.

Materials and Equipment

The materials utilized for this study include monocrystalline PV panels, polycrystalline PV panels, Raspberry Pi computer systems, Pelican cases, CAT cables, power cables, and SD cards. All of these items comprise the test system at each location. Some locations also needed an external power source, in the form of a battery.

Methodology

Previous researchers designed and constructed 40 test systems for this study [3-4], [9]. The systems are designed to measure solar PV panel power output, temperature, humidity, and other data. This data was recorded at each site, for two panel types, every 15 minutes using a Raspberry Pi computer system and an SD card. The computer system is stored in a waterproof Pelican case and plugged into the PV panels and the temperature/humidity probe. These test systems were distributed to Air Force installations around the world in order to understand the effects different climate characteristics, locations, and ambient conditions have on horizontal solar PV panel power output. Data from a weather station near each test system location was also incorporated into the analysis.

Using the collected data, two types of models were developed: multivariate linear (MLR) regression and random forest (RF) machine learning. The validation R-squared values of each model were compared in order to determine which model accounts for more variance in its predictions. MLR was utilized to determine the relationship between multiple independent variables and one dependent variable: instantaneous power output. The following dependent variables were used to predict power output: latitude, month, hour, humidity, ambient temperature, wind speed, cloud cover, and altitude. The RF model was used to determine which input factors are most effective in reducing the mean squared error when predicting solar panel power output.

Research Questions

This study consists of three research questions. Each question's objective is to contribute to current research in order to further the field's understanding of horizontal solar PV panels and photovoltaic modeling techniques.

1. Which model type accounts for more variance in power output of horizontal solar PV panels when irradiation is unknown: linear regression or random forest machine learning?
2. Which input variables reduce mean squared error the most in predicting the power output of horizontal solar panels when irradiance is unknown?
3. What is the relationship of input factors, such as temperature and humidity, with power output of horizontal solar panels?

Assumptions

Assumptions were made to facilitate the study's operation and achieve research objectives. The equipment will perform in a manner consistent with the manufacturer's specifications for the duration of the study. Each of the geographical site's weather activity will meet expectations for the particular location's climate for the duration of the study. It is also assumed that data gaps will not adversely affect the model. Examples of potential causes are power outages, extremely cold temperatures, and system damage. It is also assumed that the site monitors will be reactive and responsive to any issues regarding the test systems they are monitoring. The site monitors will inform the researchers of any issues with the systems and assist in rectifying the situation in a timely matter.

Scope

Two types of photovoltaic panels were tested out of many available in the market, as the two types comprised a majority of the solar panel market at the time of this study [10]. Due to funding limitations, spare parts were not always supplied and some systems had to be decommissioned due to irreparable damage. This was accounted for by utilizing redundant sites.

As mentioned earlier, there are many renewable energy technology options available for use. Not only are there other ways to harness solar energy aside from PV panels, there are also ways to harness other forms of energy, such and tidal and wind energy. The scope of this study only covers solar energy being harnessed via PV panels.

II. LITERATURE REVIEW

Introduction

Chapter II discusses solar photovoltaic panels, the concept and application of horizontally oriented solar photovoltaic (PV) panels, and current studies and models surrounding them. This research is important because solar energy technology traditionally requires an abundance of land to be utilized. However, not everyone has an abundance of land available; therefore, other options for using solar energy technology should be sought. This chapter provides an overview of the general issues and objectives surrounding the subject of this research. It will also give background on what solar PV panel technology is and how it is used today. Additionally, photovoltaic pavements will be discussed as an application for solar PV technology. The reader can expect an overview of what factors affect the performance of solar PV panels and current models incorporating these factors.

General Issues

Fossil fuel energy dependence raises concerns regarding energy security for the Department of Defense (DoD). The need to reduce the DoD's dependence on fossil fuels was identified in a 2006 study conducted by JASON, an independent scientific advisory group [11]. The study concluded that there is a need to reduce fuel use due to the logistical requirements of procuring the fuel [11]. In conjunction with the current use of fossil fuels and future energy demand reduction, renewable energy technology implementation can assist countries to achieve energy security at a national level. Fossil fuel supplies are finite; one method to mitigate future energy shortages is supplementing power production with renewable sources of energy [12]. As the cost of renewable energy components continues to decrease and their performance continues to increase, implementation of renewable energy alongside fossil fuels is more cost efficient and effective than ever before [12].

Energy usage, efficiency, resilience, and sustainability are critical issues for the DoD. In 2009, the DoD was the largest consumer of energy in the world [13]. Furthermore, DoD energy use accounts for 80% of all US federal government energy use, and energy sustainability is also a DoD strategic goal [14]. According to the US Department of Energy, energy sustainability is defined as the ability to operate without a decline in operational capability [15]. Renewable energy resources are sustainable, due to their theoretically endless supply, and this goal can be achieved by utilizing renewable energy technologies across a wide range of locations and facilities.

Air Force Energy Objectives

The US Air Force will play a large part in the achievement of the DoD’s energy goals due to the magnitude of energy that the Air Force requires for its operations. The US Air Force makes up 48% of the total DoD energy use and 11% of that is allotted for installation energy [1]. The installations mentioned are located both in the continental United States and overseas. Homeland installations primarily rely on commercial electricity grids, which are vulnerable to a number of threats including natural hazards and terrorism [14]. The military is addressing these vulnerabilities by implementing plans for increased use of renewable energy and ensuring access to reliable energy [14]. Installation energy, while vulnerable to electricity grid risks, is the primary beneficiary of solar energy initiatives within the Air Force. To achieve base energy resilience, installation dependence on electrical grids must be mitigated, which occurs through the use of renewable energy such as solar PV panels.

One of the Air Force's energy goals is to assure its energy supply. A strategic objective relating to that goal is to increase facility clean energy use by 25% by FY 2025 [1]. This objective aligns with the Air Force Strategic Master Plan vector to ensure a full-spectrum-capable, high-end force. The Air Force has made efforts to decrease energy consumption in recent years. In fiscal year (FY) 2016, Air Force square footage increased while energy production and intensity decreased from FY 2015 baselines [16]. Additionally, 6.8% of the Air Force's total energy consumption was supplied by renewable energy sources that year. This percentage continued to grow as over 300 renewable energy projects at over 100 different sites were in construction or in operation as of the start of FY 2017 [16]. Even with these accomplishments, additional work is required to meet the DoD's and Air Force's energy goals.

Clean energy contains many mechanisms, such as wind, solar, and hydropower. This study focuses on harnessing solar energy as a form of clean energy production for the Air Force to help the department meet its energy goals. By focusing on developing clean energy on-site, the Air Force can protect its facilities from grid failure and energy supply disruptions. On-site clean energy use and production can provide economic and environmental benefits as well by providing consistency in energy pricing and avoiding greenhouse gas emissions [1].

Air Force Solar Energy Initiatives

Nellis Air Force Base, located in Nevada, completed a 140-acre solar array in 2007 [7]. This tilted, open-frame array uses single-axis tracking that allows the panels to laterally track the sun across the sky [7]. The array also uses cleaning robots as opposed to manual cleaning conducted by human workers [17]. Following installation, the array was expected to save the base $1M in energy costs annually and provide 25% of the base's power needs [7].

A 180-acre solar array was installed in 2016 at Air Force Plant 42, a manufacturing plant in Palmdale, California [18]. The array is similar to the Nellis array in that the panels are tilted and have single-axis tracking. The array is designed to produce 20 megawatts of power per year, which would power over 3,000 homes. This project was part of a lease agreement between the Air Force and NRG Solar Oasis LLC at no cost to the Air Force [18].

Davis-Monthan Air Force Base, located in Arizona, unveiled a 170-acre solar array in 2014 [19]. This array was expected to generate 35% of the base's electricity energy needs and exceeded expectations a year later at over 40%. The base's electricity costs went from 8.6 cents per kilowatt-hour to only 4.6 cents per kilowatt-hour [19].

Each of these projects was power purchase agreements (PPAs) between the Air Force and a private company. There are multiple types of PPAs; the PPA mentioned above can be described as a government agency allowing the use of government land by a private company [20]. This private company uses the land to produce power and sell some of that power to the government agency. In this case, the government agency is the Air Force and the power produced from solar PV panels is installed and maintained by a private company.

In these examples, the Air Force provided land to solar energy companies in exchange for some of the energy produced by the array at a lower price than purchasing from the local energy grid. All of these arrays utilize a considerable amount of area. These panels are elevated and tilted; therefore, each panel must be spaced far enough from the others to avoid shading, which can significantly decrease the panel output [21]. Solar panels arranged horizontally can use ground area more efficiently because they can be positioned directly next to each other. This feature validates the practice of installing such panels on bases that do not have hundreds of unused acres of land.

Photovoltaic Technology Overview

This section presents an overview of how solar energy is harnessed using solar PV panels. The exact science of what is occurring internally for each type of panel used in this study will not be discussed.

The sun emits energy in the form of radiation [22]. The sun's radiation reaches the earth's atmosphere and some of it strikes the surface of the earth. The amount of solar radiation that reaches the surface of the earth depends on many factors. Some of those factors include the amount of water vapor in the atmosphere and the angle of the sun's position in the sky in relation to the horizontal plane of the surface of the earth. Photovoltaic arrays are made up of solar panels, with each panel composed of numerous solar cells. A silicon solar cell is a semiconductor that can harness a fraction of the sun's energy. The sun's energy striking the surface of the panel allows for an internal flow of electric current through the panel. This current can be directed to an external circuit to supply power [22].

Types of Solar PV Panels

Crystalline silicon is the most common material solar PV cells are made out of today, with almost 90% of the world market's photovoltaics made with silicon [10]. The two types of solar PV cells used in this study are types of crystalline silicon: monocrystalline silicon and polycrystalline silicon. At the time of this publication, solar PV cells made of monocrystalline silicon are typically 15-20% efficient in converting the sun's energy to usable electrical energy. Polycrystalline silicon solar PV cells are 13-16% efficient and less expensive than monocrystalline silicon. Furthermore, polycrystalline silicon PV cells have worse performance in higher temperatures when compared to monocrystalline silicon cells. This is a result of fabricating monocrystalline cells with higher grade silicon. The third type of solar PV cell is thin-film, which is made of various materials, instead of primarily silicon. Thin-film solar PV panels typically need more surface area to produce the same amount of power as monocrystalline silicon PV panels due to a lower efficiency of range of 7-13%. Thin-film is less expensive than both of its silicon counterparts [10]. Thin-film is not utilized in this study due to its lower efficiency and large space requirements.

Orientation of Solar PV Panels

Solar irradiance describes the amount of solar energy incident to the surface of the earth on a particular area [2]. A solar panel produces the most power when the panel is perpendicular to the solar irradiance being cast upon it by the sun during solar noon [23]. This orientation involves positioning the panel at a specific tilt angle relative to the horizontal plane and rotating the panel to track the sun's path of travel throughout the day

[23]. A rule of thumb for panel orientation was established for locations in the northern hemisphere. This standard, dating back to the 1980s, stated to tilt the panel the same number of degrees from horizontal as the latitude of the panel location and to face the panel due south [24-25].

Cloudy Days and Overcast Conditions

The ideal orientation mentioned previously for a solar PV panel refers to clear sky conditions. Studies have found that diffuse solar irradiation, which is irradiation scattered by water droplets suspended in the atmosphere, is best captured when a solar PV panel is oriented horizontally [23], [26-27].

The continental United States was the focus of another study that concluded that the optimal tilt angle is also a function of cloudiness, in addition to latitude. This finding challenged the latitude tilt angle rule of thumb and found that some optimal tilt angles can be up to 10 degrees less than latitude for locations in the continental United States due to seasonal weather patterns, such as winter clouds [23].

A study in Milford, MI stated that when the majority irradiance in an area is diffused, horizontally aligned solar PV panels will provide maximum solar irradiance when compared to tilted panels [26]. Using a horizontally oriented panel alongside tracking panels, the researchers predicted that positioning panels horizontally during overcast conditions could increase yearly yield by 1% [26]. A study in Europe found that a PV system that tracks the sun’s path will have a 3% increase in annual yield if the panels are moved to a horizontal position during overcast conditions [27].

Array Self-Shading

A study investigating a phenomenon known as self-shading discusses the issue of panels in an array shading each other [27]. This issue causes damaging hot spots and lower performance. One way to solve this problem is to increase the distance between each of the array's rows so they do not shade each other during sunrise and sunset. Since arrays currently use an abundance of land, this solution would decrease the energy density of an array's area, declaring it infeasible. The researchers solved this problem by positioning the panels horizontally during sunrise and sunset because shadows affect yield more than non-optimal tilt angles [27]. This is possible because the array in the study has a sun tracking system that can be programmed to a certain angle at a specific time of day.

Horizontal Solar Panel Application: Pavements

Solar PV technology may be used in ways that will not take up as much space as a traditional array. Solar panels can be arranged more densely when placed flat on the ground. When installed in such a manner, this technology may be used as solar pavements. The remainder of this section discusses solar PV technology used as pavement systems.

A study conducted in 2017 aimed to evaluate the feasibility of solar roads as a sustainable energy source [28]. Two prototypes, each of which used monocrystalline silicon panels, were constructed with two different cover layers: a polycarbonate sheet and a porous rubber sheet. These cover layers were necessary because they increased the friction on the surface of the panel to mimic the friction needed to prevent slipping while driving on roads. The supplied energy, surface safety movement, and structural performance were evaluated for each design. Both designs performed well in the skid resistance test and one design handled more load than the other. Each design had a percent decrease in power generation of 26 percent and 50 percent. This decrease is due to the cover layers that allows the panels to be used as pavements. This study concluded that solar roadways can provide benefits in the form of reduced fossil fuel consumption, reduced pollution, and access to electricity generated from solar energy [28].

A 70-meter solar bike path was installed in the Netherlands in 2015 [29]. The solar bike path consisted of 54 polycrystalline silicon modules embedded in concrete and covered in an anti-skid layer. The power output of the solar road was estimated to produce an annual efficiency of 9.08%. This prediction was slightly lower than that of a polycrystalline panel installed in the same area as the bike path at optimal tilt and orientation but is still an appealing amount of PV potential [29]. The researchers noted that using monocrystalline silicon cells could potentially increase the annual yield by 1.5 times. Furthermore, annual yield can also increase for different locations other than where this bike path was installed. Due to this application being a bike path, as opposed to a road for motorized vehicles, structural and weight bearing considerations were not considered. Using this application, the researchers concluded that the energy density potential justifies further exploration of solar roadway technology [29].

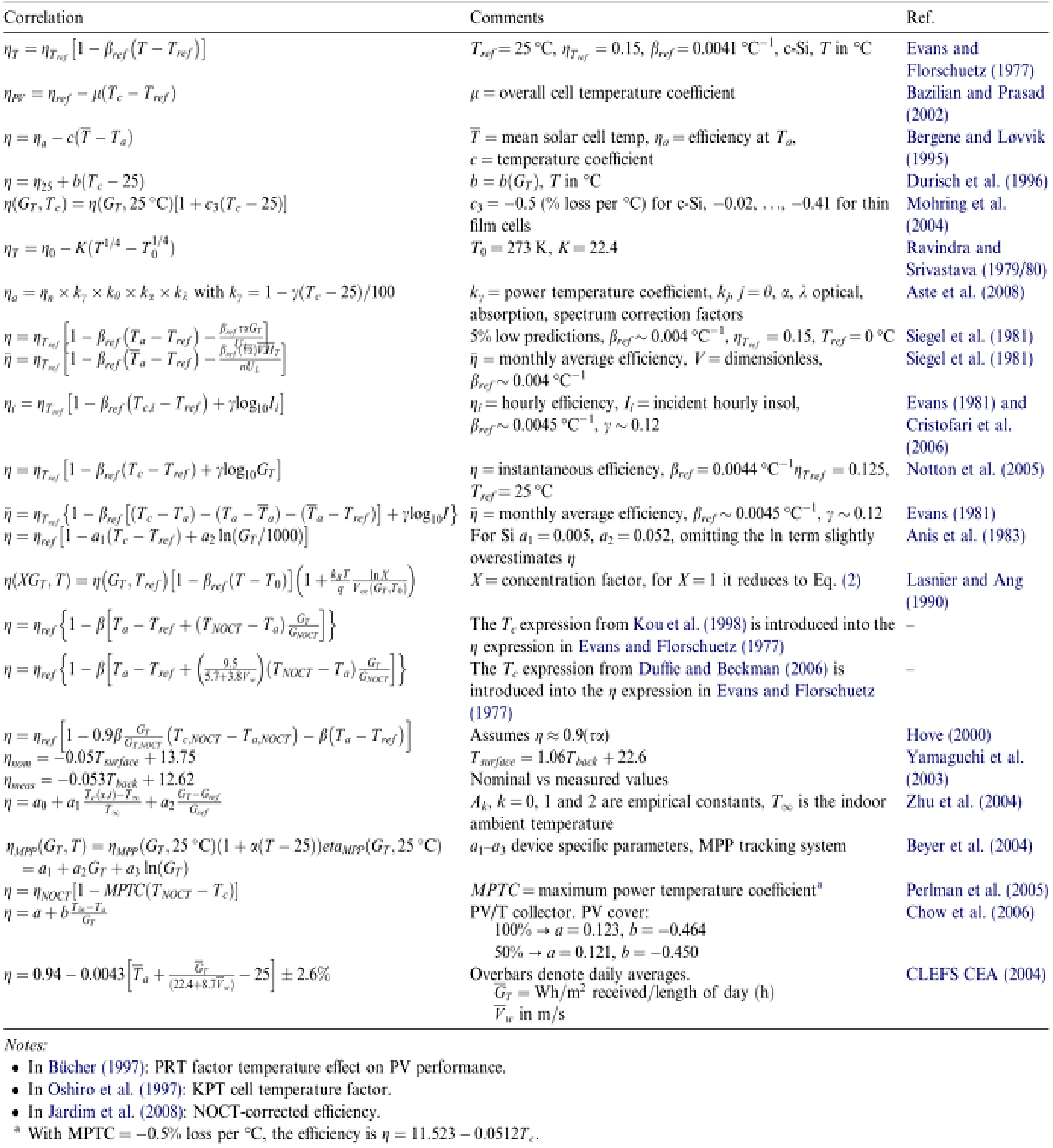
Temperature and Humidity Effects on Solar PV Panels

Existing solar PV panel power prediction equations use temperature as a function to describe efficiency and power output. The functions that incorporate ambient temperature account for more variance in the prediction models than the functions that omit ambient temperature [30]. Current literature highlighted 24 efficiency functions and 27 power output functions that utilize temperature as one of their factors [30]. Both efficiency and power equations are described in detail in Tables 1 and 2, respectively.

In 2008, a study compared the prediction performance of current photovoltaic models to empirical data [8]. The study included four radiation models, three module performance models, an inverter model, the PVWATTS model and PVMod models. After a temperature coefficient was incorporated into the models, solar panel output prediction accuracy improved between 2.1% and 10.4%, respectively [8].

Humidity is known for influencing PV performance. A primary example of this is the direct effect of water droplets diverting incoming sunlight through refraction, diffraction, and reflection [31]. Indirectly, humidity also affects dust build-up on panels due to the formation of dew increasing coagulation of dust. Mekhilef et al. (2012) reviewed previous studies conducted across multiple climates and reported differing solar PV panel performance drops for the studied climates [31]. As an example, a study conducted in the United States reported a peak efficiency drop of 4.7 % while a study in Saudi Arabia reported as much as 40 % performance degradation due to dust build-up on the surface of the panels [31].

