January 2017

A Data Driven Approach to Solar Generation Forecasting

Guangyuan Shi
Syracuse University

ABSTRACT

With the developments in renewable energy resources, more Photovoltaic (PV) generators are being built. Compared to traditional generators, a PV generator is less controllable which will adversely impact power system operation and planning. To ensue seamless operation of power systems, PV forecasting is essential and necessary. A challenge in PV forecasting is that PV generation behavior differs in different regions due to the fact that PV generation is highly dependent on weather conditions, in particular solar irradiance. This makes it important to study the power output data based on a specific region. In this thesis, I first analyze how PV forecasting will affect system planning by calculating probabilistic power flow (PPF). By using a variety of probabilistic models that can estimate solar irradiance, the PPF of each model is calculated and compared. The PPF will give us an idea of how accurate and inaccurate forecasts will affect power system operations and planning. I then seek to find out which method can forecast the power output more accurately. I used several methods such as Linear Regression, Artificial Neural Network (ANN), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), and compared these methods in 3-hours ahead forecasting. In addition, these methods are analyzed for future use, as the dataset used is constantly growing. Through my analysis of the data, I found out that, based on a small dataset, linear regression works better and as the dataset grows larger, the error for K-Nearest Neighbor reduces dramatically. In addition, a new approach named Symbolic Aggregate approximation (SAX) was used when an extremely large dataset was used to increase calculation speed and reduce dimensionality.

A Data Driven Approach to Solar Generation Forecasting

by

Guangyuan Shi

B.E., North China Electrical Power University, 2015

Thesis
Submitted in partial fulfillment of the requirements for the degree of
Master of Science in Electrical Engineering.

Syracuse University
May 2017

Acknowledgments

There are a lot of people I want to thank and, without them, I could not possibility have finished this thesis.

To my advisor, Doctor Sara Eftekharnejad, whom led my way during my research and gave me a lot of advice and resources to guide me. She is such an excellent professor that I can feel her full support for both my academic studies and my personal life.

I would like to thank Syracuse University's Center for Advanced Systems and Engineering (CASE) and Greenview EMS Company for their support of this thesis. I thank them for providing the funding and the data for this research.

I would also like to thank my defense committee, Doctor Bujanovic, Doctor Gursoy and Doctor Zafarani, for providing valuable comments and feedback on my thesis. I appreciate them spending their valuable time to serve on my committee. Also to my professors in Syracuse University, Doctor Chen, Doctor Ghosh, Doctor Lee and Doctor Mojdehi, who helped me not only in my academic studies but also to prepare for my future career. I am also grateful to have had the opportunity to work with all the members in Syracuse University Power System & Renewable Energy Lab: Mirjavad, Rui, Sagnik and Wolf Peter. Although we came from different countries and have different culture backgrounds, we share the same enthusiasm to solve the unknowns in power field.

Finally and most importantly, to my parents, Huiping Shi and Wenzhou Fan, who encouraged me to study far away from home and who, I know, made a lot of sacrifices along the way.

Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

In the past decade, solar power has taken more share of energy usage percentage. In 2008, renewable energy accounted for 9.3% of total generation in the US and reached 13.1% in 2013 [1]. National Renewable Energy Laboratory (NREL) estimated that by 2050, 80 percent of U.S. energy will be provided by renewables [2]. Compared to traditional energy resources such as natural gas or coal, solar power has the following characteristics: renewable, environmental friendly, and requires little maintenance. However, compared to traditional energy resources, solar power is generally expected to cost more and most of all, solar power is weather dependent which makes it hard to control. The uncontrollable photovoltaic (PV) generator will cause grid uncertainty and may even cause blackout. Therefore, it is important to forecast PV generation a few hours ahead to give system operators time to operate and effectively dispatch the power.