Table 1. Competing equations describing solar PV panel efficiency with respect to temperature [30]**.**



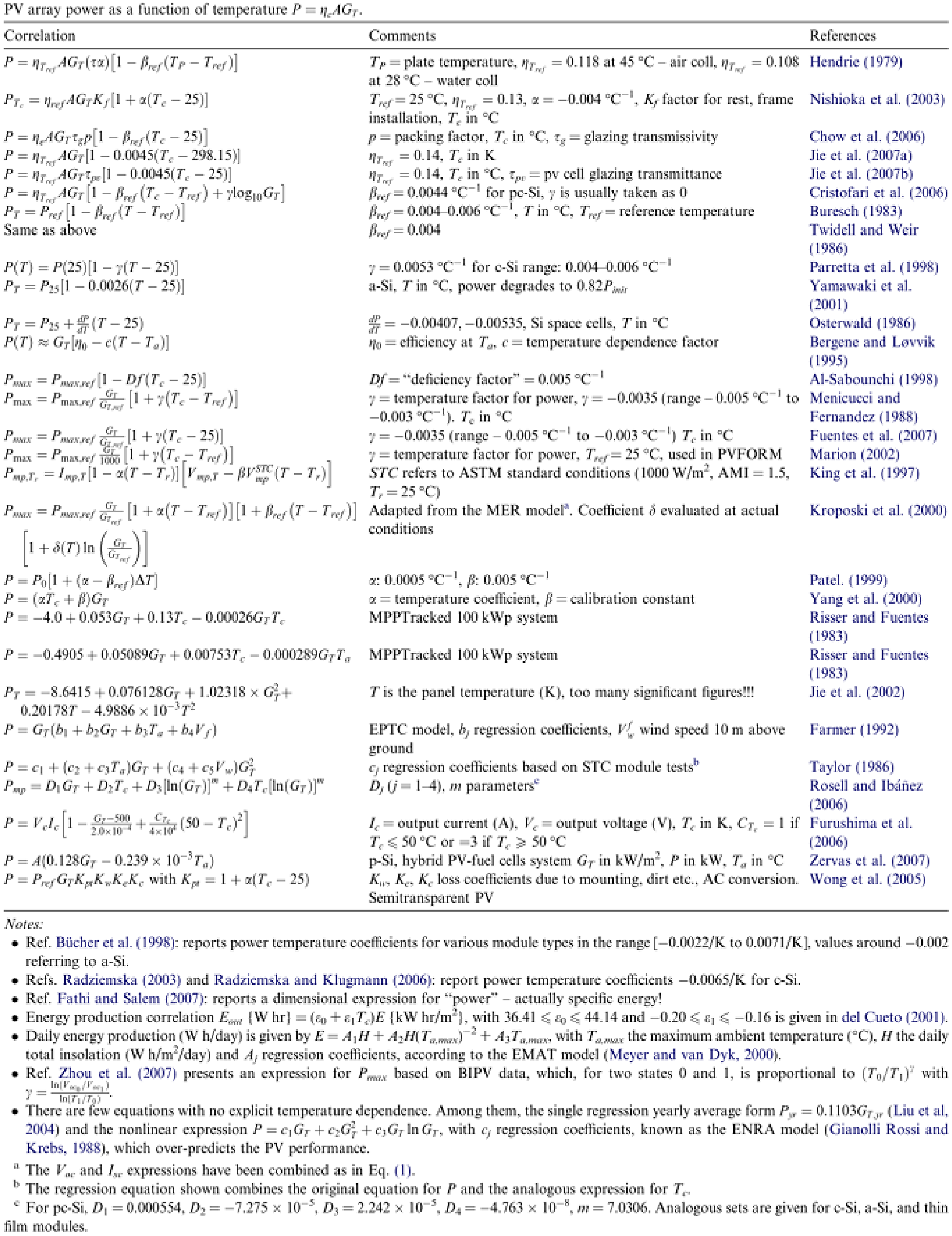
Correlation	Comments	Ref.
$\eta_T = \eta_{T_{ref}}\left[1-\beta_{ref}\left(T-T_{ref}\right)\right]$	$T_{ref} = 25\,°C$, $\eta_{T_{ref}} = 0.15$, $\beta_{ref} = 0.0041\,°C^{-1}$, c-Si, T in °C	Evans and Florschuetz (1977)
$\eta_{PV} = \eta_{ref} - \mu(T_c - T_{ref})$	μ = overall cell temperature coefficient	Bazilian and Prasad (2002)
$\eta = \eta_a - c(\overline{T} - T_a)$	$\overline{T}$ = mean solar cell temp, η_a = efficiency at T_a, c = temperature coefficient	Bergene and Løvvik (1995)
$\eta = \eta_{25} + b(T_c - 25)$	$b = b(G_T)$, T in °C	Durisch et al. (1996)
$\eta(G_T, T_c) = \eta(G_T, 25\,°C)[1 + c_3(T_c - 25)]$	$c_3 = -0.5$ (% loss per °C) for c-Si, −0.02, …, −0.41 for thin film cells	Mohring et al. (2004)
$\eta_T = \eta_0 - K(T^{1/4} - T_0^{1/4})$	$T_0 = 273$ K, $K = 22.4$	Ravindra and Srivastava (1979/80)
$\eta_a = \eta_n \times k_\gamma \times k_\theta \times k_\alpha \times k_\lambda$ with $k_\gamma = 1 - \gamma(T_c - 25)/100$	k_γ = power temperature coefficient, k_j, $j = \theta$, α, λ optical, absorption, spectrum correction factors	Aste et al. (2008)
$\eta = \eta_{T_{ref}}\left[1-\beta_{ref}\left(T_a - T_{ref}\right) - \frac{\beta_{ref}\tau\alpha G_T}{U_L}\right]$	5% low predictions, $\beta_{ref} \sim 0.004\,°C^{-1}$, $\eta_{T_{ref}} = 0.15$, $T_{ref} = 0\,°C$	Siegel et al. (1981)
$\bar{\eta} = \eta_{T_{ref}}\left[1-\beta_{ref}\left(\overline{T}_a - T_{ref}\right) - \frac{\beta_{ref}\left(\frac{\overline{\tau\alpha}}{\tau\alpha}\right)V\overline{H}_T}{nU_L}\right]$	$\bar{\eta}$ = monthly average efficiency, V = dimensionless, $\beta_{ref} \sim 0.004\,°C^{-1}$	Siegel et al. (1981)
$\eta_i = \eta_{T_{ref}}\left[1-\beta_{ref}\left(T_{c,i} - T_{ref}\right) + \gamma\log_{10}I_i\right]$	η_i = hourly efficiency, I_i = incident hourly insol, $\beta_{ref} \sim 0.0045\,°C^{-1}$, $\gamma \sim 0.12$	Evans (1981) and Cristofari et al. (2006)
$\eta = \eta_{T_{ref}}\left[1-\beta_{ref}(T_c - T_{ref}) + \gamma\log_{10}G_T\right]$	η = instantaneous efficiency, $\beta_{ref} = 0.0044\,°C^{-1}\eta_{T_{ref}} = 0.125$, $T_{ref} = 25\,°C$	Notton et al. (2005)
$\bar{\eta} = \eta_{T_{ref}}\{1-\beta_{ref}[(T_c - T_a) - (T_a - \overline{T}_a) - (\overline{T}_a - T_{ref})] + \gamma\log_{10}I\}$	$\bar{\eta}$ = monthly average efficiency, $\beta_{ref} \sim 0.0045\,°C^{-1}$, $\gamma \sim 0.12$	Evans (1981)
$\eta = \eta_{ref}\left[1 - a_1(T_c - T_{ref}) + a_2\ln(G_T/1000)\right]$	For Si $a_1 = 0.005$, $a_2 = 0.052$, omitting the ln term slightly overestimates η	Anis et al. (1983)
$\eta(XG_T, T) = \eta(G_T, T_{ref})\left[1-\beta_{ref}(T - T_0)\right]\left(1 + \frac{k_B T}{q}\frac{\ln X}{V_{oc}(G_T, T_0)}\right)$	X = concentration factor, for $X = 1$ it reduces to Eq. (2)	Lasnier and Ang (1990)
$\eta = \eta_{ref}\left\{1-\beta\left[T_a - T_{ref} + (T_{NOCT} - T_a)\frac{G_T}{G_{NOCT}}\right]\right\}$	The T_c expression from Kou et al. (1998) is introduced into the η expression in Evans and Florschuetz (1977)	–
$\eta = \eta_{ref}\left\{1-\beta\left[T_a - T_{ref} + \left(\frac{9.5}{5.7+3.8V_w}\right)(T_{NOCT} - T_a)\frac{G_T}{G_{NOCT}}\right]\right\}$	The T_c expression from Duffie and Beckman (2006) is introduced into the η expression in Evans and Florschuetz (1977)	–
$\eta = \eta_{ref}\left[1 - 0.9\beta\frac{G_T}{G_{T,NOCT}}\left(T_{c,NOCT} - T_{a,NOCT}\right) - \beta\left(T_a - T_{ref}\right)\right]$	Assumes $\eta \approx 0.9(\tau\alpha)$	Hove (2000)
$\eta_{nom} = -0.05T_{surface} + 13.75$	$T_{surface} = 1.06T_{back} + 22.6$	Yamaguchi et al. (2003)
$\eta_{meas} = -0.053T_{back} + 12.62$	Nominal vs measured values	
$\eta = a_0 + a_1\frac{T_c(x,t) - T_\infty}{T_\infty} + a_2\frac{G_T - G_{ref}}{G_{ref}}$	A_k, $k = 0$, 1 and 2 are empirical constants, T_∞ is the indoor ambient temperature	Zhu et al. (2004)
$\eta_{MPP}(G_T, T) = \eta_{MPP}(G_T, 25\,°C)(1 + \alpha(T - 25))eta_{MPP}(G_T, 25\,°C) = a_1 + a_2G_T + a_3\ln(G_T)$	a_1–a_3 device specific parameters, MPP tracking system	Beyer et al. (2004)
$\eta = \eta_{NOCT}[1 - MPTC(T_{NOCT} - T_c)]$	$MPTC$ = maximum power temperature coefficient[a]	Perlman et al. (2005)
$\eta = a + b\frac{T_{in} - T_a}{G_T}$	PV/T collector. PV cover: 100% → $a = 0.123$, $b = -0.464$ 50% → $a = 0.121$, $b = -0.450$	Chow et al. (2006)
$\eta = 0.94 - 0.0043\left[\overline{T}_a + \frac{\overline{G}_T}{(22.4+8.7\overline{V}_w)} - 25\right] \pm 2.6\%$	Overbars denote daily averages. $\overline{G}_T$ = Wh/m^2 received/length of day (h) $\overline{V}_w$ in m/s	CLEFS CEA (2004)

Notes:

- In Bücher (1997): PRT factor temperature effect on PV performance.
- In Oshiro et al. (1997): KPT cell temperature factor.
- In Jardim et al. (2008): NOCT-corrected efficiency.

[a] With MPTC = −0.5% loss per °C, the efficiency is $\eta = 11.523 - 0.0512T_c$.

Table 2. Competing equations describing solar PV panel power output with respect to temperature [30]**.**



PV array power as a function of temperature $P = \eta_c A G_T$.

Correlation	Comments	References
$P = \eta_{T_{ref}} A G_T (\tau\alpha)\left[1 - \beta_{ref}(T_P - T_{ref})\right]$	T_P = plate temperature, $\eta_{T_{ref}} = 0.118$ at 45 °C – air coll, $\eta_{T_{ref}} = 0.108$ at 28 °C – water coll	Hendrie (1979)
$P_{T_c} = \eta_{ref} A G_T K_f [1 + \alpha(T_c - 25)]$	$T_{ref} = 25$ °C, $\eta_{T_{ref}} = 0.13$, $\alpha = -0.004$ °C^{-1}, K_f factor for rest, frame installation, T_c in °C	Nishioka et al. (2003)
$P = \eta_c A G_T \tau_g p\left[1 - \beta_{ref}(T_c - 25)\right]$	p = packing factor, T_c in °C, τ_g = glazing transmissivity	Chow et al. (2006)
$P = \eta_{T_{ref}} A G_T [1 - 0.0045(T_c - 298.15)]$	$\eta_{T_{ref}} = 0.14$, T_c in K	Jie et al. (2007a)
$P = \eta_{T_{ref}} A G_T \tau_{pv} [1 - 0.0045(T_c - 25)]$	$\eta_{T_{ref}} = 0.14$, T_c in °C, τ_{pv} = pv cell glazing transmittance	Jie et al. (2007b)
$P = \eta_{T_{ref}} A G_T \left[1 - \beta_{ref}(T_c - T_{ref}) + \gamma \log_{10} G_T\right]$	$\beta_{ref} = 0.0044$ °C^{-1} for pc-Si, γ is usually taken as 0	Cristofari et al. (2006)
$P_T = P_{ref}\left[1 - \beta_{ref}(T - T_{ref})\right]$	$\beta_{ref} = 0.004$–0.006 °C^{-1}, T in °C, T_{ref} = reference temperature	Buresch (1983)
Same as above	$\beta_{ref} = 0.004$	Twidell and Weir (1986)
$P(T) = P(25)[1 - \gamma(T - 25)]$	$\gamma = 0.0053$ °C^{-1} for c-Si range: 0.004–0.006 °C^{-1}	Parretta et al. (1998)
$P_T = P_{25}[1 - 0.0026(T - 25)]$	a-Si, T in °C, power degrades to $0.82P_{init}$	Yamawaki et al. (2001)
$P_T = P_{25} + \frac{dP}{dT}(T - 25)$	$\frac{dP}{dT} = -0.00407, -0.00535$, Si space cells, T in °C	Osterwald (1986)
$P(T) \approx G_T[\eta_0 - c(T - T_a)]$	η_0 = efficiency at T_a, c = temperature dependence factor	Bergene and Løvvik (1995)
$P_{max} = P_{max,ref}[1 - Df(T_c - 25)]$	Df = "deficiency factor" = 0.005 °C^{-1}	Al-Sabounchi (1998)
$P_{\max} = P_{\max,ref} \frac{G_T}{G_{T,ref}} \left[1 + \gamma(T_c - T_{ref})\right]$	γ = temperature factor for power, $\gamma = -0.0035$ (range – 0.005 °C^{-1} to -0.003 °C^{-1}). T_c in °C	Menicucci and Fernandez (1988)
$P_{max} = P_{max,ref} \frac{G_T}{G_{T,ref}} [1 + \gamma(T_c - 25)]$	$\gamma = -0.0035$ (range – 0.005 °C^{-1} to -0.003 °C^{-1}) T_c in °C	Fuentes et al. (2007)
$P_{\max} = P_{\max,ref} \frac{G_T}{1000} \left[1 + \gamma(T_c - T_{ref})\right]$	γ = temperature factor for power, $T_{ref} = 25$ °C, used in PVFORM	Marion (2002)
$P_{mp,T_r} = I_{mp,T}[1 - \alpha(T - T_r)]\left[V_{mp,T} - \beta V_{mp}^{STC}(T - T_r)\right]$	STC refers to ASTM standard conditions (1000 W/m^2, AMI = 1.5, $T_r = 25$ °C)	King et al. (1997)
$P_{max} = P_{max,ref} \frac{G_T}{G_{T_{ref}}} \left[1 + \alpha(T - T_{ref})\right]\left[1 + \beta_{ref}(T - T_{ref})\right]$ $\left[1 + \delta(T) \ln\left(\frac{G_T}{G_{T_{ref}}}\right)\right]$	Adapted from the MER model[a]. Coefficient δ evaluated at actual conditions	Kroposki et al. (2000)
$P = P_0[1 + (\alpha - \beta_{ref})\Delta T]$	α: 0.0005 °C^{-1}, β: 0.005 °C^{-1}	Patel. (1999)
$P = (\alpha T_c + \beta)G_T$	α = temperature coefficient, β = calibration constant	Yang et al. (2000)
$P = -4.0 + 0.053G_T + 0.13T_c - 0.00026G_T T_c$	MPPTracked 100 kWp system	Risser and Fuentes (1983)
$P = -0.4905 + 0.05089G_T + 0.00753T_c - 0.000289G_T T_a$	MPPTracked 100 kWp system	Risser and Fuentes (1983)
$P_T = -8.6415 + 0.076128G_T + 1.02318 \times G_T^2 +$ $0.201787T - 4.9886 \times 10^{-3}T^2$	T is the panel temperature (K), too many significant figures!!!	Jie et al. (2002)
$P = G_T(b_1 + b_2 G_T + b_3 T_a + b_4 V_f)$	EPTC model, b_j regression coefficients, V_w^f wind speed 10 m above ground	Farmer (1992)
$P = c_1 + (c_2 + c_3 T_a)G_T + (c_4 + c_5 V_w)G_T^2$	c_j regression coefficients based on STC module tests[b]	Taylor (1986)
$P_{mp} = D_1 G_T + D_2 T_c + D_3[\ln(G_T)]^m + D_4 T_c[\ln(G_T)]^m$	D_j (j = 1–4), m parameters[c]	Rosell and Ibáñez (2006)
$P = V_c I_c \left[1 - \frac{G_T - 500}{2.0 \times 10^{-4}} + \frac{C_{T_c}}{4 \times 10^4}(50 - T_c)^2\right]$	I_c = output current (A), V_c = output voltage (V), T_c in K, $C_{T_c} = 1$ if $T_c \leqslant 50$ °C or =3 if $T_c \geqslant 50$ °C	Furushima et al. (2006)
$P = A(0.128G_T - 0.239 \times 10^{-3}T_a)$	p-Si, hybrid PV-fuel cells system G_T in kW/m^2, P in kW, T_a in °C	Zervas et al. (2007)
$P = P_{ref} G_T K_{pt} K_w K_e K_c$ with $K_{pt} = 1 + \alpha(T_c - 25)$	K_w, K_e, K_c loss coefficients due to mounting, dirt etc., AC conversion. Semitransparent PV	Wong et al. (2005)

Notes:

- Ref. Bücher et al. (1998): reports power temperature coefficients for various module types in the range [−0.0022/K to 0.0071/K], values around −0.002 referring to a-Si.
- Refs. Radziemska (2003) and Radziemska and Klugmann (2006): report power temperature coefficients −0.0065/K for c-Si.
- Ref. Fathi and Salem (2007): reports a dimensional expression for "power" – actually specific energy!
- Energy production correlation E_{out} {W hr} $= (\varepsilon_0 + \varepsilon_1 T_c)E$ {kW hr/m^2}, with $36.41 \leqslant \varepsilon_0 \leqslant 44.14$ and $-0.20 \leqslant \varepsilon_1 \leqslant -0.16$ is given in del Cueto (2001).
- Daily energy production (W h/day) is given by $E = A_1 H + A_2 H(T_{a,max})^{-2} + A_3 T_{a,max}$, with $T_{a,max}$ the maximum ambient temperature (°C), H the daily total insolation (W h/m^2/day) and A_j regression coefficients, according to the EMAT model (Meyer and van Dyk, 2000).
- Ref. Zhou et al. (2007) presents an expression for P_{max} based on BIPV data, which, for two states 0 and 1, is proportional to $(T_0/T_1)^\gamma$ with $\gamma = \frac{\ln(V_{oc_0}/V_{oc_1})}{\ln(T_1/T_0)}$.
- There are few equations with no explicit temperature dependence. Among them, the single regression yearly average form $P_{yr} = 0.1103G_{T,yr}$ (Liu et al, 2004) and the nonlinear expression $P = c_1 G_T + c_2 G_T^2 + c_3 G_T \ln G_T$, with c_j regression coefficients, known as the ENRA model (Gianolli Rossi and Krebs, 1988), which over-predicts the PV performance.

[a] The V_{oc} and I_{sc} expressions have been combined as in Eq. (1).

[b] The regression equation shown combines the original equation for P and the analogous expression for T_c.

[c] For pc-Si, $D_1 = 0.000554$, $D_2 = -7.275 \times 10^{-5}$, $D_3 = 2.242 \times 10^{-5}$, $D_4 = -4.763 \times 10^{-8}$, $m = 7.0306$. Analogous sets are given for c-Si, a-Si, and thin film modules.

The study conducted by Mekhilef et al. (2012) concluded that dust accumulation is affected by both humidity and wind speed, with higher humidity increasing the accumulation and higher wind speed decreasing accumulation. Solar PV panels with higher tilt angle have less dust accumulation than those with lower tilt angles, which is a drawback for horizontal PV arrays [31].

The United States Department of Energy supports models produced by Sandia National Laboratories that predict the performance of PV panels based on solar radiation and weather data. Tests used to further validate these models against empirical data use panels tilted at latitude and facing toward the sun. Therefore, these models may not be sufficient for use with horizontally oriented panels because horizontal panels may respond differently to various weather effects. For example, wind may cool tilted panels more effectively than horizontal panels because the horizontal panels would be placed close to the ground where the air from the wind may not reach the back side of the panel as easily.

Other Factors Affecting Solar Panels

In order to predict the power output or efficiency of a panel, researchers have modeled experimental data utilizing the following input variables: irradiation, temperature, humidity, solar elevation angle, wind speed, wind direction, month, and others [31–34]. Furthermore, every model has not focused on the same factors, which can impact the results on short, medium, and long timescales.

Table 3 summarizes four photovoltaic studies mentioned throughout this document, including the study discussed, for comparison purposes. Table 3 highlights that numerous factors are the subject of any single photovoltaic study, depending on the research objectives surrounding the work.

Busquet et al. (2015) studied how the environment and the age of the solar panel affected the daily energy output using factors such as aging and soiling, with aging not commonly used by other studies [32]. Aging describes the amount of time the panel has been installed and exposed to the elements. Soiling describes the dust build-up of the panel's surface. Kayri et al. (2017) and Lahouar et al. (2017) forecasted solar panel power output and used short-term factors such as solar elevation angle and wind direction, they, however, did not include longer-term factors such as aging [33-34]. Mekhilef et al. conducted a review primarily interested in the effects of dust, humidity, and air velocity, such as water droplets trapped inside the cell and dew causing dust accumulation [31].

Solar irradiance is one common factor that the four studies used. The research this paper describes does not use solar irradiation as an input but instead uses other predictor variables to account for that unknown. For instance, the *hour of day* may account for solar elevation angle and latitude. Climatic variables, such as pressure, may indicate the presence of rain, which decreases the amount of solar irradiation a panel receives. Solar irradiation data is not widely available and often needs to be found using previously developed models and global radiation data as an input. This study aims to investigate the viability of predicting horizontal solar panel power output using available data, such as position, time, and weather. If the power output of a solar panel can be reasonably predicted without including irradiation as an input, then assessing the possibility of utilizing horizontal panels in any global location becomes possible. Such a model would accurately predict power output for horizontal panels with the use of readily available inputs, such as location and weather data.

Table 3. A summary of various studies factors of interest and other details of the studies, such as the type of panel and location. Data from Busquet [32], Kayri [33], Lahouar [34], and Mekhilef [31].

study author		GP3L Study Model	Busquet	Kayri	Lahouar	Mekhilef
analysis type		random forest linear regression	least-square linear regression	multiple linear regression random forest artificial neural network	random forest forcasting	case study
type of panel		monocrystalline	many	unknown	unknown	many
orientation		horizontal	20 deg tilt	unknown	unknown	many
location		over 20	Hawaii	Turkey	Australia	many
output		power	daily energy	power	max power	efficiency
factors	impact					
hour of day	short	x				
month	med	x			x	
humidity	short	x		x	x	x
ambient temperature	short	x	x	x	x	
wind speed / air velocity	short	x	x	x	x	x
visibility	short	x				
atmospheric pressure	short	x				
cloud ceiling	short	x				
altitude	long	x				
latitude	long	x				
soiling (dust)	med		x		x	x
aging	long		x			
solar elevation angle	short			x		
solar irradiation	short		x	x	x	x

Irradiation

In the numerous photovoltaic studies that utilize modeling, irradiation is found to be the most important factor on solar panel power output [33-34]. Cloud cover and the angle of the sun in relation to the Earth affects the irradiation of that area at a specific time [35]. Once the irradiation reaches a panel, other factors will then affect the power output of that panel. The panel's individual efficiency (as determined by the technology used), the age of the panel, and the soiling of the panel will all affect the power output [32]. Furthermore, the ambient temperature surrounding the panel will influence the surface temperature of the panel and high temperatures generally lead to decreased panel efficiency [36-37].