Unlike traditional generators, due to sudden changes in weather, PV generation can vary dramatically. Traditional power flow solutions are extremely sensitive to these PV variations, where a small change can cause significant difference in power flow calculations. To alleviate this problem, probabilistic power flow (PPF) is used to focus on grid uncertainty and access a range of loads and generations. In 1974, Borkowa introduced the concept of PPF [3]. Later, Heydt and Allan introduced two methods to calculate PPF: the Stochastic Load Flow (SLF) and the Probabilistic Load Flow (PLF) [4], [5]. SLF considers short-term uncertainties while PLF considers long-term uncertainties. Further developments include point estimate scheme method introduced in reference [6]. Recent advancements in PPF includes combining the concept of cumulants and different types of expansions to calculate cumulative distribution function (CDF) and probabilistic density function (PDF) of power flows. This approach is in general more accurate, and requires less computing time. In recent study [7], Fan et. al. compared three different types of expansion: Gram-Charlier, Cornish-Fisher and Edgeworth. He and his team found out that Gram-Charlier operates better during extreme conditions while other two are more accurate at other conditions.

Two types of PV forecasting techniques exist in the literature: long term and short term. Long term refers to one day or more ahead forecasting, and short term refers to a few hours ahead forecasting. In 2009, an Artificial Neural Network (ANN) based method was proposed by Adel Mellit, and since then, this method has been widely used in PV generation forecasting [8]. In [9], Ding and their team applied an improved back-propagation (BP) learning algorithm to overcome slow convergence and used a similar day method to improve accuracy. In addition, instead of forecasting solar irradiance, this work used a Feed-forward Neural Network (FNN) to forecast power output directly. In [10], a self-organized map (SOM) is used to classify the weather based on ANN to improve the accuracy. Other methods, such as Support Vector Machine (SVM), are also used in PV forecasting in [11]–[13]. The difference is that [11] focuses on finding the best kernel functions between Radial Basis Function (RBF) kernel, Linear Kernel, and Polynomial Kernel. The authors in [12] combine weather classification with SVM. In [13], weight factors are assigned in the SVM method. The author first selected five similar days as training samples and then weighted them based on the combined similarities and time interval. Other data mining methods, such as decision trees [14], and k-means clustering [15] are also used for PV forecasting. The difference between these two methods is how data is classified. Mandal in [16] brought up the idea of wavelets. Their method has three steps: (1) The wavelets decomposed the PV power into four components; (2) The decomposed signal is analyzed by Neural Networks; and (3) The forecasted signal is reconstructed using wavelets. Similar to wavelets, a faster and less dimensional method called Symbolic Aggregate approXimation (SAX) was proposed by Lonardi in 2002 [17], and modified by Lin in 2003 [18]. However, this new method has not yet been used in PV forecasting. Some applications of this method analyze PV impact such as voltage rise, reverser power flow, variation of feeder power loss, voltage unbalance, and change in tap operations [19]. Reference [20], [21] used the position of the sun and global solar radiation to forecast PV power output. Compared to traditional data mining forecast approaches, they add geography factors to their methods. Other proposed methods include autoregressive [22], extreme learning machine [23], K-Nearest Neighbor (KNN) [24], and naïve Bayes classifier [25].

The aforementioned methods include the following characteristic: First, they use a time series forecasting data mining technology. Then, they add other requirements such as classifications or geography locations. Currently, none of these works can accurately forecast weather situations such as rainy or cloudy conditions. For example, in the Syracuse area, due to the lake effect, it is often cloudy, rainy, or snowy. Therefore, my approach in this thesis is to focus on rainy or cloudy weather conditions forecasting, i.e., the sample days chosen in this thesis include the aforementioned type of weather conditions. In addition, most of these forecasts have large datasets, thus I aim to find out the best forecast method not only for large datasets but also for limited datasets and extremely large datasets. Moreover, I used linear regression and the SAX method in PV generator forecasting. In addition, I combined classification with the KNN method to improve accuracy. Methods such as ANN are used for comparison purposes, and the errors are calculated by parameters such as root-mean-square-error (RMSE) and mean-absolute-error (MAE). In this thesis, the focus is on 3-hour ahead forecasting and the results of each method are compared.