In order to predict the irradiation that strikes the horizontal plane in a specific area of the earth’s surface, studies have developed models that use various datasets as input factors including ground weather, satellite remote-sensing, and sun position. Unlike studies concerned with the efficiency of a solar panel and its performance once the irradiation reaches the panel, these studies forecast the amount of irradiation that reaches the tilted panels.

A study in Taiwan used surface solar radiation forecasting to estimate solar irradiation for solar panels [38]. The study developed machine learning models to forecast surface solar radiation and the solar irradiation received by solar panels at different tilt angles. Alongside the satellite and sun position data, the following ground weather variables were used: atmospheric pressure, wind speed, precipitation, temperature, and relative humidity. The researchers also included radiation, an objective weather variable. This data was recorded on an hourly basis using a weather station on the ground. The researchers used these variables in their models to predict both the direct and diffuse horizontal irradiance striking the surface of the earth. They then estimated the global irradiance within the next hour with the solar panel set at a tilted position. The model was validated using a solar panel positioned at the same tilt angle of the estimated model. Using the study’s global predicted values of the total annual global irradiance received by panels at various tilt angles between 0 degrees and 41 degrees, the researchers were able to determine the optimal tilt angle for panels in that region [38]. This experiment excluded an empirical study involving input data from a physical solar panel, but developed its models using available weather and radiation data then compared the model outputs to panels in the field.

Another study estimating solar radiation was conducted in Sfax, Tunisia [35]. Using a numerical method, the researchers aimed to calculate the global hourly irradiance for a range of tilted surfaces in Sfax. The range of tilts was from 0 degrees to 90 degrees increasing in increments of 10 degrees. The inputs included the monthly average solar radiation on a horizontal surface, the tilt angle of the panel, the panel location (latitude and longitude), and the albedo of the area surrounding the panel. The monthly average solar radiation on a horizontal surface was obtained from NASA, Surface Meteorology and Solar Energy. The predicted average global irradiance outputs for each tilt angle were compared to the Photovoltaic Geographical Information System (PVGIS) and there was a 3% difference [33]. This study did not use the weather variables the previous study did because the values for the solar irradiation on a horizontal surface were measured values from NASA instead of being obtained via model predictions.

Studies conducted in 1991 by Faine et al. and in 2016 by Eke et al. discuss the factors that have the most profound effect on solar spectral irradiance [39-40]. Spectral irradiance is the power density at a particular wavelength and affects the performance of solar panels. Spectral irradiance is primarily affected by the following variables: air mass (AM), clouds, turbidity, water vapor, and surface pressure. Turbidity describes aerosol effects. AM describes the path length of the solar beam through the atmosphere. The standard test conditions for a solar panel include an AM value of 1.5. Under that condition, the sun is at a deflection angle of 48.2 degrees [39]. In real-world outdoor conditions, the AM value will deviate from the standard condition and the solar panel performance will be influenced by the real-time AM at any given moment [40]. AM values increase with longer path lengths. The longer the path length of the sun’s radiation, the more likely scattering and absorption of solar radiation will occur by variables such as clouds, aerosols, and water vapor. The path length is determined by the sun angle which varies with latitude, time of day, and the day of the year [39-40]. Therefore, controlling for AM in the LM and RF models to the best extent possible will be imperative to each model’s accuracy and predictability.

External Effects and Internal Losses

Busquet et al (2015), was primarily interested in the power output of the panel in relation to both external and physical factors affecting its ability to output power [32]. These factors included the age of the panel, the soiling level of the panel, and temperature and wind speed of the area surrounding the panel. These factors, along with irradiation as an input, gave Busquet insight into how such factors will affect the panels ability to convert the irradiation it receives into usable electricity. Factors such as aging describe the long-term degradation of the panel’s power output over its lifetime. While temperature and humidity affect the instantaneous energy output of the panel, soiling decreases the energy output of the panel and is improved when rain washes off the dust and debris causing the soiling. By using irradiation as an input, Busquet was able to focus on parameters that are affecting the panel’s performance.

A study conducted in Hong Kong in 2007 developed a model that is dependent on solar intensity and model temperature. The experimental set-up involved a solar panel placed under a 3-phase array of lamps that simulates sunlight. The incident solar irradiance on the plane of the solar panel was measured using a pyrometer. The following six parameters were measured: solar irradiance, solar panel temperature, short-circuit current, open circuit voltage, and the maximum power point current and voltage. Using an experimental set-up, the researchers were able to isolate the panel affects to only that of the energy coming into the panel and the temperature of the panel. When the model, using the size parameters mentioned above, was validated against real-world solar panels in Hong Kong, and an R-squared of 0.98 was achieved for sunny conditions and 0.96 for cloudy conditions. This study confirmed that weather, especially panel temperature and solar irradiance, strongly influences the irradiation that affects solar PV panel performance [41].

NOAA Weather Data

The National Oceanic and Atmospheric Administration (NOAA) hosts a website where weather and climatic data from all over the world can be accessed [42]. This extensive repository of data can be used to determine weather norms in an area, track weather trends, plan for future renewable energy projects, predictive modeling, and a multitude of other applications. The NOAA has access to airport weather stations in the United States and some overseas locations. The following weather variables were gathered from airport weather stations for this study: wind direction, wind speed, cloud ceiling, visibility, outdoor temperature, dewpoint temperature, and station pressure. This work uses NOAA data in its analysis and how it is extracted and used will be discussed in Chapter III. A description of each weather variable, as described by the NOAA, can be seen in Table 4. The starred variables were chosen as initial input variables.

Table 4. Data descriptions for NOAA weather data used in this study. Seven variables were used from the dataset. The starred variables were chosen as input variables.

Measurement	Description
Wind Direction	Wind direction in compass degrees, 990 = variable, 0 when air is calm (speed will then be 0)
*Wind Speed	Wind speed in miles per hour
*Cloud Ceiling	Cloud ceiling--lowest opaque layer with 5/8 or greater coverage, in hundreds of feet, 722 = unlimited
*Visibility	Visibility in statute miles to nearest tenth
Outdoor Temperature	Temperature in Fahrenheit
Dewpoint Temperature	Dew point in Fahrenheit
*Atmospheric Pressure	Station pressure in millibars to nearest tenth

Photovoltaic Statistical Modeling

Multivariate Linear Regression

Regression is a statistical evaluation that focuses on the relationship between independent and dependent variables A regression is done by employing a model developed using observed variable values [43]. One regression technique, known as linear regression, is used for variable relationships that are linear between the independent and dependent variables. Multivariate linear regression (MLR) is used when more than one independent variable is included in the model. In an MLR model, the dependent variable (Y) is described as a linear function of the independent variables (X_n). Regression coefficients (b_n) for each independent variable is computed by the model as well as the regression line intercept (b_o). The subscript *n* represents the number of independent variables used to produce the model. The function is as follows:

$$Y = b_o + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

The value of each regression coefficient (b_n) is the amount of change in the dependent variable (Y) for every unit change in the independent variable (X_n) [43].

Linear models have been applied in many ways since their conception and are widely used. One example of its applications in the photovoltaic field is in a study by Hammad et al. published in 2018 [44]. The research was on the effects of dust accumulation and ambient temperature on photovoltaic performance. An MLR model was developed to estimate the solar panel conversion efficiency given experimental data of exposure time to natural dust and ambient temperature. The model was then applied to recommend the cleaning frequency needed for the optimal power output of a PV system installed in North Africa [44].

One advantage of an MLR model is that coefficients given for each predictor variable can reveal how that variable affects the independent variable. For example, in an MLR model with solar panel power output as the response variable and six different predictor variables, including temperature, Kayri et al. (2017) found that the temperature coefficient was positive [33]. The researchers concluded that temperature has a positive effect on solar panel power output for their experimental context. Wind speed, on the other hand, had a negative coefficient and as airspeed increased, power output decreased, with all other variables held constant [33].

A disadvantage to MLR is its sensitivity to outliers. If one or more outliers are present in the dataset used to develop the model, it may disproportionally affect the regression and may also improperly suggest a lack of fit [45]. This is primarily a concern in smaller datasets where one outlier can have a more profound effect on the model [45].

Furthermore, multiple assumptions regarding the predicted variables error terms must be met in order for the model to be considered valid. If some or all of the assumptions are not met, more data may need to be gathered or different modeling techniques may need to be considered. The definition of these assumptions and how they are tested will be discussed in Chapter III.

Random Forest Regression

Random forest (RF) regression is a nonparametric machine learning model formulation technique developed by Breiman [45-46]. A nonparametric model does not have assumptions regarding the form of the function the model is fitting. Therefore, an RF model has the possibility to fit a wider range of functions than a parametric model [48]. An RF model is composed of an ensemble of decision trees. Each decision tree uses binary splitting in order to output a numerical estimation of the output value given the inputs. The sample is chosen at random from the dataset, with replacement. Given the number of total input variables, p, each decision tree will try a number of input variables less than p at each node in order to create a split. This number, *mtry*, is typically $p/3$ depending on the problem and can be seen as a tuning parameter [48]. The value of the predictor is the average of the numerical outcome given by the decision trees. With hundreds of decision trees using a set of variables less than the total amount of variables to make predictions, the model is robust and overfitting is seldom [49]. Robust models perform well with data drawn from a distribution other than normal [45]. Overfitting is seldom because the decision tree is not always given the option to use the best predictor to split the first node of a tree and it forces other predictors to be used. Thus, the trees are uncorrelated and the average predicted output has less variation and higher reliability than highly correlated trees that used the primary predictor variable to make an estimate [48]. An example of a decision tree that could be used to classify an animal based on information given regarding the physical characteristics of the animal can be seen in Figure 1. The characteristics are both numerical, such as how big the animal is, and categorical, such as whether or not the animal has horns.

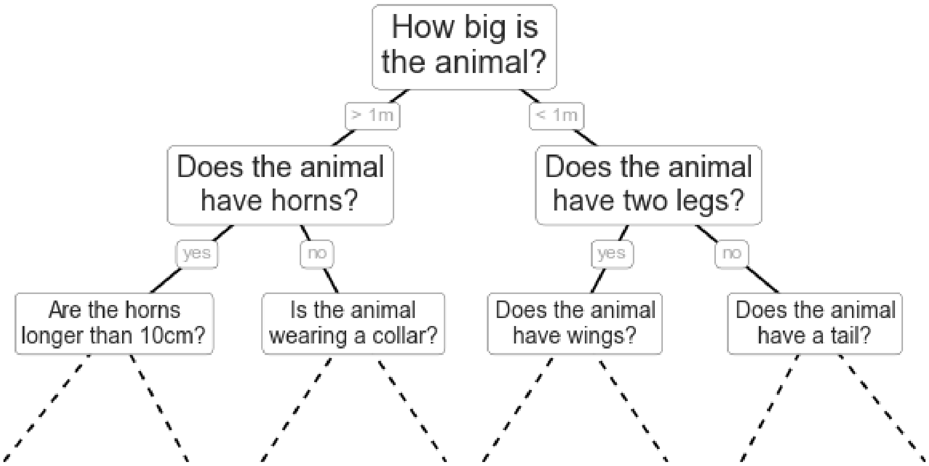

How big is the animal?
> 1m
< 1m
Does the animal have horns?
Does the animal have two legs?
yes
no
yes
no
Are the horns longer than 10cm?
Is the animal wearing a collar?
Does the animal have wings?
Does the animal have a tail?

Figure 1. A decision tree used to classify an animal based on information given regarding the physical characteristics of the animal. The characteristics are both numerical, such as how big the animal is, and categorical, such as whether or not the animal has horns. [46]

In addition to the advantages mentioned above, RF models output relative variable importance for all model input variables [49]. The model will output the rank of every input variable based on how well each variable decreases the mean squared error. The primary disadvantage of RF models is their interpretability. The results of an RF model are not easily interpretable and conclusions cannot easily be drawn regarding the meaning of the RF regression model [46].

Studies Comparing Modeling Techniques

A study done in Turkey by Kayri et al. (2017) compared three modeling techniques using photovoltaic and atmospheric data [33]. The photovoltaic data consisted of the power output of a panel located in Turkey. The three modeling techniques were MLR, RF, and artificial neural network (ANN). The researchers used the following input factors in order to predict the power output of a photovoltaic module using each of the three models: global radiation, ambient temperature, humidity, wind speed, wind direction, and solar elevation angle. The researchers found that when comparing the correlation coefficients of the real and estimated values of each model, the ANN performed the best (r = 0.997) and the RF (r = 0.986) outperformed the MLR (r = 0.963).

In a study conducted by Lahouar et al. (2017), photovoltaic power forecasting was carried out using an RF regression model [34]. Using PV sites located at the University of Queensland in Australia, the researchers compared the power output of the actual panels to their forecast model's predicted power output. Of the model's inputs, solar irradiance was the most important variable followed by humidity and then temperature. This was a forecasting model and the researchers used only future temperature, future humidity, and current power values to forecast future power output. The forecasted output was compared to the actual data. In this case, the prediction results for three different techniques were compared: persistence (PER), ANN, and RF. Without any parameter tuning or optimization, the RF outperformed the PER and ANN when comparing the model's mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error.

Researchers in Taiwan compared four different modeling techniques used to forecast solar irradiation striking a tilted panel. The four techniques were: multilayer perception (MLP), RF, k-nearest neighbor (kNN), and MLR. When the RMSE, the MAE, and the correlation coefficient (r) were compared for every model, it was determined that MLR had the worst prediction performance and did not apply to such a study. It was thought that this was due to linear models creating errors when applied to nonlinear problems. The researchers did not specify which component or variable from this study was not linear. MAE values for RF and kNN were similar and MLP had the least stable prediction results as the forecast horizon increased.

Photovoltaic Model Variable Importance

Kayri et al. (2017) noted the most important variables in their MLR, RF, and multilayer perception artificial neural network (MLP-ANN) models [33]. For this study, photovoltaic power of a solar panel is the dependent variable and six independent atmospheric factors were used: global radiation, temperature, wind speed, wind direction, relative humidity, and solar elevation angle. For the three models, all six predictors were meaningful and global radiation was the most important factor. For the MLR, humidity was the next important, then temperature. Elevation angle was the least important factor for the MLR. The RF and MLP-ANN had different factor rankings. Solar elevation angle followed and wind direction was the least important for each of those two models. Furthermore, temperature was more important than humidity in each, as opposed to the opposite in the MLR [33].

III. METHODOLOGY

Introduction

The purpose of this investigation is to provide further insight into the most important input factors for predicting power output for solar PV panels in a horizontally oriented position. Of the input factors, irradiation will not be included. This investigation will explore the relationships between numerous climatic factors and power output for solar PV panels across multiple locations using one year of empirical field data. The benefits of investigating these input factors and their relationships enable future researchers to determine the viability of solar pavement technology and begin developing predictive performance models for PV technology.

Materials and Equipment

The hardware and software for this study's test system were designed by Captain John Nussbaum and the Electrical Engineering Department located at the Air Force Institute of Technology (AFIT) [3]. Captain Joseph Applebee and Captain Cory Booker manufactured the 40 test systems at Tec^Edge Works located in Dayton, OH [4-5]. An example of one of the test systems assembled in the field can be seen in Figure 2. Within this system, an external power source was not present to power the recording devices. Therefore, an external battery (bottom right corner of the figure) and a third panel to charge the battery (top right corner) were included. The systems were sent to Air Force installations around the world. Site monitors at each location set up their respective system. The global location of the 37 initial test sites can be seen in Figure 3.

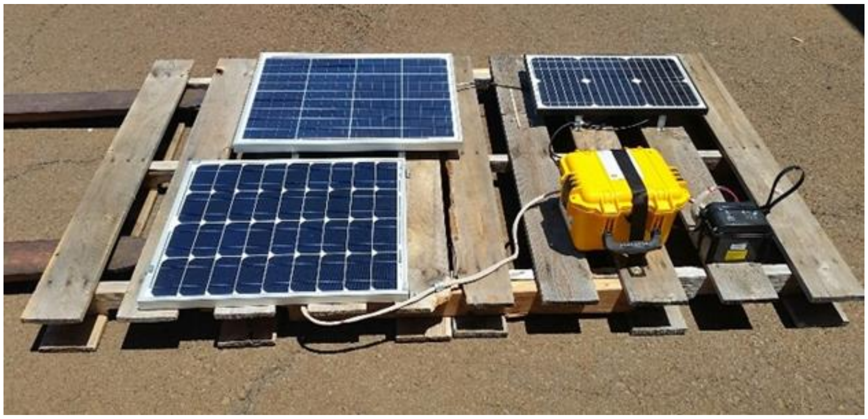


Figure 2. The assembled test system in the field. The polycrystalline solar panel is positioned at the top right with the monocrystalline panel below it. The yellow waterproof case is present as well. The external battery and the third panel to charge the battery are each in the bottom right and top right of the photo, respectively. [4], [9]

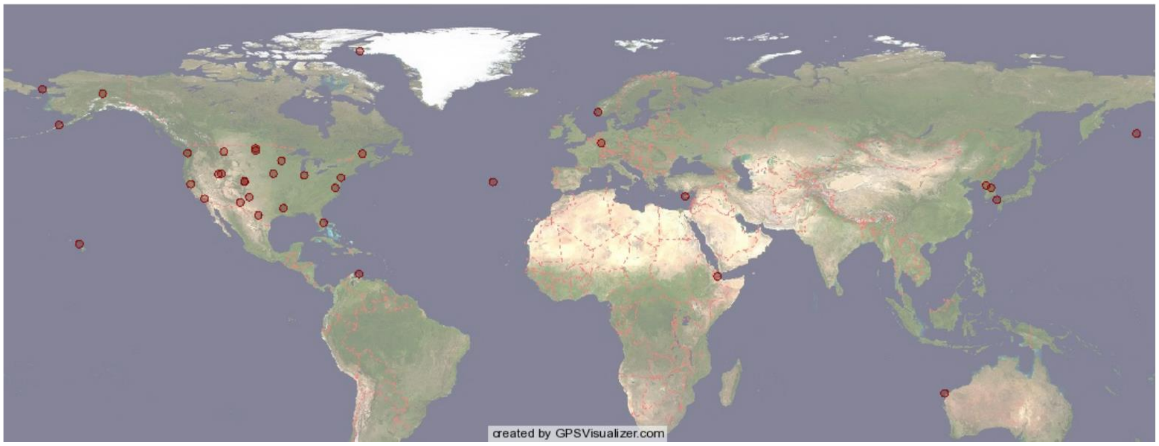


Figure 3. The planned location of each test system at the inception of the study. 37 test sites are shown in this figure. [9]

The test systems were comprised of the following equipment: a Renogy 50-watt, 12-volt, monocrystalline solar PV panels; an ALEKO 25-watt, 12-volt, polycrystalline solar PV panels; Raspberry Pi 3, model B, version 1.2 computer systems; waterproof Pelican cases, CAT cables, power cables, and SD cards. The Raspberry Pi computer system gives the test system the ability to record the panels power outputs, temperature and humidity readings, date, and time, in 15-minute intervals. The SD card in the computer was retrieved by the site monitors and downloaded every month and the dataset was sent to the researchers. Site monitors at each location were given instruction to clean off the panel whenever dust or snow cove was observed. Although this was performed daily for some locations, others were cleaned off less frequently. The infrequency of panel cleaning at some locations may have an affect on the modeling efforts being pursued by this study. Not having the ability to account for this affect has been identified as a limitation of this research.

Procedures and Processes

The geographical location of each test systems was also determined by Captain John Nussbaum. He conducted a statistical Analysis of Variance (ANOVA) of the latitude and longitude coordinates of 1,763 Air Force installations, dividing the world into five latitude and five longitude bins [3]. A Pareto analysis was conducted on all of the sites using the Koppen-Geiger climate classification system and the 25 regions identified by an ANOVA. The Pareto analysis allowed for prioritized placement of test systems by identifying which regions represented a majority of Air Force installations [3]. Therefore, the effects of climatic conditions are being determined in the regions where most Air Force installations exist.

During the experimental set-up phase, a few incidents occurred. One site received a cracked panel which was replaced soon after notification of the damage [4], [9]. As a result, this site started gathering data at a later date. Two sites decided to forego participation in the study.

Data Compilation and Exclusions

Both panels for each location's system outputted the power data as separate measurements of voltage and current. For every fifteen-minute interval, voltage and current were simultaneously measured every ten seconds, 64 times. Power, in watts, was calculated by multiplying voltage and current together. The maximum value from the 64 measurements for each fifteen-minute interval was found. This resulted in four power measurements for every hour. The date (YYYYMMDD), time (HHMM), ambient temperature, and humidity was recorded once every 15 minutes and corresponded to each power measurement. Each location's name, latitude, and longitude were added to the dataset. Latitude and longitude were recorded to two decimal places. It was determined which locations had at least eight months of recordings. All locations that met the cutoff were combined into one dataset. At this point, it was determined that the monocrystalline solar panel measurements were inconsistent and unreliable and would not be used for modeling. For example, the monocrystalline panel at the United States Air Force Academy, in Colorado, recorded data above 400 watts, much outside of the panel's 25-watt rating. The two sites in Hawaii and Florida did not record any power values above 10 watts, which is an unexpectedly low output for locations known for sunny weather. Other sites, such as Camp Murray, had measurements that were almost double the panel rating. The distributions for these four sites can be seen in Figure 4. After identifying these discrepancies across many sites, it was decided to proceed using only the polycrystalline panel data.