1.2 Thesis Outline

This thesis is organized as follows: in chapter II, PPF calculation is explained the effect of forecast accuracy on PPF calculation and system planning is shown. In chapter III, some challenges in PV forecasting are listed, and the solution to these challenges are provided. In addition, three different PV forecasting methods are used in different sizes of datasets. Finally, the conclusion and future work are presented.

CHAPTER 2
PROBABILISTIC POWER FLOW AND SYSTEM PLANNING

In this chapter, the concept and calculation of PPF is introduced. Also, the relationship between PPF and system planning is demonstrated. This relationship proves how important accurate forecast is.

2.1 Probabilistic Power Flow Calculation

As mentioned in chapter 1, there are many ways to calculate PPF. In this thesis, I combined 6th order of Gram-Charlier expansion with cumulants to calculate PPF [26]. This method takes two steps, first step is to linearize the power flow equations and second step is to combine the expansion with cumulants.

2.1.1 Linearized Power Flow Equations

Traditional power flow equations are as follows :

$$P_i = V_i \sum_{k=1}^{n} V_k (G_{ik} \cos\theta_{ik} + B_{ik} \sin\theta_{ik}) \tag{2.1}$$

$$Q_i = V_i \sum_{k=1}^{n} V_k (G_{ik} \sin\theta_{ik} - B_{ik} \cos\theta_{ik}) \tag{2.2}$$

$$P_{ik} = -t_{ik} G_{ik} V_i^2 + V_i V_k (G_{ik} \cos\theta_{ik} + B_{ik} \sin\theta_{ik}) \tag{2.3}$$

$$Q_{ik} = t_{ik} B_{ik} V_i^2 - B_{ik}^{'} V_i^2 + V_i V_k (G_{ik} \sin\theta_{ik} - B_{ik} \cos\theta_{ik}) \tag{2.4}$$

$$Q_{i(sh)} = V_i^2 B_{i(sh)} \tag{2.5}$$

Where B_{ik} represents the imaginary part of element *ik* of admittance matrix, G_{ik} represent the real part of element *ik* of admittance matrix.

Since the equations are not linearized, the following assumptions are made to linearize the equations:

$$V_i = V_k = 1\ p.u. \tag{2.6}$$

$$G_{ik} = 0 \tag{2.7}$$

$$sin\vartheta_{ik} = \vartheta_{ik} \tag{2.8}$$

The assumptions are made due to the fact that line resistance and angle ϑ_{ik} are very small. Therefore, the equations are changed to:

$$\theta_i = \sum_{k=1}^{n-1} L_{ik} P_k \quad (i = 1, \cdots, n-1) \tag{2.9}$$

$$P_{ik} = \frac{1}{X_{ik}} \sum_{j=1}^{n-1} (L_{ij} - L_{kj}) P_j \tag{2.10}$$

$$P_s = \sum_{i=1}^{n} P_i \tag{2.11}$$

where L_{ik} represent the matrix of $1/X_{ik}$. Then by assuming $V_i^2 = V_i$, $V_i V_k = V_k$ equations (2.4-2.5) will become as follows (2.12-2.13):

$$Q_{ik} = \omega V_i + \chi V_k \tag{2.12}$$

$$Q_{i(sh)} = V_i B_{i(sh)} \qquad (2.13)$$

where *ω=t_{ik}B_{ik}- B_{ik}′* and *Ҳ= G_{ik} sinϑ_{ik}- B_{ik}cosϑ_{ik}*. By making *V_i =1 p.u.* Equation (2.2) is changed to:

$$Q_i = \sum_{k=1}^{n} \chi V_k \qquad (2.14)$$

Reference [27] gives details on how these equations are linearized and how the calculation time is reduced.