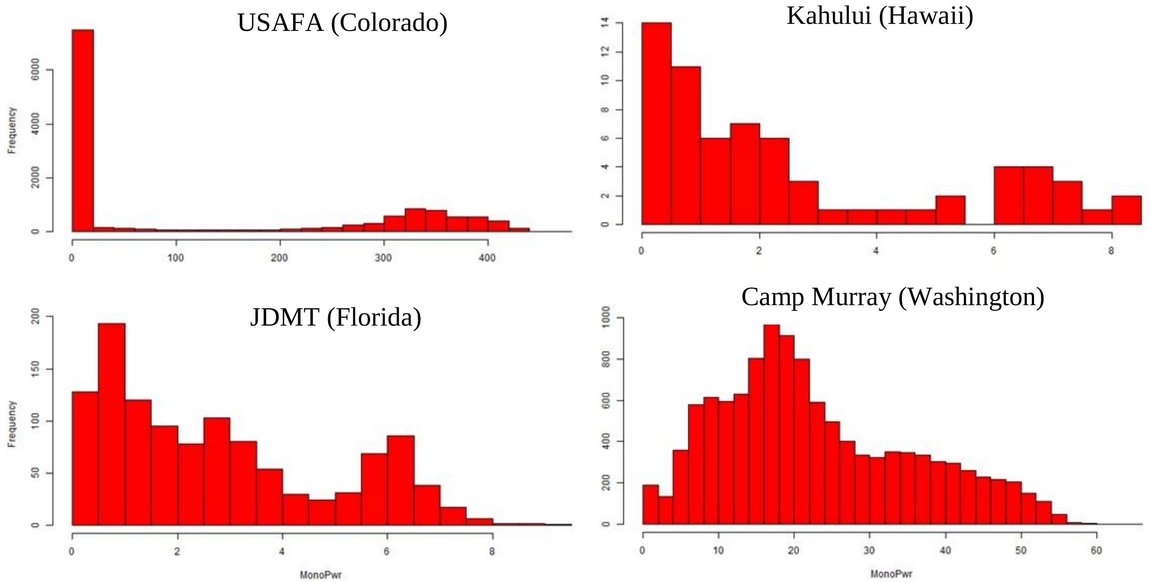


Figure 4. Examples of Monocrystalline power output distributions. USAFA had extremely high measurements (> 400 watts). Kahului and JDMT had unusually low measurements (<10 watts). Camp Murray had power outputs more than double the panel's 25-watt rating.

Each location's time was recorded using Greenwich Mean Time (GMT) and the time column for every location was adjusted to its respective local time. Using the date and time column, hour and month columns were added to the dataset. A time stamp in the following format was also calculated: YYYYMMDDHHMM. This timestamp was needed to add weather variables from the National Oceanic and Atmospheric Administration's (NOAA) datasets, which use the same timestamp. The dataset discussed above will be referred to as the *system dataset* for the remainder of this chapter because all measurements in this dataset were derived from the test systems in the field at each location.

One site is located in the southern hemisphere and was removed from the system dataset. This removal was because the seasons are inverted for the southern hemisphere.

The variation between the two hemispheres could not be accounted for in the models and the scope was limited to the northern hemisphere.

Temperature values unexplainedly jumped from -28.3 degrees Celsius to -39.3 degrees. Furthermore, the temperature will not read temperatures less than -40 degrees Celsius. Once -40 was reached, humidity values and power output values were predominantly recorded as zeros. There was no way of knowing whether or not the zeros were a legitimate measurement for power values less than -39, so those data points were removed.

A few of the locations had an initial period in data collection where the power output was improbably high. This phenomenon occurred for Camp Murray, Curacao, Grissom, Lajes, MNANG, Offutt, and Spangdahlem. An example of this can be seen in a plot of Camp Murray's months versus power output in Figure 5. This plot was built in JMP Pro statistical software [50]. June - October all outputted power values above the rest of the year. Furthermore, the panel is rated for 50 watts and June was outputting power values near 70 watts. Between 3-5 months of data was removed from the sites mentioned above due to high outputs such as the ones produced by Camp Murray.

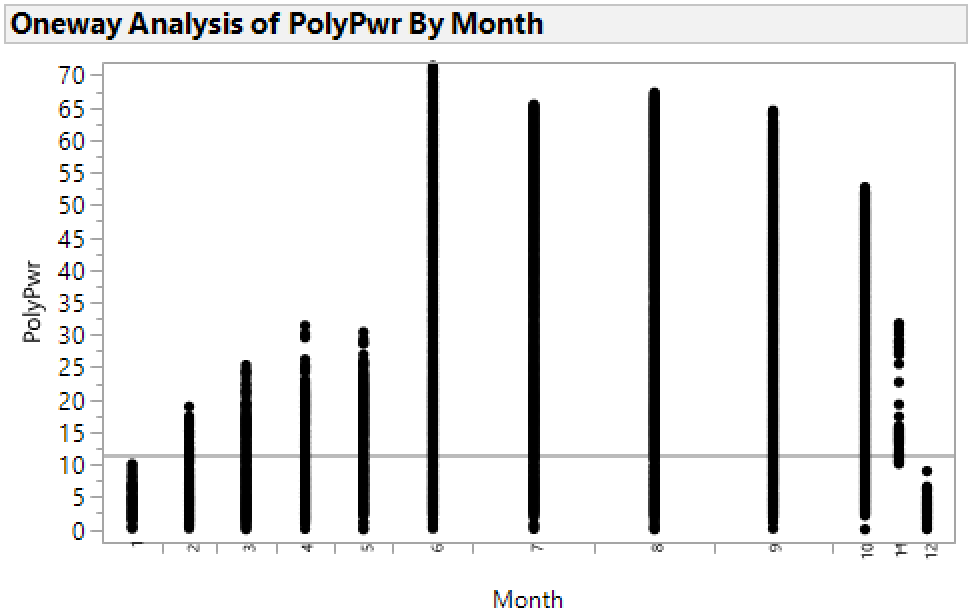


Figure 5. Camp Murray months vs. power output. June – October all showed abnormally high power outputs and datasets in those months were ultimately removed from the final dataset.

For the remaining locations, the nearest available airport weather station was found from the NOAA’s website and a weather dataset containing the following variables was downloaded: wind direction, wind speed, cloud ceiling, visibility, temperature, dewpoint temperature, and atmospheric pressure. Each weather station was within five miles of the test system location. There was no weather station near Curacao and as a result that location was also excluded from the final dataset.

The weather data timestamp column was also recorded in GMT and was adjusted in the same manner the system dataset times were. Each weather station recorded data at different times and time intervals. This resulted in a time mismatch between when the weather stations recorded data and when the test systems recorded data. The programming language known as R was used to combine the system dataset and the weather station data [51]. The timestamps were used to determine the weather data points that were within plus or minus seven minutes of the system data points. For example, if a site in the system dataset recorded on 201802071345 and the closest reading for weather data occurs at 201802071342, then that weather data point will be loaded in. If a weather data point was recorded more than 7 minutes before or after the test system, then no weather variables were matched to that data point. Subsequently, test system data points missing any weather variables from the weather station were removed from the final dataset. This resulted in Spangdahlem having zero weather data points and it was fully removed.

A time window was chosen to ensure that power readings were only occurring when the sun was up at every location. Therefore, only data for times between 1000 and 1545 were used for model building. A wider window was chosen initially (0600 – 1945) that was later tightened due to short winter days in the northern areas.

Furthermore, data points with power output readings less than 1% of the polycrystalline rating were removed. Low solar panel power output may be caused by legitimate events, such as heavy overcast. However, low power output could also be a result of snow cover, excessive dust, low temperatures, electrical shorts, and system malfunction. It could not be reasonably determined whether or not low power values were actual measurements or errors. As a result, power outputs less than 0.25 watts would not be included in predictive modeling. After all exclusions mentioned above, Lajes was left with only 2 data points and was subsequently removed as well. The full dataset was split 50/50 at random, with half of the data being used to develop the model and half of the data used to validate the model. These two datasets will respectively be identified as the *training dataset* and *validation dataset* for the remainder of this study.

All data exclusions, as well as the training and validation dataset split, are listed in Table 5. A tabulated list of the locations in the final dataset, the states each are located in, and their respective latitudes and longitudes can be seen in Table 6. The location of each site whose data was used in the study can be seen in Figure 6. There are two sites in Colorado that are near each other and appear as only one red dot on the map in Figure 6. The remaining 12 sites were all in the northern hemisphere and were predominately located in the continental United States.

Table 5. Data exclusions that led to the final dataset used for modeling.

	Data Points
Initial compiled (16 sites)	528569
Exclusion 1: Southern Hemisphere	-21087
Exclusion 2: Ambient Temperature < = -39	-3802
Exclusion 3: Hours outside of 6 - 19 (10 hours)	-214338
Exclusion 4: Bad months	-38901
Final without weather station data	**250441**
Weather station data added	
Exclusion 1: Datapoints missing weather variables	-193425
Exclusion 2: Hours outside 10 - 15 (8 hours)	-32278
Exclusion 3: Power <= 0.25 Watts	-3691
Exclusion 4: Lajes	-2
Final dataset with weather station variables (12 sites)	**21045**
Training dataset	10522
Validation dataset	10523

Table 6. Final locations, the state each is located in, and their respective latitudes and longitudes.

Data Name	State	Lat	Long
Camp Murray	Washington	47.11	-122.57
Grissom	Indiana	40.67	-86.15
JDMT	Florida	26.98	-80.11
Kahului	Hawaii	20.89	-156.44
Malmstrom	Montana	47.52	-111.18
March	California	33.9	-117.26
MNANG	Minnesota	44.89	-93.2
Offutt	Nebraska	41.13	-95.75
Peterson	Colorado	38.82	-104.71
Hill Weber	Utah	41.15	-111.99
Travis	California	38.16	-121.56
USAFA	Colorado	38.95	-104.83

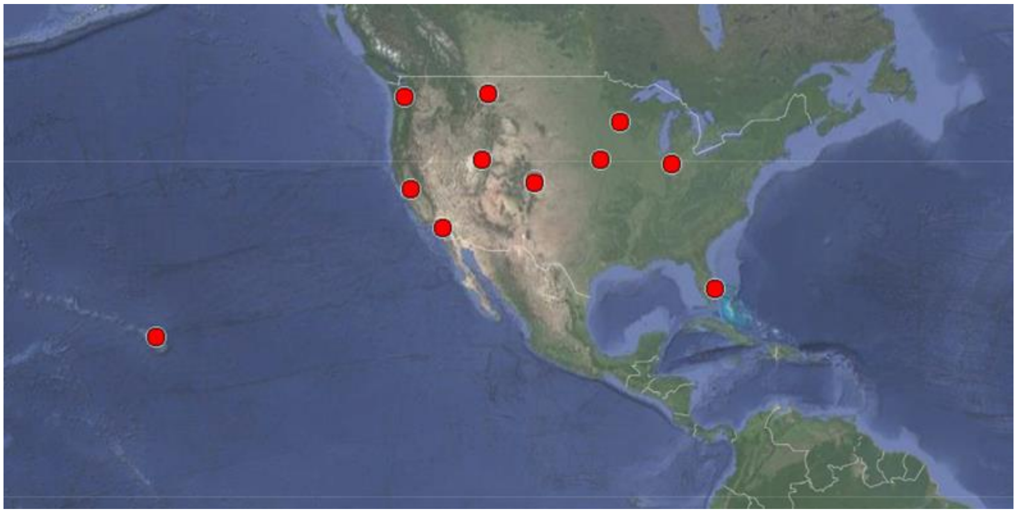


Figure 6. Each site whose data was used in the study is denoted by a red dot on the map. There are two sites in Colorado that are near each other and appear as only one red dot.

Model Development

Two types of models were developed: multivariate linear regression (MLR) and random forest regression (RF). The R programming language was used for model development [51]. RStudio is an R script editor used for the coding process [52]. Within

R, a package known a Rattle was used to initiate the model construction. Rattle stands for the R Analytical Tool To Learn Easily and is an open source graphical user interface (GUI) meant to facilitate data mining in R without requiring extensive knowledge of programming and/or statistics [53]. While the interface is used to analyze data, a log of the R code used in the background of the GUI can be viewed and copied to RStudio. The R commands from the Rattle interface log was put into RStudio and altered by the researcher. Through accessing the log, all the packages and processes used by the Rattle package in forming an RF can be known to the researcher and necessary alterations can be implemented.

Model Input Variable Selection

Based on the information discussed in Chapter II, output variables from the experimental set-up, and available weather variables, ten predictor variable candidates were chosen. A list of the variable and why each was chosen is as follows:

- Latitude: the latitude of each location will dictate the sun deflection angle. In clear sky conditions, the sunlight deflection angle will affect the amount of sunlight the panel receives. This variable controls for the sun angle as it relates to the panel's position on the globe;

- Month: when the sun rises and sets and how high it will appear in the sky at any location on the earth is, in part, determined by the time of year at that location. This phenomenon is what month is meant to control for in the model;

- Hour: the time of day determines how high the sun is in the sky, or whether or not it is present at all. Hour controls for the sun's position in relation to the time of day;

- Humidity: as mentioned previously, humidity can physically affect a solar panel's power output. Humidity can also indicate the possible presence of rain and/or clouds. This variable is controlling for both how the panel is physically affected by humidity and for certain weather phenomenon that may be occurring in the area;

- Temperature: the power output of a panel can be affected by temperature. If the temperature is very high, power output may actually decrease. This effect is meant to be controlled for in the model by including temperature as an input variable;

- Wind speed: the temperature of the panel may be affected by the speed of the wind surrounding the panel, which will subsequently affect the power output of the panel due to power output's dependence on panel temperature;

- Visibility: this variable is a measurement of the distance at which a light can be seen and identified [54]. Visibility will primarily affect how much irradiation reaches the panel and can have a negative effect on power output if visibility is low during daylight hours;

- Pressure: this variable has not been extensively explored in solar panel power output literature. Pressure may have an effect on the panel by indicating a weather occurrence, such as a storm [55];

- Cloud Ceiling: the presence of clouds in the sky above the panel will affect the irradiation that reaches the panel. A cloud ceiling measurement will occur when at least 5/8th of the sky contains clouds [56]. Therefore, this form of measurement also accounts for cloud cover; and

- Altitude: there is less atmosphere for the sun to travel through at locations with higher altitudes. Those locations may be receiving more irradiation on clear days than locations closer to sea level.

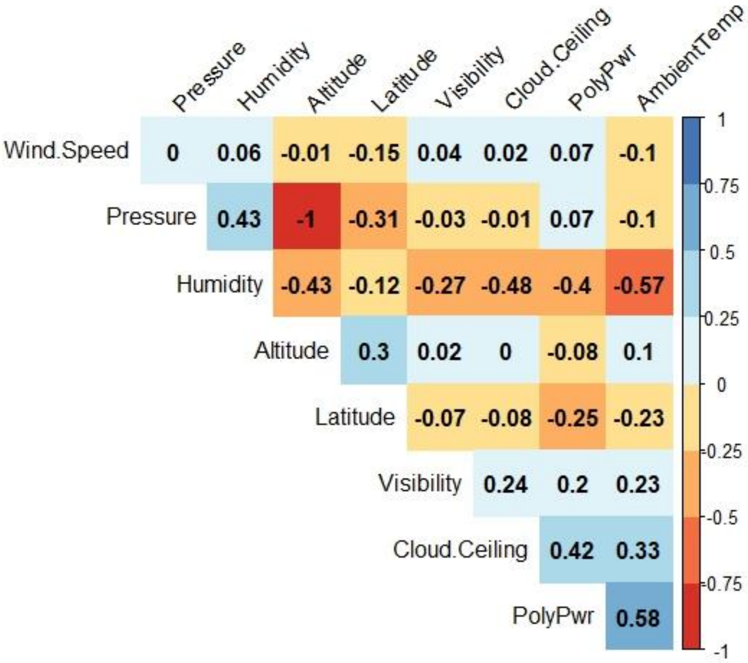
For model simplicity, every available weather variable was not used. Three examples of this are the following variables that could not be justified by theory to be included in the model or were redundant measurements:

- Outdoor Temperature measured at the weather station: each solar panel set-up already has an ambient temperature measurement;

- Dewpoint Temperature: due to the time period chosen for the study, 1000-1500, the dewpoint will not be reached [57]. Including the dewpoint temperature cannot be justified for this model; and

- Wind Direction: the presence of wind speed controls for the effects of wind and wind direction is not necessary for this investigation.

In order to address correlated variables, the correlation of every numeric input variable was checked using R prior to model development. Removing highly correlated variables ensures that the effect of the highly correlated variables is not overemphasized in the model. The only variables with a high correlation coefficient were altitude and pressure with a coefficient of -0.997. Pressure and power output had a correlation coefficient of 0.07. Altitude and power output had a correlation coefficient of -0.08. Out of the two highly correlated variables, pressure was subsequently removed due to its lesser correlation with power output. The correlation coefficients for all numeric input variables are in Figure 7.



	Pressure	Humidity	Altitude	Latitude	Visibility	Cloud.Ceiling	PolyPwr	AmbientTemp
Wind.Speed	0	0.06	-0.01	-0.15	0.04	0.02	0.07	-0.1
Pressure		0.43	-1	-0.31	-0.03	-0.01	0.07	-0.1
Humidity			-0.43	-0.12	-0.27	-0.48	-0.4	-0.57
Altitude				0.3	0.02	0	-0.08	0.1
Latitude					-0.07	-0.08	-0.25	-0.23
Visibility						0.24	0.2	0.23
Cloud.Ceiling							0.42	0.33
PolyPwr								0.58

Figure 7. Correlation coefficients for all numeric input variables. Pressure and altitude and highly negatively correlated.

Multivariate Linear Regression

Data analysis was conducted prior to building the model. Histograms of each variable may indicate how each variable could be treated in the model. Scatterplots of the response variable versus each of the predictor variables can help determine their univariate relationships, as well as each predictor variable against each other [45].

Hypothesized Model

Scatterplots and correlation coefficients of the full dataset were used to understand how each variable may be related to power output. Humidity versus power output was fairly spread-out but may have a negative relationship. As humidity increases, power output decreases. The humidity versus power output scatter plot can be seen in Figure 8. Wind speed, shown in Figure 9, also appeared to have a negative effect on power output. Ambient temperature (Figure 10) and visibility (Figure 11) appeared to each have a positive relationship with power output. As temperature increased and/or visibility improved, power output increased. The relationship between power output and altitude (Figure 12), cloud ceiling (Figure 13), and latitude (Figure 14) were not clear in the plots. The correlation coefficients for every numeric variable can be seen in Figure 7. Altitude’s correlation coefficient revealed a slightly negative correlation of -0.08. Cloud ceiling had a positive correlation coefficient of 0.42; the higher the cloud ceiling, the more power output. Latitude had a correlation coefficient of -0.42 with power output. As the latitude increased (moves north of the equator), power output decreased.

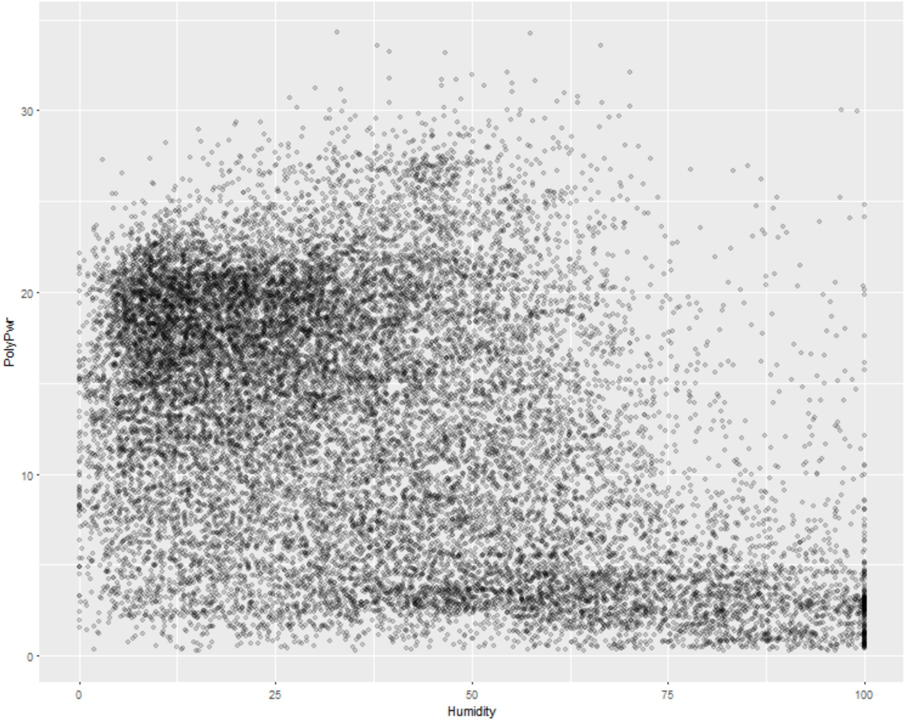


Figure 8. Scatterplot of humidity vs. power output using the full dataset. The plot indicates a possible negative relationship.

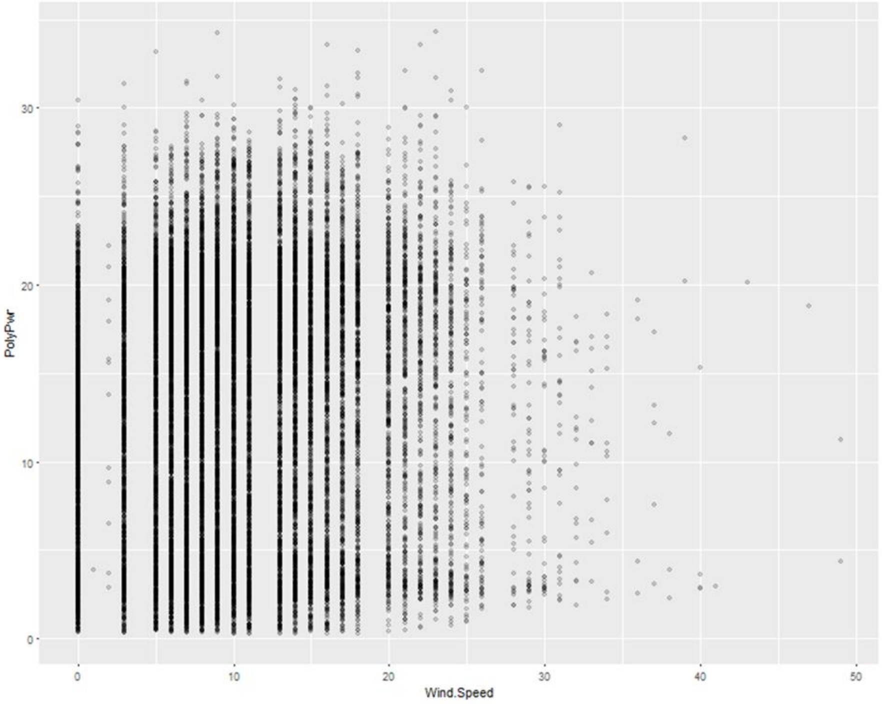


Figure 9. Scatterplot of wind speed vs. power output using the full dataset. The plot indicates a possible negative relationship.

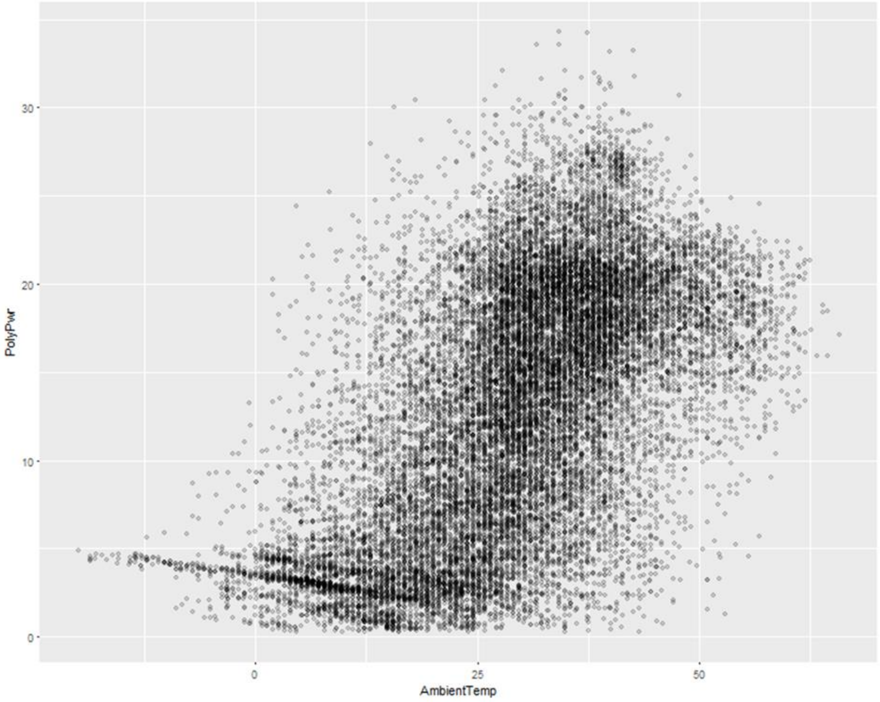


Figure 10. Scatterplot of ambient temperature vs. power output using the full dataset. The plot indicates a positive relationship.

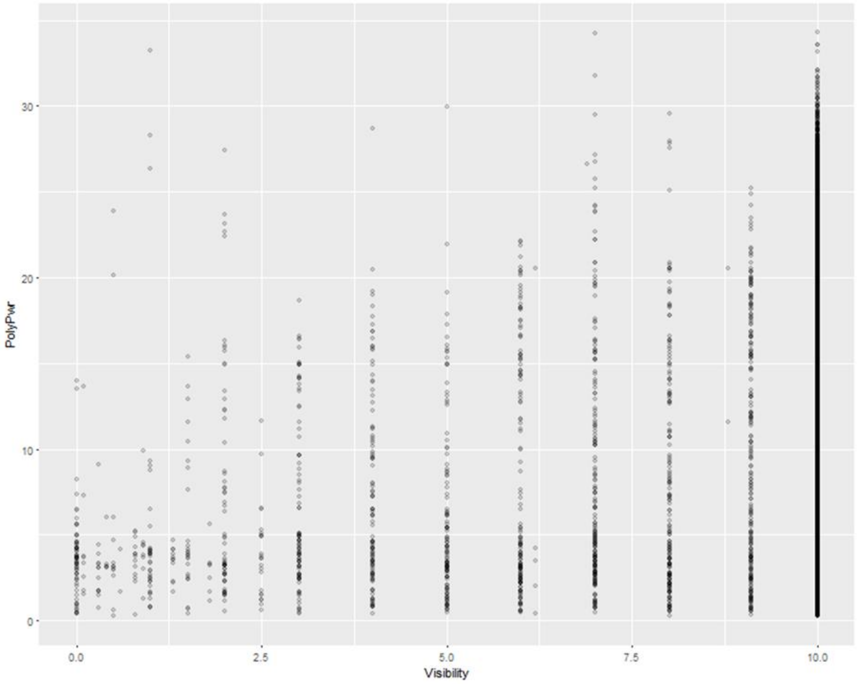


Figure 11. Scatterplot of visibility vs. power output using the full dataset. The plot indicates a positive relationship.

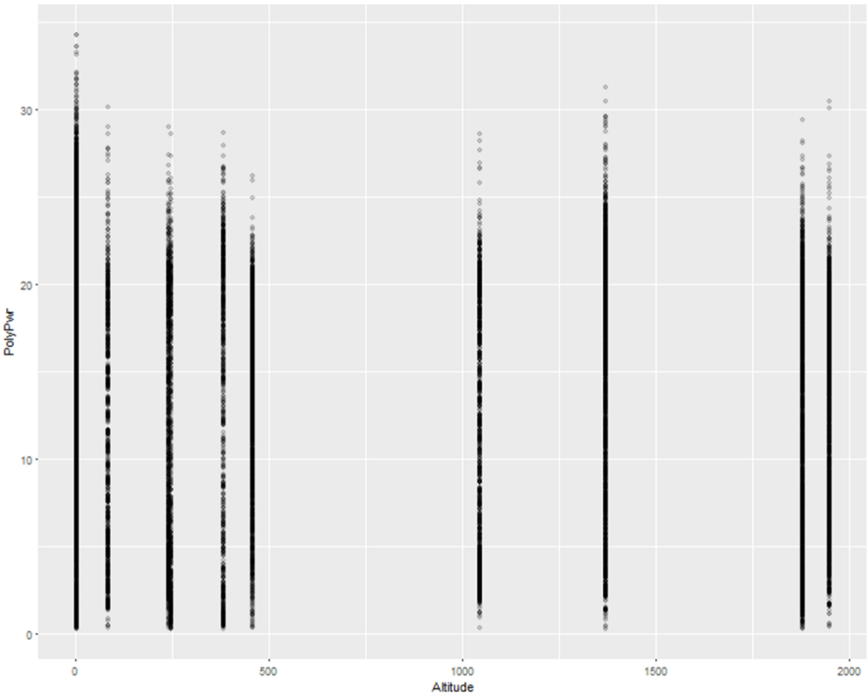


Figure 12. Scatterplot of altitude vs. power output using the full dataset. The relationship between the variables in this plot is not clear.

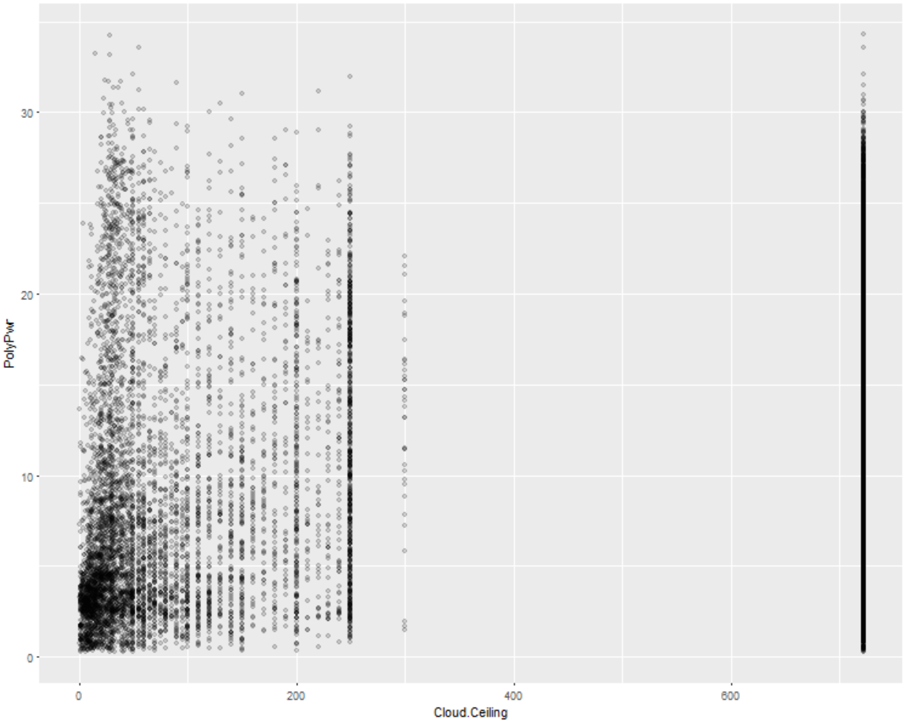


Figure 13. Scatterplot of cloud ceiling vs. power output using the full dataset. The relationship between the variables in this plot is not clear.

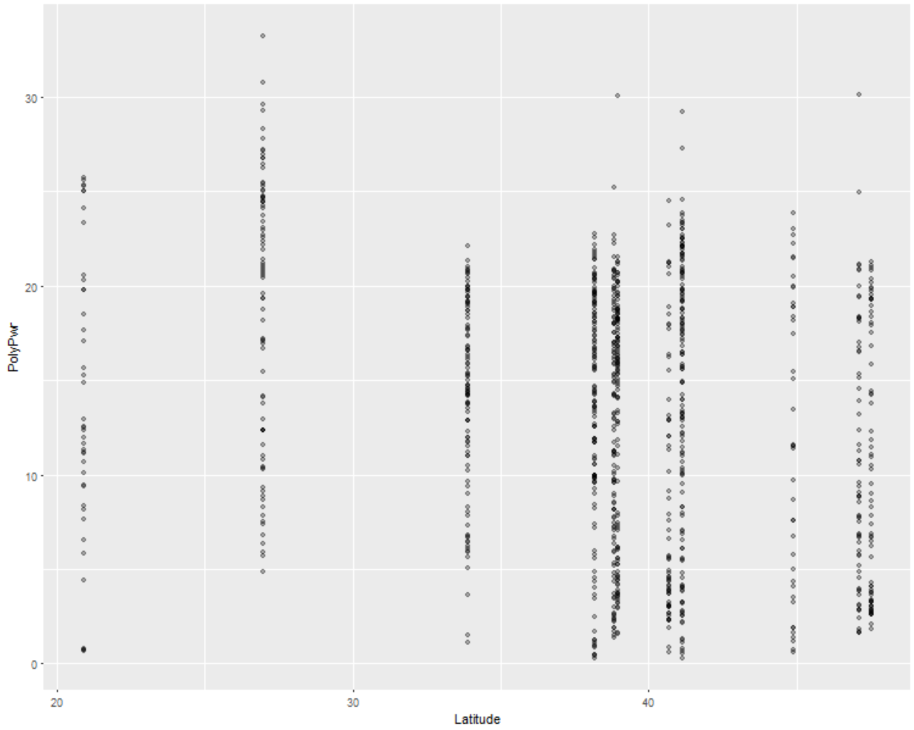


Figure 14. Scatterplot of latitude vs. power output using the full dataset. The relationship between the variables in this plot is not clear.

These relationships were used to develop the hypothesized model in Equation 1. It was hypothesized that the estimated coefficients for latitude, humidity, and altitude would all be negative. Month and Hour were each coded as categorical variables in the model because they are controlling for sun position. It was assumed that sun position did not need to be known for every day or minute, but that each month and hour would suffice. Due to these variables being classified as categorical, they each had their own coefficient associated with every month and hour value. There will be 11 coefficients for month because month 1 will be the baseline for the model and will not be assigned a coefficient. The same is true for hour 10.

$$\text{Power Output} = \beta_o + \beta_1 Lat \pm + Month \pm \beta_{13-17} Hour + \beta_{18} Humidity + \beta_{19} AmbientTemp + \beta_{20} Wind\ Speed + \beta_{21} Visibility + \beta_{22} Cloud\ Ceiling + \beta_{23} Altitude$$

Equation 1. The hypothesized linear model including all initial input variables.

Assumption Checks

Because MLR is a parametric model, key assumptions regarding the residuals of the model must be met in order to apply the MLR to this investigation. The residuals are the error values of the predicted power value versus the actual values, given the input variables. There are a number of assumptions that can be checked qualitatively and/or quantitatively. The first assumption addressed is that of constant variance of the residuals. Qualitative tests for this assumption involve plotting the residuals versus predicted power values and observing whether or not the variance is increasing or decreasing in a systematic manner. If such a pattern is observed, then a quantitative test will be conducted to confirm or dispute the qualitative test. If the residuals appear to be normal, then the Breusch-Pagan will be conducted. The null hypothesis for that test states that the variance of the residuals is constant. A p-value less than 0.05 will reject the null because a significance level of 0.05 was chosen for this model.

Lastly, coerrelation of the residuals must be addressed. This is done by checking the residuals for autocorrelation. If the residuals are correlated, then they are not independent of each other. Autocorrelation is common in time series data similar to the data used in this research. A residual versus time plot is used to evaluate autocorrelation qualitatively and a Durbin-Watson test is used for quantitative assessment. The null hypothesis for the Durbin-Watson test is that autocorrelation of the residuals is zero and a p-value close to zero rejects the null. If autocorrelation of the residuals exists then a Cochrane-Orcutt transformation can be attempted. A Cochrane-Orcutt transformation is meant to correct for first-order autocorrelation [45]. First-order meaning that each residual is only correlated with the residual immediately proceeding it.

If either assumption mentioned above is not met, then the estimated regression coefficients are still unbiased, but may no longer be efficient [45]. By not being efficient the coefficients no longer have minimum variance and the MSE may underestimate the variance of the error terms. The true standard deviation of the estimated regression coefficient will also be underestimated [45]. Furthermore, the t distributions used to determine the significance of the input variables are no longer applicable. In place of the underestimated standard errors outputted from the model, robust standard errors can be used to allow for model inferences that would have otherwise been invalid. Robust standard errors can be found in R and allow for valid t-tests. The robust standard error estimation procedure in R uses the sandwich package and is called the Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation [58].

Normality of the residuals is another assumption that applies to linear models. In this case, however, any deviation from normality will not be a concern because the dataset for this study is considered large [44].

Model Iterations

Model iterations were conducted to meet model assumptions, decrease prediction standard error, or increase variance accounted for. After addressing model assumptions, insignificant variables were removed from the model. A significance level of 0.05 was chosen for this model, which dictated which predictor variables were significant. Then dependent variable transformations and interactions were explored in order to optimize the model's predictability. The finalized model was validated using the validation dataset.

Random Forest Regression

RF is a nonparametric modeling technique and no assumptions were required to be met. The same training dataset used to develop the MLR was used for the RF. Iterations for the RF model involved adjusting two model tuning parameters. The first tuning parameter is how many trees need to be developed in order to reduce prediction error as much as possible. The second tuning parameter, known as mtry, is how many variables will be tried at each node within every tree. Tuning these two parameters will reduce model error and increase the amount of variance the model can account for. Variable importance is an output of the model and can help dictate whether or not certain variables can be removed from the model with little change to model performance. Removing variables that are not helping the model is desired in order to simplify the model. The finalized model was validated using the same validation dataset that was used for the MLR.

Comparing the Two Models

The same training dataset and validation dataset were used for development and validation of the MLR and the RF. Each model began its development with the same initial predictor variables. Within each model's development process, predictor variables may be removed due to insignificance in the model. Therefore, some variables may remain in one model and not in the other. Furthermore, MLR models will often involve variable transformation or interactions while a random forest model will not change its input variables in this manner.

Each model was used to predict the power output of the validation set. Each model's prediction will estimate an R-squared value. This output describes the percentage of variance in the data the model can account for. For instance, if the R-squared value is 0.3, then 30% of variance in the output is accounted for in the model. These R-squared values were compared to decide which model better predicted the power output for the validation dataset.

Sources of Error

The data measurement equipment is a source of error in this study. Sensitive components, such as the computer system, was in a weatherproof case at each site. However, condensation may still get inside the case and cause malfunctions in the system. Additionally, extensive snow coverage and/or extremely cold temperatures may cause data outages. Some sites did not begin reading consistent data until a few months into the investigation and erroneous data from that break-in period may have not been filtered out of the dataset completely. The weather stations were not co-located with the test systems and were up to five miles away. Additionally, a slight measurement error between the weather station data points and the system data points may be present. This error was because some of the measurements at the weather station were recorded up to seven minutes before or after the system data was measured.

IV. RESULTS AND ANALYSIS

Quality of Data

The final dataset used for analysis has 9 input variables and is composed of data from 12 different locations. Each location had a minimum of 780 data points and as many as 2,746. The number of data points for each location can be seen in Table 7.

Table 7. The number of data points each location has in the final dataset. The locations are arranged in descending of data points order from left to right.

Travis	Peterson	USAFA	Hill Weber	March AFB	JDMT	Malm-strom	Grissom	Camp Murray	Kahului	Offutt	MNANG
2746	2640	2573	2384	2204	1779	1517	1487	1113	941	881	780

Descriptive statistics for each variable are shown in Table 8. Visibility and cloud cover are predominantly reporting clear skies (722) and high visibility (10). This large number of repeated values could lead to the variable being insignificant in either model.

Table 8. Descriptive statistics for each quantitative variable.

	Power Output	Latitude	Humidity	Temp	Wind Speed	Visibility	Cloud Cover	Altitude
Units	watts	degrees	percent	Celsius	mph	miles	100 feet	meters
Min.	0.26	20.89	0	-19.98	0	0	0	1
1st Qu.	6.40	38.16	17.53	21.92	6	10	140	2
Median	13.80	38.95	33.12	30.29	9	10	722	458
Mean	12.98	38.12	37.12	29.29	10.32	9.7	516	798.8
3rd Qu.	18.86	41.15	52.59	37.47	14	10	722	1370
Max.	34.29	47.52	99.99	65.74	49	10	722	1947

Data point exclusions affected the number of data points for each hour and month. The earliest hour in the timeframe, hour 10, has the least amount of data points, as can be seen on the bar chart in Figure 15. The amount of data points for each month appears to be affected the most by data point exclusions. October (month 10) has the least amount of data points with 903 while July (month 7) had the most with 2,929 data points. It is not known why October had the least data points. However, it may be expected for fall and winter months to have the least due to data errors associated with cold temperature, snow cover, and system dropouts. The bar chart for months can be seen in Figure 16.

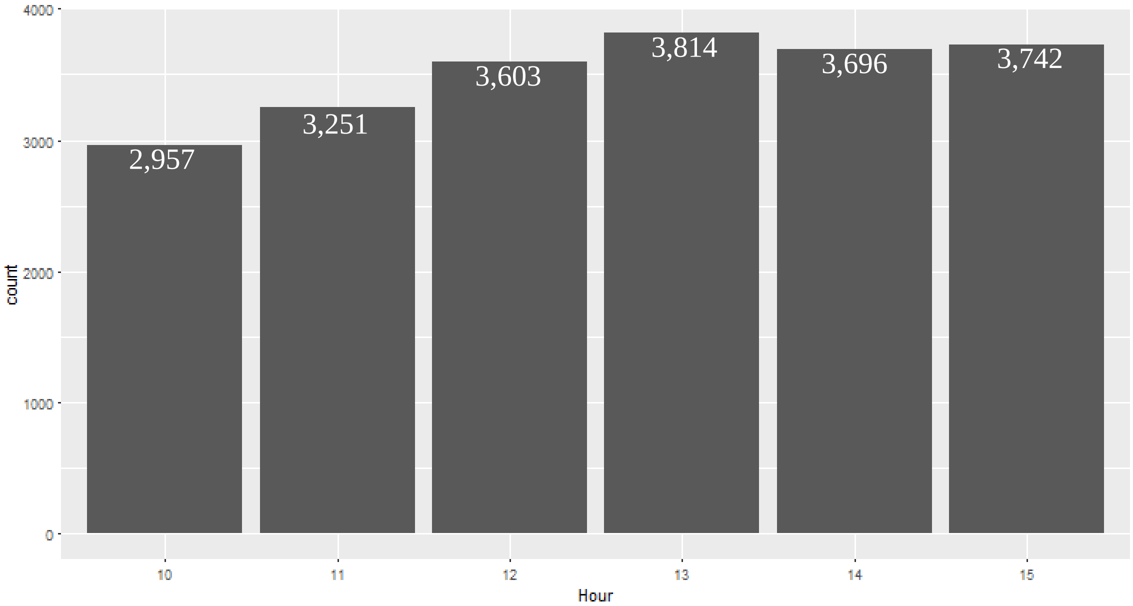


Figure 15. Bar chart of hours within the timeframe set for the analysis. The hours are shown in succession from left to right. The dataset count is shown near the top of each bar.

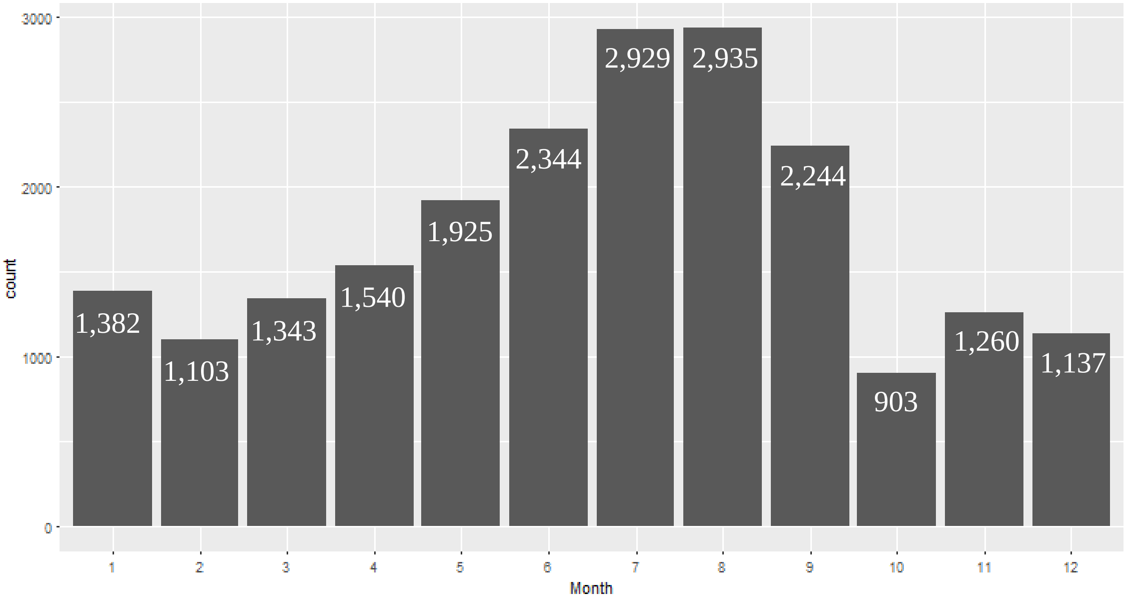


Figure 16. Bar chart of months. October has the least amount of data points and July has the highest number of data points.

Modeling

Linear Regression Model

The hypothesized model was run in R using the stats package [59]. The initial R-squared value for the hypothesized model was 0.525. The residual plots of the hypothesized model indicated that the residual variance is not constant. According to the residual versus predicted plot in Figure 17, as the predicted power values increased, so did the variance of the residuals. The constant variance assumption had to be addressed before optimizing the model prediction.

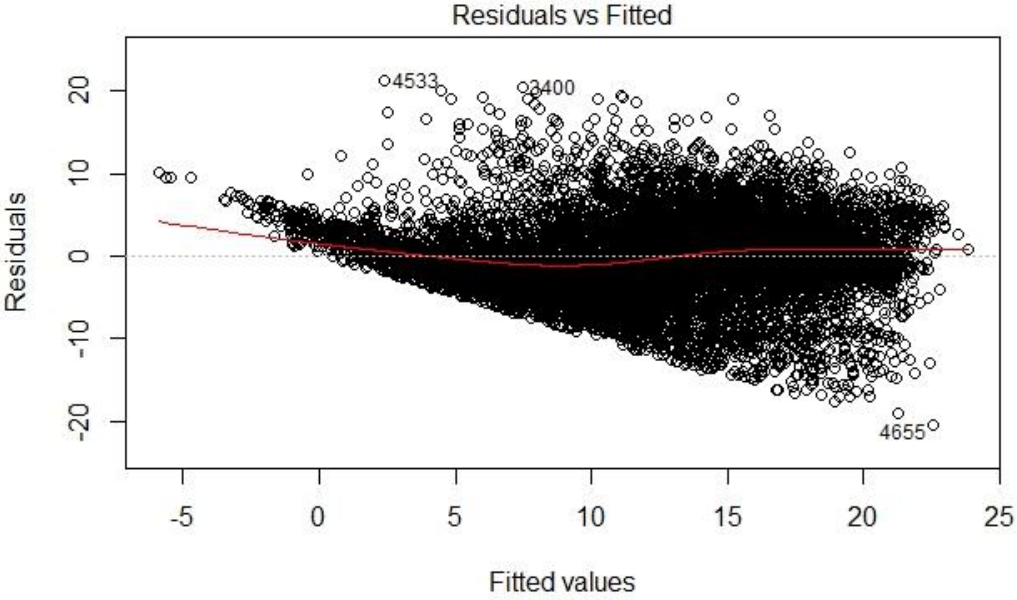


Figure 17. Residual versus fitted plot for the hypothesized MLR model. Non-constant variance of the residuals can be seen in this plot.

In order to confirm what is being observed in the residual plots, a Breusch-Pagan test was conducted in R using the lmtest package [60]. The null hypothesis for this test states that the variance of the residuals is constant. The p-value for the Breusch-Pagan test for this model was < 2.2e-16. This p-value is below 0.05 so the null hypothesis is rejected and the variance of the residuals is non-constant.

One option used to remedy non-constant variance is transforming the dependent variable, power output. Three transforms were considered based on their normalizing influence on the distribution of power output. The distribution of power output can be seen in Figure 18. Power output is slightly bimodal and appears to be right-skewed. The distribution of each transform applied to the power output are also presented in the titles of Figure 18. The first transform was logarithmic. This transform made the distribution unimodal but caused a left skew. Taking the square root of power output moved the distribution closer to normal with less of a left skew than a logarithmic transform. The third transform was decided using a Box-Cox transformation method that estimates which power value would be the most effective in stabilizing the residuals [45]. The Box-Cox method was performed in R using the caret package [61]. The power value was found to be 0.7. Each of these transformations was considered because they each transformed the power output distributions differently.

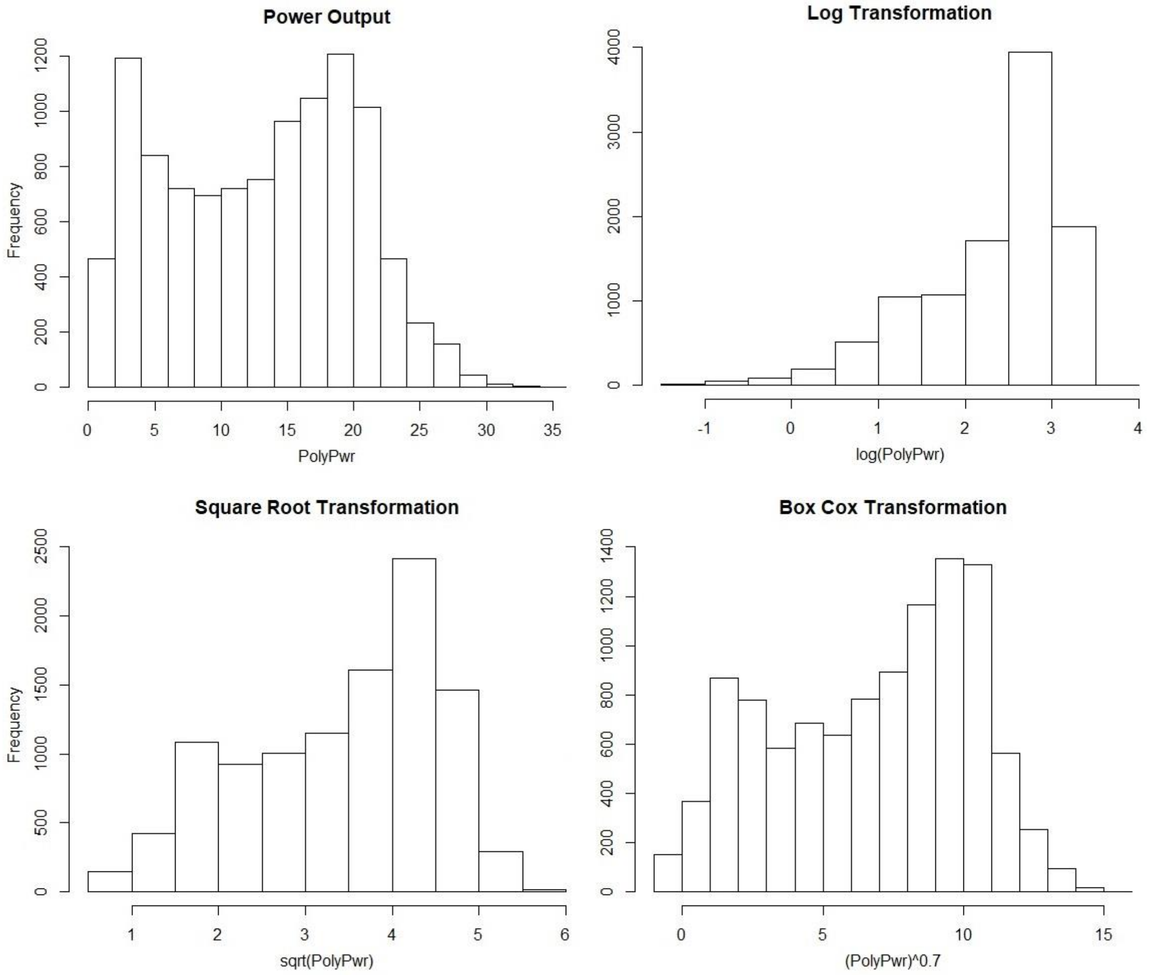


Figure 18. Distributions of power output untransformed and power output with three transformations applied.

All three of the residuals versus fitted plots were similar to the hypothesized model. It was evident that non-constant variance was still present in each of the transformed models. Shown by the residuals versus fitted plots in Figure 19, the logarithmic transform over-corrected and the two other transforms still showed increased variance for increasing fitted values.

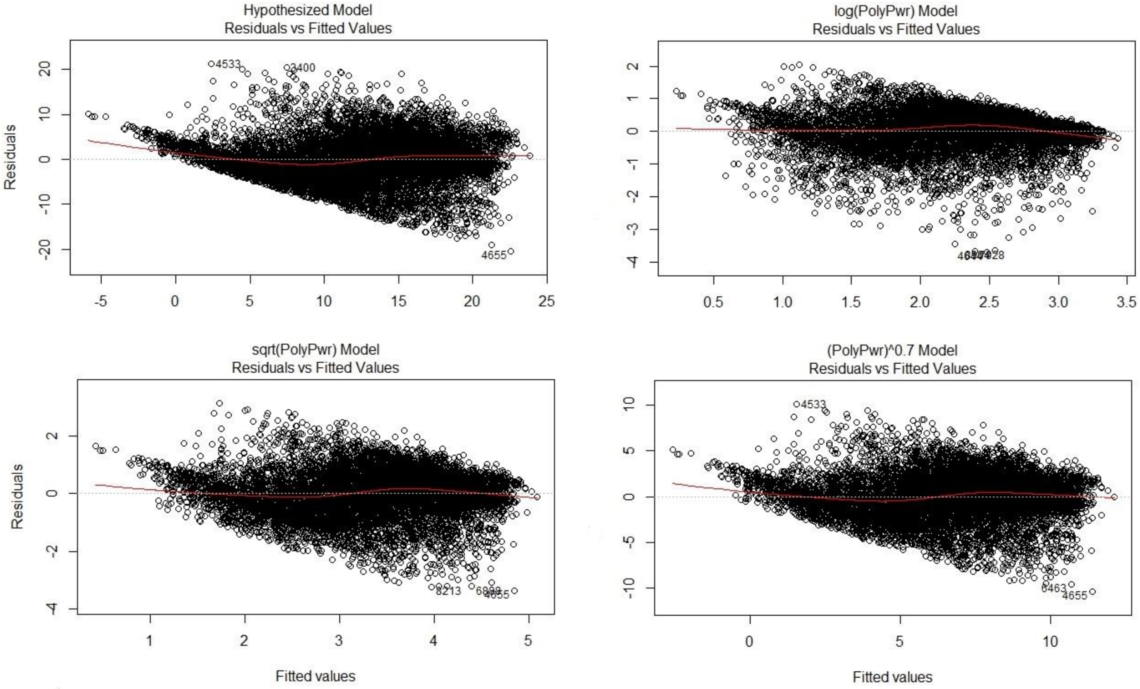


Figure 19. Hypothesized model residual versus fitted values plots for all transformed and untransformed power outputs.

Correcting for non-constant variance by transformation proved ineffective. Robust standard errors will have to be applied before making any inferences. The effects of non-constant variance will be addressed after checking the rest of the assumptions.

Autocorrelation of the residuals was assessed next. Evidence of autocorrelation can be seen in the residual versus index plot in Figure 20. This plot was produced in R using the ExPanDar package [62]. The data points are indexed by location alphabetically. Within each location, the data points are ordered by date/time from earliest to latest. For example, all data points for the location *Camp Murray* appear first in the index and correspond to X=1 to X=1113 and all Camp Murray data points are ordered by date/time. Following Camp Murray are all data points for Grissom, also ordered by date/time.

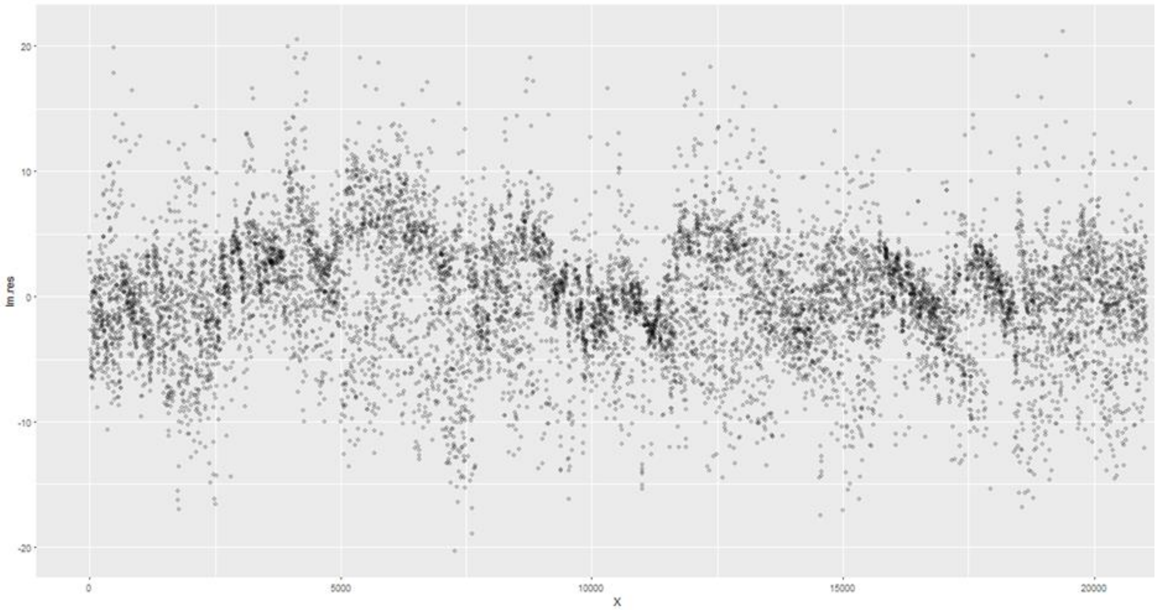


Figure 20. Hypothesized model residual vs. index plot. A pattern can be seen, indicating autocorrelation of the residuals.

The possibility of autocorrelation was confirmed using the Durbin-Watson (DW) test in R with the lmtest package [60]. The null hypothesis for the DW test states that the residual autocorrelation is zero. The alternative hypothesis states that the autocorrelation is greater than zero. The p-value for the Durbin-Watson test was < 2.2e-16. This value is close to zero and the null hypothesized is rejected. The autocorrelation of the residuals is not zero.

One cause of autocorrelation in the residuals is the absence of one or more key predictor variables that have time-ordered effects on power output [45]. Given the scope of this research, additional predictor variables could not be added to the model. Another remedial measure is to transform the model variables using a procedure known as Cochrane-Orcutt. The procedure will only work for first-order autoregression; each residual is correlated with the residual immediately proceeding it [45]. The autocorrelation for this model is more complicated than a first-order autoregression. Autocorrelation is present in the residuals, but the exact structure of the data causing the autocorrelation is unknown. Therefore, the autocorrelation cannot be remedied using Cochrane-Orcutt.

Because the model has failed two assumptions, the model outputs must be adjusted to account for both the non-constant variance and autocorrelation of the residuals. Once the t-tests must be assessed, the new standard errors will be used. In the meantime, model formulation continued.

Although non-constant variance could not be remedied by transforming power output, the model adjusted R-squared value can be increased using a y-transformation. Adjusted R-squared takes into account the number of variables in the model. The square-root transformation increased the hypothesized model adjusted R-squared value by 0.0117. Table 9 shows the adjusted R-squared value of each transformation.

Table 9. Adjusted R-squared values are the hypothesized model and each of the transformed models.

	PolyPwr	log(PolyPwr)	sqrt(PolyPwr)	(PolyPwr) $^{0.7}$
Adjusted R-Squared	0.525	0.5041	0.5367	0.5356

Once an appropriate Y-transform was applied, the significance of the variables must be assessed in order to determine if each predictor should remain in the model. However, the t-statistic of each variable is not valid because the residuals have non-constant variance and are autocorrelated. Robust standard errors were used to perform a t-test of the coefficient in order to determine which variables are significant in determining the power output. The robust standard error estimation was performed in R and does not change the coefficient estimates, just the standard errors and the t-test p-values. From the robust standard error, visibility was shown to have a p-value of more than 0.05; the significance level chosen for this study. The t-test null hypothesis states that the relationship between the dependent and independent variables is zero. When the p-value is above 0.05, then the null hypothesis is failed to be rejected and there is not enough evidence to conclude that a non-zero relationship exists. As a result of the hypothesis test visibility was removed from the model and the model was re-fit.

With insignificant variables now removed from the model, independent variable transformations were explored. Transforming the independent variables can increase the adjusted R-squared value. The scatter plots of each dependent variable versus the square-root of the power output revealed possible transformations. By transforming a dependent variable, the relationship between the independent and the dependent variable may become more linear and increase the adjusted R-squared. Humidity showed a tighter grouping in its scatter plot when a square root transformation was applied. The scatterplots of humidity against the square root of power output before and after the transformation can be seen in Figure 21 and Figure 22, respectively. Transforming humidity caused the R-squared to decreased by 0.011 and error increased by 0.009. As a result, humidity remained untransformed for the remainder of model optimization.

Input variable transformations were also pursued in an effort to further optimize the model. In the scatterplot of humidity versus ambient temperature, seen in Figure 23, the two variables appeared to be related. This was also confirmed with their correlation of -0.57 in Figure 7. As a result, ambient temperature and humidity were interacted and this improved the model. R-squared increased by 0.0031 and error decreased by 0.0028.

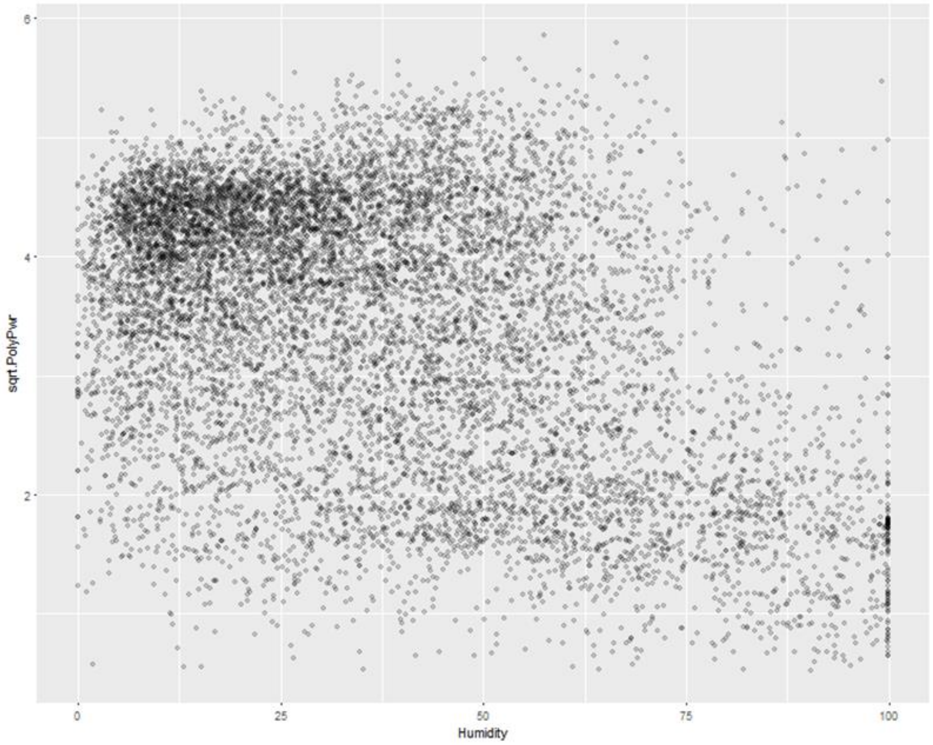


Figure 21. Humidity vs. sqrt(PolyPwr).

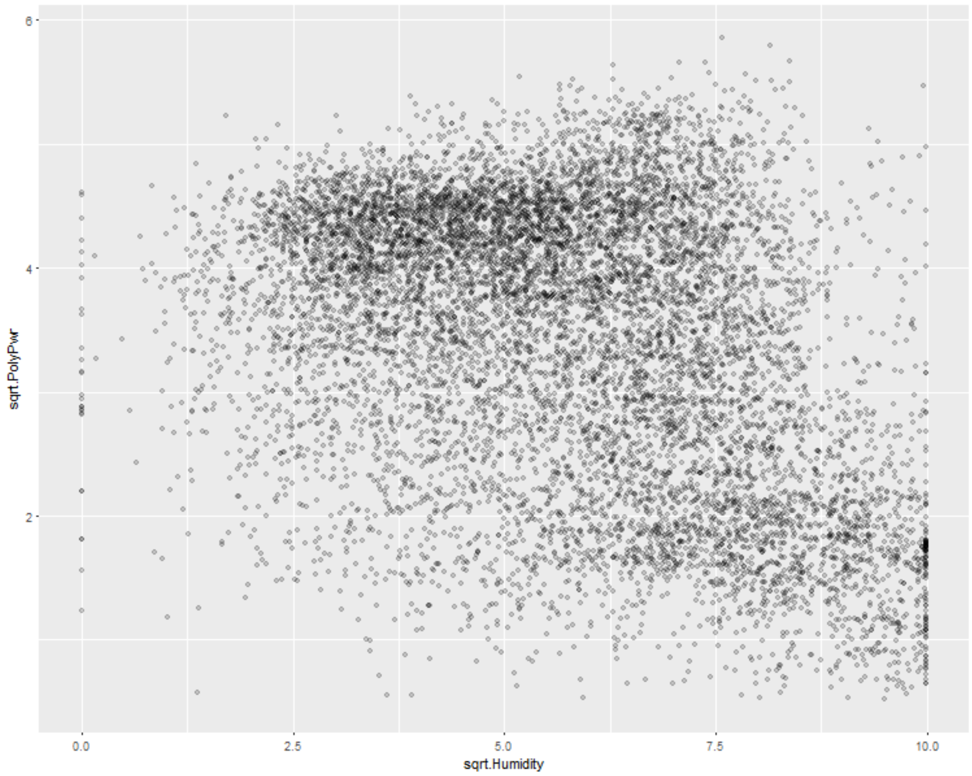


Figure 22. Sqrt(humidity) vs. sqrt(PolyPwr). The data points appear to have moved slightly closer together and the relationship may be more linear than before humidity was transformed.

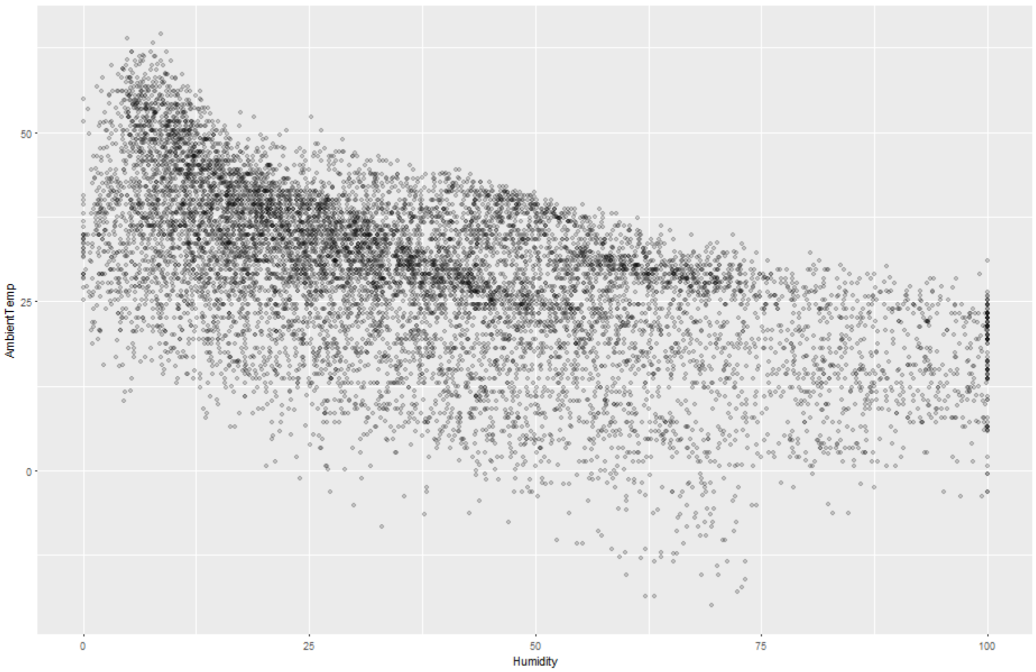


Figure 23. Scatterplot of humidity vs. ambient temperature. There appears to be a slightly negative relationship between the two variables.

Latitude and altitude are both variables meant to control for the amount of irradiance the panel receives. Latitude is controlling for the angle of the sun in relation to each location, which affects how much atmosphere the sun's energy must travel through before reaching the ground. Altitude also controls for how much atmosphere the sun's energy must go through based on the fact that higher altitude locations have less atmosphere. When these two terms are interacted, the model improves by an R-squared increase of 0.0118 and an error decrease of 0.01.

This interaction marked the final manipulation of the model. The final model R-squared value was 0.5516, meaning 55.16% of the variance in the power output was accounted for in the model. Each coefficient estimate, its standard error, and its robust standard error can be seen in Table 10. The percent difference between the standard errors and the robust standard errors are also included in Table 10. Many robust standard errors increased from the original standard errors in order to account for the non-constant variance and autocorrelation of the residuals. The standard errors for all quantitative variables increased by at least 50% and as much as 82.9%. This change signifies how much of an effect the non-constant variance and autocorrelation has on the estimation of the population distribution.

Table 10. Coefficient estimates, standard errors, robust standard errors, and the percent difference between the standard errors. Standard error increased for all quantitative variables by at least 50% and as much as 82.9%.

Coefficient	Estimate	Standard Error	Robust Std. Error	% Diff.
Intercept	3.6607	0.1021	0.1693	65.8
Latitude	-0.0393	0.0018	0.0019	5.6
Month 2	0.2986	0.0429	0.0787	83.4
Month 3	0.6587	0.0404	0.0802	98.5
Month 4	0.7559	0.0394	0.0772	95.9
Month 5	0.7233	0.0404	0.0749	85.4
Month 6	0.7249	0.0407	0.0708	74.0
Month 7	0.5993	0.0412	0.0733	77.9
Month 8	0.5029	0.0407	0.0723	77.6
Month 9	0.2845	0.0402	0.0710	76.6
Month 10	0.1529	0.0468	0.0775	65.6
Month 11	-0.0573	0.0414	0.0742	79.2
Month 12	-0.2229	0.0422	0.0765	81.3
Hour 11	0.2845	0.0270	0.0239	-11.5
Hour 12	0.4109	0.0270	0.0271	0.4
Hour 13	0.4172	0.0271	0.0293	8.1
Hour 14	0.3169	0.0275	0.0308	12.0
Hour 15	0.0573	0.0274	0.0315	15.0
Humidity	-0.0162	0.0008	0.0008	0.0
Ambient Temp	0.0197	0.0014	0.0012	-14.3
Wind Speed	0.0057	0.0012	0.0018	50.0
Cloud Ceiling	0.0009	0.00003	0.00005	66.7
Altitude	-0.0025	0.0001	0.0001	0.0
Ambient Temp * Humidity	0.0002	0.00003	0.00005	66.7
Latitude*Altitude	0.00006	0.000004	0.000006	50.0

The hypothesized model began with 9 input variables and the final model eliminated one, visibility. Power output was transformed and two sets of variables were interacted with each other in the final model. Due to these changes, model interpretation is no longer straight forward. As an example, this is an explanation of the relationship ambient temperature and humidity have with power output in the final model: with all else held constant, for every unit increase of both ambient temperature and humidity, the square root of the power output will increase by 0.00023 watts$^{0.5}$. The model does not have any meaningful interpretation regarding its relationships due to how much the model had to be manipulated to produce better predictability. However, a comparison between this model and the RF model can be conducted. Also, the general relationships in the model can be discussed.

Temperature, wind speed, and cloud ceiling each have a positive relationship with power output. Latitude, humidity and altitude each have a negative relationship with power output. As temperature and humidity both increase, power output follows. For locations with higher altitude and at higher latitude, power output increases. Months 11 and 12 (November and December) were the only months with negative coefficients in the model. This may indicate that these months produced the least amount of power output than the rest of the year.

The final MLR model was validated using the validation dataset. The MLR model was used to predict the square root of the power output due to the y-transformation that was applied. The resulting R-squared value describes how well the model predicted the power output of the validation dataset. The R-squared value of the validation was 0.5620. This means that the model was able to account for 56.20% of the variance in the validation dataset. The R-squared of the validation was 1.04% higher than the final model's R-squared value. Figure 24 shows the model validation predicted versus observed plot. The solid red line is the fit line for the prediction. The dashed red line represents a theoretical perfect-fit with an R-squared value of 1.0. Predicted values in the top left and bottom right of the plot pulled the fit line away from the theoretical fit.

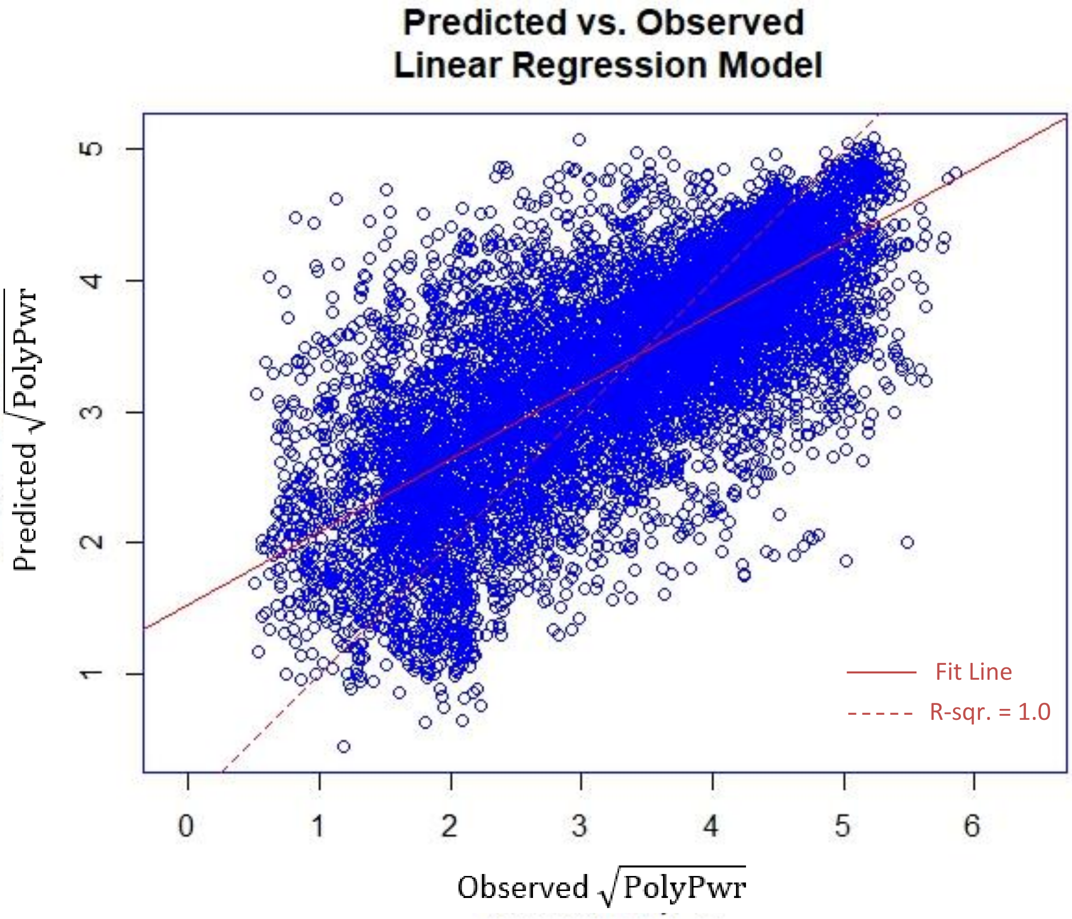


Figure 24. Linear regression validation predicted vs. observed plot. The solid red line represents the fit line. The dashed red line represents the fit line if the model R-squared value was 1.0.

Random Forest Model

The initial RF model included the same input variables as the hypothesized MLR. Month and hour were also coded as categorical variables. The initial model was built using 500 trees and an mtry of 3. A forest of 500 trees sufficiently decreased model error as shown in the number of trees versus error rate plot in Figure 25. Adding more than 500 trees would not reduce the error any further.

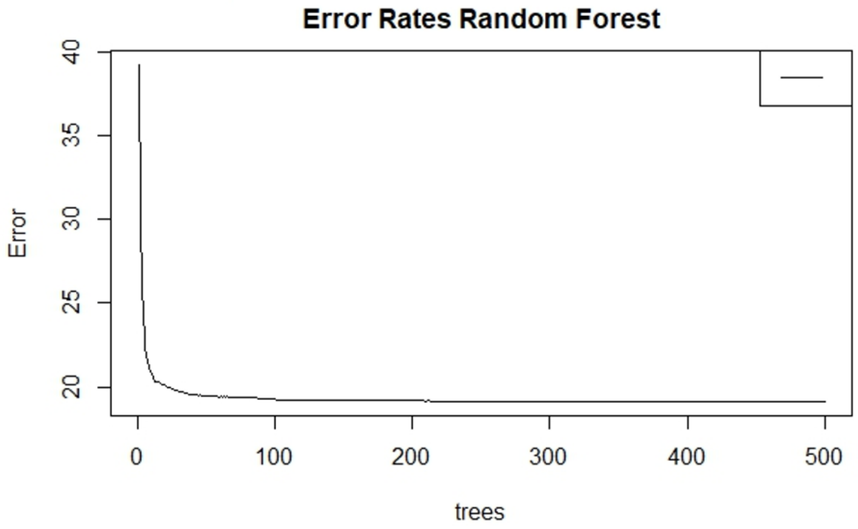


Figure 25. The number of trees versus error plot. 500 trees are shown to sufficiently reduce the error of this model. Adding more than 500 trees would not significantly reduce the error.

Other mtry values were tried in order to reduce model error. Each mtry value and the associated model mean squared error (MSE) is shown in

Table 11. MSE was minimized when mtry was 3. Model analysis continued with 500 trees and an mtry of 3.

Table 11. Model MSE for mtry from 1-5. An mtry of 3 had the highest reduction in MSE.

mtry	1	2	3	4	5
MSE	20.56	17.58	17.54	17.62	17.76

Variable importance is listed in Table 12. Variable importance was measured by how much each variable decreased the model MSE. This measure is found by calculating the MSE of the portion of the data not used to build the decision tree. The MSE is calculated again after permuting each predictor variable. Then each MSE are subtracted from each other. This is done for every tree in the forest. The average change in MSE for each predictor variable permutation is found for the whole forest. By this measurement, ambient temperature is the most important variable and visibility is the least important. Ambient temperature decreased the prediction MSE by 23.87% when it was not permuted. Visibility decreased the prediction MSE by 0.45%

Table 12. Variable importance ranked in descending order by reduction in MSE.

Variable	MSE Reduction
Ambient Temp	23.87
Month	15.72
Humidity	11.91
Cloud Ceiling	11.83
Latitude	9.7
Altitude	7.41
Hour	5.59
Wind Speed	1.63
Visibility	0.45

Another way to rank the variables is to divide the MSE reduction by the variable's standard error. By scaling the values, the ranking takes the error of each variable into account and may change each variable's relative importance. The scaled rankings can be seen in Table 13. Cloud ceiling is now the most important variable in reducing model error and visibility is still the least important. This raking is by relative importance and each individual variable may remain quintessential in reducing model error.

Table 13. Variable importance scaled by dividing the MSE reduction by each variable's standard error.

Variable	Scaled Importance
Cloud Ceiling	132.57
Month	131.34
Hour	110.34
Ambient Temp	69.76
Humidity	80.49
Latitude	59.23
Altitude	54.79
Wind Speed	43.52
Visibility	24.14

How visibility affects the full model MSE can be determined by re-running the model without it and comparing the MSE to the model with it included. Removing visibility from the model increased MSE from 17.56 to 17.60. That increase in error is small and for the purposes for model simplicity, visibility was dropped from the final model. Removing the next lowest ranking variable, wind speed, was also considered. However, wind speed was ultimately left in the model due to its 0.3 error increase when removed.

The final model was developed using 8 of the initial 9 input variables: latitude, humidity, ambient temp, month, hour, wind speed, cloud ceiling, and altitude. The number of variables tried at each node was 3 and the number of trees in the forest was

500. MSE was 17.60 and the variance accounted for was 65.34%. Variable importance was recalculated after removing visibility and can be seen in Table 14. The scaled variables importance rankings can be seen in Table 15.

Table 14. Final model variable importance ranked in descending order from the top.

Variable	MSE Reduction
Ambient Temp	24.7
Month	16.01
Humidity	12.37
Cloud Ceiling	11.01
Latitude	10.06
Altitude	7.29
Hour	5.66
Wind Speed	1.63

Table 15. Final model scaled variable importance ranked in descending order from the top.

Variable	Scaled Importance
Month	133.74
Cloud Ceiling	133.43
Hour	112
Ambient Temp	95.95
Humidity	79.89
Latitude	63.17
Altitude	56.68
Wind Speed	45.71

The top three variables most important in reducing RF model MSE were ambient temperature, month, and humidity. When the variable importance is scaled, the top three most important variables were month, cloud ceiling, and hour. Cloud ceiling also takes cloud cover into account, and is measured when at least 5/8 of the sky view above the weather station is covered by clouds – cloud ceiling data is presented in Table 4. Therefore, cloud ceiling has an effect on how much sun can reach a solar panel and that effect is realized in the RF model. Month and hour were each included in the model in order to account for seasonal weather changes and the sun's position in the sky. The three least important variables in reducing MSE were latitude, altitude, and wind speed. Latitude is known to affect the angle at which the sun's energy reaches the surface of the earth and, in turn, the amount of irradiation that reaches a horizontal solar panel [23], [62-65]. This variable may be ranked low because of the limited latitudes included in the model. Latitudes in the northern hemisphere range from 0 degrees (the equator) to 66 degrees (the start of the arctic circle). The latitude range for this study was from 20.89 degrees to 47.52 degrees. That is about 1/3 of the total latitude band from the equator to the arctic circle. Subsequently, the effect that latitude may have on power output may not be strong for this study's range of latitudes.

The final RF model was validated using the same dataset as the MLR model. The RF model was used to predict the power output of the validation dataset. With an R-squared value of 0.6580, the model was able to account for 65.80% of the variance in the validation dataset. The R-squared of the validation was 0.46% higher than the final model's R-squared value. Figure 26 shows the model validation predicted versus observed plot. The solid red line is the fit line for the prediction. The dashed red line represents a theoretical perfect-fit with an R-squared value of 1.0. Similar to what happened with the MLR, the predicted values in the top left and bottom right of the plot pulled the fit line away from the theoretical fit.

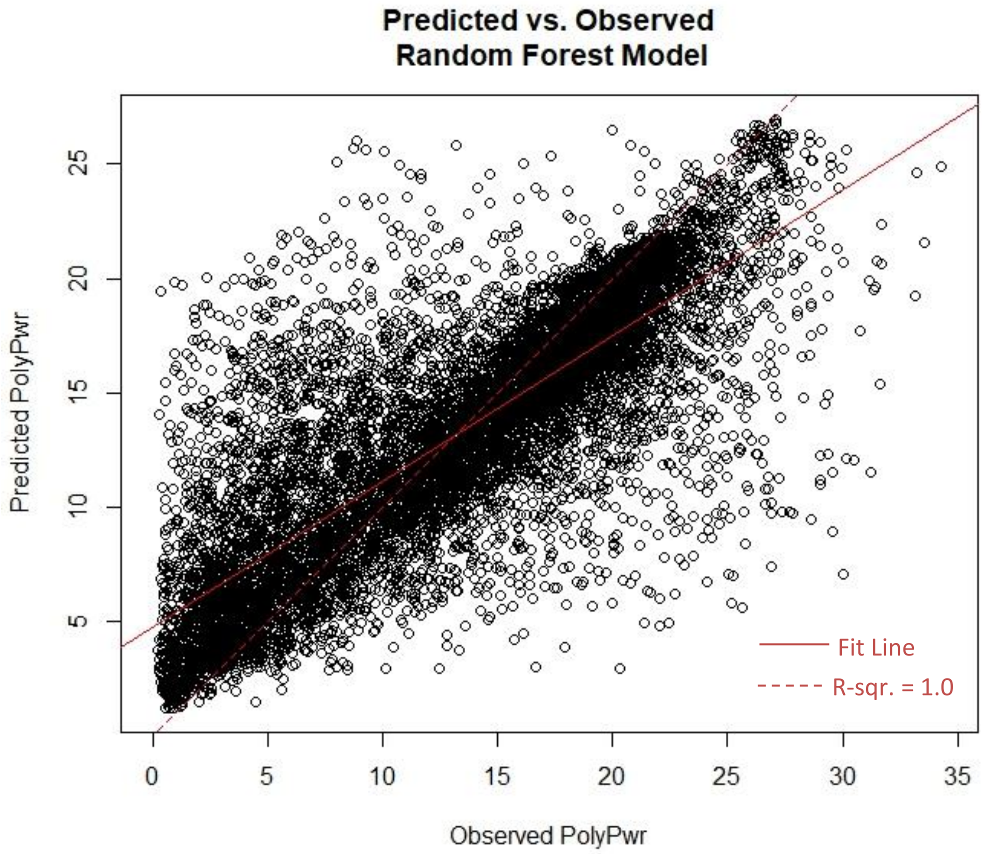


Figure 26. Random forest validation predicted vs. observed plot. The solid red line represents the fit line. The dashed red line represents the fit line if the model R-squared value was 1.0.

Model Comparison

Following the MLR model iterations and RF model tuning, each model used the same 8 predictor variables to predict power output. The MLR model accounted for 56.20% of the variance in the power output of the validation dataset. The RF model accounted for 9.6% more variance, outperforming the MLR. When each model was used to predict the power output for every location, the RF model was able to account for as much 13.24 % more variance than the MLR model. Travis Air Force base in California had the highest predictability using the RF model with 78.91% variance accounted for. The location with the lowest predictability was Kahului airport with 39.32% variance accounted for. Every locations predictability using each model can be seen in Table 16.

Table 16. Each locations predictability using each model as well as the difference in percent variance accounted for.

	Camp Murray	Grissom	Hill Weber	JDMT	Kahului	Malm-strom	March AFB	MNANG	Offutt	Peterson	Travis	USAFA
RF	74.99	64.87	67.35	49.53	39.32	62.94	60.03	71.08	56.60	60.74	78.91	48.87
MLR	73.73	51.99	65.07	38.10	26.08	55.43	53.06	63.18	54.27	50.90	74.41	38.46
Difference	1.26	12.88	2.28	11.43	13.24	7.51	6.97	7.90	2.33	9.84	4.50	10.41

An example of a 3-day prediction summary for Travis Air Force base using each model can be seen in Figure 27. This prediction took place from 1000 on July 2nd, 2018 to 1500 July 4th, 2018. The RF model slightly overpredicted on July 2nd, was close to the observed power outputs on July 3rd, and slightly underpredicted on July 4th. The MLR model slightly underpredicted a majority of the time for all three days.

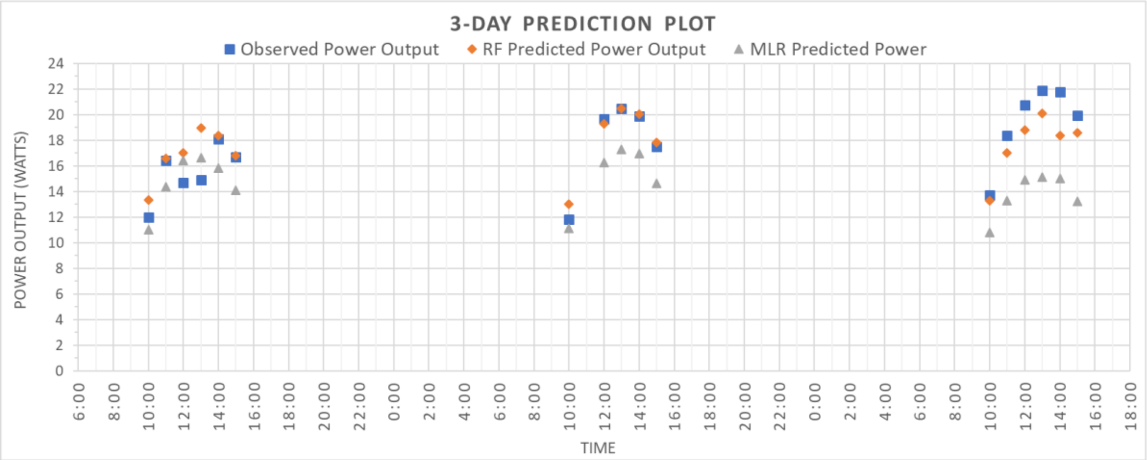


Figure 27. A 3-day prediction summary for Travis Air Force base using each model from 1000 on July 2nd, 2018 to 1500 July 4th, 2018.

V. CONCLUSIONS

The researched aimed to contribute to the US Air Force goals of increased clean and renewable energy implementation in the future. Horizontal solar panel implementation may help achieve renewable energy goals for the future and increase energy resilience across the DoD. This aim is partially achieved by contributing to future modeling efforts of horizontal solar panels. This modeling effort focused on predicting power output without irradiation as an input. If a horizontal solar panel model can predict power output with the use of readily available inputs, such as location and weather data, then assessing the possibility of utilizing horizontal panels in many global locations becomes possible.

Research Question 1: Prediction ability based on model type

The random forest (RF) machine learning outperformed the multivariate linear regression (MLR) model. The RF model accounted for 65.8% of the power output in the validation dataset while the MLR model accounted for 56.2%, a difference of 9.6%. The RF model also accounted for up to 13.24% more variance when predicting power output for each location in the study.

Research Question 2: Input variable importance ranking

Given the 8 input variables used, ambient temperature is the most important input variable in reducing the mean squared error of the RF model. When the importance rankings are scaled, the month of the year is the most important variable in predicting power output of a horizontal solar panel when irradiation is not present. Month was one the top three most important variables for both the scaled and unscaled importance measurements. Therefore, month is controlling for factors that have a high impact on the panel performance. Cloud cover closely followed month in scaled variable importance. This result makes intuitive sense because the amount of clouds present above a panel will affect how much of the sun's energy it will receive.

Research Question 3: Relationships between input factors and power output

The correlation coefficients of humidity, latitude, and altitude in Figure 7 indicate a negative relationship with power output. As humidity increases, power output decreases. More northern locations with higher latitudes have lower power outputs than locations closer to the equator. Locations with higher altitudes have lower power output than locations closer to sea level. Wind speed and cloud ceiling each have a positive relationship with power output. Increased wind speed increases power output. Higher cloud ceilings also lead to higher power outputs. These relationships are also present in the estimated coefficients of the MLR model. The model's interacted variables estimated coefficients indicate that as humidity and temperature each increase, so does power output. Also, power output increases for locations with increased latitude and altitude.

Recommendations for Future Research

The month of the year is an important variable in the RF model. Further work can be done to study which additional factors month may be controlling for within the model aside from sun position. Exploring the possibility of adding more predictor variables to account for irradiation may help increase model accuracy. Solar irradiation data was not available for use in this study. For future studies, adding an instrument to measure irradiation to the experimental set up will allow for the efficiency of the panel in various weather conditions to be analyzed. This can also point to meaningful relationships between the weather conditions and the panel performance. Adding parameters that account for soiling may also be beneficial in future work. Factors such *as time since last rainfall* may indicate the last time dust was cleaned off of the panel by rain. Including a factor that controls for soiling may increase model predictability by controlling for low power output that is due to dust cover on the panel.

Additionally, other modeling techniques and parameters can be used in an effort to better capture the relationship between the input variables and the power output of a horizontal solar panel. Within the MLR model, there may be fixed effects or random effects that the model may be able to control for. Future researchers can consider the possibility of utilizing a fixed effect or random effect model for this experimental context. Researching modeling techniques for time series data and/or panel data may also help better characterize the relationships within the data.

References

[1] Department of the Air Force, “United States Air Force Energy Flight Plan,” 2017.

[2] Greg Stickler, “Educational Brief - Solar Radiation and the Earth System,” *National Aeronautics and Space Administration*, 2016. [Online]. Available: https://web.archive.org/web/20160425164312/http://education.gsfc.nasa.gov/experimental/July61999siteupdate/inv99Project.Site/Pages/science-briefs/ed-stickler/ed-irradiance.html.

[3] John H. Nussbaum, “Analyzing the Viability of Photovoltaic Pavement Systems: A Study in Structural Testing Methods, Measuring Potential Power, and Quantifying the Risks of Implementation,” Air Force Institute of Technology, 2017.

[4] Cory J. Booker, “Analysis of Temperature and Humidity Effects on Horizontal Photovoltaic Panels,” Air Force Institute of Technology, 2018.

[5] Micheal Eckhart, Mohamed El-Ashry, David Hales, Kirsty Hamilton, and Peter Rae, “Renewables 2018 Global Status Report,” National Technical University of Athens, 2018.

[6] “U.S. Air Force Academy Targets Net Zero Energy Goals With a 6 MW Solar System,” *Sun Power*, 2015. [Online]. Available: https://us.sunpower.com/sites/sunpower/files/usairforceacademy-casestudy-2016-r6.pdf. [Accessed: 05-Feb-2018].

[7] “Nellis Air Force Base Builds Largest Solar Photovoltaic Power Plant in North America with SunPower,” *Sun Power*, 2007. [Online]. Available: https://us.sunpower.com/sites/sunpower/files/media-library/case-studies/cs-nellis-air-force-base-builds-largest-solar-photovoltaic-power-plant-north-america-sunpower.pdf. [Accessed: 05-Feb-2018].

[8] Christopher P. Cameron, William E. Boyson, and Daniel M. Riley, “Proceedings of the Comparison of PV System Performance-Model Predictions with Measured PV System Performance,” in *Conference Record of the IEEE Photovoltaic Specialists Conference*, 2008.

[9] Joseph A. Applebee, “Determining the Viability and Efficiency of GP3L Photovoltaic System Study at Air Force Installations in Various Climate Regions,” Air Force Institute of Technology, 2017.

[10] Mathius Maehlum, “Which Solar Panel Type is Best? Mono- vs. Polycrystalline vs. Thin Film,” *Energy Informative*, 2017. [Online]. Available: http://energyinformative.org/best-solar-panel-monocrystalline-polycrystalline-thin-film/. [Accessed: 02-Apr-2018].

[11] Paul Dimotakis, Robert Grober, and Nate Lewis, “Reducing DoD Fossil-Fuel Dependence,” The MITRE Corporation, 2006.

[12] John A Mathews and Hao Tan, “Economics: Manufacture Renewables to Build Energy Security,” *Nature*, vol. 513, no. 7517, pp. 166–168, 2014.

[13] Jerry Warner and P.W. Singer, “Fueling the Balance,” Foreign Policy at Brookings, 2009.

[14] Sierra Hicks, “Powering the Department of Defense,” America Security Project, 2017.

[15] Department of Energy, “2016 Strategic Sustainability Performance Plan,” 2016.

[16] Office of the Assistant Secretary of Defense, “Department of Defense Annual Energy Management and Resilience (AEMR) Report Fiscal Year 2016,” 2017.

[17] NV Energy, “Nellis Solar Array II Generating Station,” 2017. [Online]. Available: https://www.nvenergy.com/publish/content/dam/nvenergy/brochures_arch/about-nvenergy/our-company/power-supply/Nellis-Fact-Sheet.pdf. [Accessed: 02-Mar-2018].

[18] “Solar Energy Project Completed at Plant 42,” *Air Force Materiel Command*, 2016. [Online]. Available: http://www.edwards.af.mil/News/Article/828400/solar-energy-project-completed-at-plant-42/. [Accessed: 02-Apr-2018].

[19] Kevin Elliot, “AF’s Largest Solar Array Celebrates First Anniversary,” *Air Force Civil Engineer Center*, 2014. [Online]. Available: http://www.af.mil/News/Article-Display/Article/558474/afs-largest-solar-array-celebrates-first-anniversary/. [Accessed: 02-Apr-2018].

[20] “Power Purchase Agreements (PPAs) and Energy Purchase Agreements (EPAs),” *World Bank Group*, 2017. [Online]. Available: https://ppp.worldbank.org/public-private-partnership/sector/energy/energy-power-agreements/power-purchase-agreements. [Accessed: 02-Apr-2018].

[21] “Installation Advice, Positioning Solar PV Panels,” *Solar Choice*, 2016. [Online]. Available: https://www.solarchoice.net.au/blog/partial-shading-is-bad-for-solar-panels-power-systems/. [Accessed: 02-Apr-2018].

[22] Elizabeth Morse and Andrew Turgeon, “Solar Energy,” *National Geographic*, 2012. [Online]. Available: https://www.nationalgeographic.org/encyclopedia/solar-energy/. [Accessed: 02-Apr-2018].

[23] Matthew Lave and Jan Kleissl, “Optimum Fixed Orientations and Benefits of Tracking for Capturing Solar Radiation in The Continental United States,” *Renew. Energy*, vol. 36, no. 3, pp. 1145–1152, 2011.

[24] HP Garg, *Treatise on Solar Energy: Volume 1: Fundamentals of Solar Energy*. New York: Wiley, 1982.

[25] John A Duffie and William A Beckman, *Solar Engineering of Thermal Processes*. New York: Wiley, 1980.

[26] Nelson A. Kelly and Thomas L. Gibson, “Increasing the Solar Photovoltaic Energy Capture on Sunny and Cloudy Days,” *Sol. Energy*, vol. 85, no. 1, pp. 111–125, 2011.

[27] J. Antonanzas, R. Urraca, F.J. Martinez-de-Pison, and F. Antonanzas, “Optimal Solar Tracking Strategy to Increase Irradiance in The Plane of Array Under Cloudy Conditions: A Study Across Europe,” *Sol. Energy*, vol. 163, no. September 2017, pp. 122–130, 2018.

[28] Azin Sadeghi Dezfooli, Fereidoon Moghadas Nejad, Hamzeh Zakeri, and Sholeh Kazemifard, “Solar Pavement: A New Emerging Technology,” *Sol. Energy*, vol. 149, pp. 272–284, 2017.

[29] Aditya Shekhar *et al.*, “Harvesting Roadway Solar Energy—Performance of the Installed Infrastructure Integrated PV Bike Path,” *IEEE J. Photovoltaics*, vol. 8, no. 4, pp. 1066–1073, 2018.

[30] E. Skoplaki and J. A. Palyvos, “On the Temperature Dependence of Photovoltaic Module Electrical Performance: A Review of Efficiency/Power Correlations,” *Sol. Energy*, vol. 83, no. 5, pp. 614–624, 2009.

[31] S. Mekhilef, R. Saidur, and M. Kamalisarvestani, “Effect of Dust, Humidity and Air Velocity on Efficiency of Photovoltaic Cells,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 5, pp. 2920–2925, 2012.

[32] Severine Busquet and Jonathan Kobayashi, “Proceedings of Modelling daily PV performance as a function of irradiation , ambient temperature , soiling , wind speed , and aging – Applied to PV modules operating in Maui,” in *IEEE 7th World Conference on Photovoltaic Energy Conversion*, 2015.

[33] Murat Kayri and I. Kayri, “The Performance Comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using Photovoltaic and Atmospheric Data,” *2017 14th Int. Conf. Eng. Mod. Electr. Syst.*, pp. 1–4, 2017.

[34] Ali Lahouar, Amal Mejri, and Jaleleddine Ben Hadj Slama, “Importance based selection method for day-ahead photovoltaic power forecast using random forests,” *Int. Conf. Green Energy Convers. Syst. GECS 2017*, 2017.

[35] Ismail Baklouti, Zied Driss, and Mohamed Salah Abid, “Estimation of solar radiation on horizontal and inclined surfaces in Sfax, Tunisia,” *2012 1st Int. Conf. Renew. Energies Veh. Technol. REVET 2012*, vol. 2, no. 1, pp. 131–140, 2012.

[36] Özge Ayvazoğluyüksel and Ümmühan Başaran Filik, “Estimation methods of global solar radiation, cell temperature and solar power forecasting: A review and case study in Eskişehir,” *Renew. Sustain. Energy Rev.*, vol. 91, no. March, pp. 639–653, 2018.

[37] Yasser Aldali, Ali Naci Celik, and T. Munee, “Modelling and Experimental Verification of Solar Radiation on a Sloped Surface, Photovoltaic Cell Temperature, and Photovoltaic efficiency,” *J. Energy Eng.*, vol. 139, no. 1, p. 49, 2012.

[38] Chih-Chiang Wei, “Predictions of Surface Solar Radiation on Tilted Solar Panels using Machine Learning Models: A Case Study of Tainan City, Taiwan,” *Energies*, vol. 10, no. 10, p. 1660, 2017.

[39] P. Faine, S. R. Kurtz, C. Riordan, and J. M. Olson, “The influence of spectral solar irradiance variations on the performance of selected single-junction and multijunction solar cells,” *Sol. Cells*, vol. 31, no. 3, pp. 259–278, 1991.

[40] R. Eke, T. R. Betts, and R. Gottschalg, “Spectral irradiance effects on the outdoor performance of photovoltaic modules,” *Renew. Sustain. Energy Rev.*, vol. 69, no. December 2014, pp. 429–434, 2017.

[41] Wei Zhou, Hongxing Yang, and Zhaohong Fang, “A Novel Model for Photovoltaic Array Performance Prediction,” *Appl. Energy*, vol. 84, no. 12, pp. 1187–1198, 2007.

[42] National Oceanic and Atmospheric Atministration, “National Center for Environmental Information,” 2019. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/.

[43] S. Kilic, “Linear regression analysis,” *J. Mood Disord.*, vol. 3, no. 2, p. 90, 2013.

[44] B. Hammad, M. Al-Abed, A. Al-Ghandoor, A. Al-Sardeah, and A. Al-Bashir, “Modeling and analysis of dust and temperature effects on photovoltaic systems’ performance and optimal cleaning frequency: Jordan case study,” *Renew. Sustain. Energy Rev.*, vol. 82, no. April 2017, pp. 2218–2234, 2018.

[45] Micheal H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li, *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill/Irwin, 2005.

[46] Jake VanderPlas, *Python Data Science Handbook*. O’Rielly Media, 2016.

[47] Leo Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[48] Gareth James, Daniela Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

[49] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning Data*, Second. Stanford: Springer, 2017.

[50] “JMP Pro.” SAS Institute Inc., Cary, NC, 2019.

[51] R Development Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, 2008.

[52] RStudio Team, “RStudio: Integrated Development for R.” RStudio, Inc., Boston, 2016.

[53] Graham J. Williams, “Rattle: a data mining GUI for R,” *R J.*, vol. 1, no. 2, pp. 45–55, 2009.

[54] International Civil Aviation Organization, “Meteorological Service for International Air Navigation,” *International Standards and Recommended Practices Annex*, 2007. [Online]. Available: https://www.wmo.int/pages/prog/www/ISS/Meetings/CT-MTDCF-ET-DRC_Geneva2008/Annex3_16ed.pdf. [Accessed: 06-Dec-2018].

[55] UCAR Center for Science Education, “The Highs and Lows of Air Pressure,” 2019. .

[56] “Surface Hourly Abbreviated Format,” 2012. [Online]. Available: ftp://ftp.ncdc.noaa.gov/pub/data/noaa/ish-abbreviated.txt. [Accessed: 01-Aug-2019].

[57] Columbia University Press, “Dew,” *The Columbia Encyclopedia*. 2000.

[58] Achim Zeileis, Thomas Lumley, Susanne Berger, and Nathaniel Graham, “sandwich: Robust Covariance Matrix Estimators.” R package version 2.5-0, 2018.

[59] R Core Team, “stats: The R Stats Package.” R Foundation for Statistical Computing, 2015.

[60] Torsten Hothorn, Achim Zeileis, Richard W. Farebrother, Clint Cummins, Giovanni Millo, and David Mitchell, “lmtest: Testing Linear Regression Models.” R package version 0.9-36, 2018.

[61] Max Kuhn, “caret: Classification and Regression Training.” R package version 6.0-81, 2018.

[62] Joachim Gassen, “ExPanDaR: Explore Panel Data Interactively.” R package version 0.3.0, 2018.

[63] Monto Mani and Rohit Pillai, “Impact of dust on solar photovoltaic (PV) performance: Research status, challenges and recommendations,” *Renew. Sustain. Energy Rev.*, vol. 14, no. 9, pp. 3124–3131, 2010.

[64] Rhythm Singh and Rangan Banerjee, “Impact of Solar Panel Orientation on Large Scale Rooftop Solar Photovoltaic Scenario for Mumbai,” *Energy Procedia*, vol. 90, no. December 2015, pp. 401–411, 2015.

[65] P. Tsalides and A. Thanailakis, “Direct computation of the array optimum tilt angle in constant-tilt photovoltaic systems,” *Sol. Cells*, vol. 14, no. 1, pp. 83–94, 1985.

[66] Caroline Housmans, Alessandro Ipe, and Cedric Bertrand, “Tilt to horizontal global solar irradiance conversion: An evaluation at high tilt angles and different orientations,” *Renew. Energy*, vol. 113, pp. 1529–1538, 2017.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)* 21-03-2019	**2. REPORT TYPE** Master's Thesis	**3. DATES COVERED** *(From – To)* Sep 2017 – March 2019

TITLE AND SUBTITLE Modeling Power Output of Horizontal Solar Panels Using Multivariate Linear Regression and Random Forest Machine Learning	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Pasion, Christil K., 2nd LT, USAF	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way, Building 640 WPAFB OH 45433-7765	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT-ENV-MS-19-M-192
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) The Civil Engineer School 2950 Hobson Way WPAFB, OH 45433-7765 (937) 255-5654 x2156 (DSN 5654 x2156) cess@afit.edu	**10. SPONSOR/MONITOR'S ACRONYM(S)** CESS **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A. Approved For Public Release; Distribution Unlimited.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
United States Air Force energy resiliency goals are aimed to increase renewable energy implementation. Researchers at the Air Force Institute of Technology distributed 37 photovoltaic test systems around the world. This research uses multivariate linear regression and random forest machine learning to determine which modeling technique will better predict power output for horizontal solar panels. If power output of a horizontal solar panel can be predicted using available weather data, then assessing the possibility of utilizing horizontal panels in any global location becomes possible. The linear model accounted for 56.2% of the variance in a validation dataset. The random forest model accounted for 65.8% variance. The most important variable in reducing the random forest model mean squared error was the month of the year, closely followed by cloud ceiling. Wind speed was the least important variable in reducing model error. More predictor variables are needed to increase predictability of horizontal solar panel power output if irradiation is not present as an input.

15. SUBJECT TERMS
Photovoltaics, multivariate linear regression, random forest machine learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Lt Col Torrey Wagner, AFIT/ENV
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U	UU	100	(937) 255-6565, ext 4611 torrey.wagner@afit.edu