2.1.2 Combine Expansion with Cumulants

Gram-Charlier Expansion PPF calculation is performed in [26]. In this thesis, this method is used with some modifications. In my proposed method, there are eight steps to calculate PPF. These steps are outlined next.

1) The v^{th} moment of generation distribution is calculated:

$$\alpha_v = E(\xi^v) = \int_{-\infty}^{\infty} x^v f(x) dx \qquad (2.15)$$

In this equation, *f(x)* represents the PDF of generation [28] and 1^{st} moment stands for the mean value.

2) Then the moments $\alpha_1 \ldots\ldots \alpha_n$ are used to express cumulants of injected power k_n:

$$\begin{aligned} k_1 &= \alpha_1 = \mu \\ k_2 &= \alpha_2 - \alpha_1^{\ 2} \\ k_3 &= \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^{\ 3} \\ k_4 &= \alpha_4 - 3\alpha_2^{\ 2} - 4\alpha_1\alpha_3 + 12\alpha_1^{\ 2}\alpha_2 - 6\alpha_1^{\ 4} \\ &\ldots\ldots \end{aligned} \qquad (2.16)$$

Where μ is the mean value of the input.

3) The cumulants related to the *i^th^* line flow are calculated:

$$k_{vp} = R_{(ik)1}{}^{v}k_{v}{}^{(1)} + R_{(ik)2}{}^{v}k_{v}{}^{(2)} + \cdots R_{(ik)n}{}^{v}k_{v}{}^{(n)} \tag{2.17}$$

where $R_{(ik)j}$ denotes $(L_{ij} - L_{kj})/X_{ik}$.

4) By repeating step (1) and (2), the moments and cumulants of bus voltage distribution are calculated to get the PPF of reactive power.

$$k_{vv} = \chi_{ik}{}^{v}k_{v}{}^{(1)} + \omega_{ik}{}^{v}k_{v}^{(1)} + \ldots + \chi_{ik}{}^{v}k_{v}{}^{(n)} + \omega_{ik}{}^{v}k_{v}^{(n)} \tag{2.18}$$

5) To get the cumulants of apparent power *S*, I combine k_{vp} and k_{vv}:

$$k_{v} = \sqrt{k_{vp}{}^{2} + k_{vv}{}^{2}} \tag{2.19}$$

6) I use the cumulants of *S* to compute the central moments of each line:

$$\begin{aligned}
\beta_1 &= 0 \\
\beta_2 &= k_2 = \sigma^2 \\
\beta_3 &= k_3 \\
\beta_4 &= k_4 + 3k_2{}^2 \\
\beta_5 &= k_5 + 10k_2k_3 \\
\beta_6 &= k_6 + 15k_2k_4 + 10k_3{}^2 + 15k_2{}^3, \ldots
\end{aligned} \tag{2.20}$$

Where σ^2 is the variance of the injected power.

7) The Gram-Charlier expansion coefficients are calculated by using the following equation [29]:

$$\begin{aligned} c_0 &= 1 \\ c_1 &= c_2 = 0 \\ c_3 &= -\frac{\beta_3}{\sigma^3} \\ c_4 &= \frac{\beta_4}{\sigma^4} - 3 \\ c_5 &= -\frac{\beta_5}{\sigma^5} + 10\frac{\beta_3}{\sigma^3} \\ c_6 &= \frac{\beta_6}{\sigma^6} - 15\frac{\beta_4}{\sigma^4} + 30, \ldots \end{aligned} \tag{2.21}$$

8) The CDF and the PDF of line apparent power are calculated:

$$\begin{aligned} F(x) &= \Phi(x) + \frac{c_1}{1!}\Phi^{'}(x) + \frac{c_2}{2!}\Phi^{"}(x) + \frac{c_3}{3!}\Phi^{(3)}(x) + \cdots \\ f(x) &= \varphi(x) + \frac{c_1}{1!}\varphi^{'}(x) + \frac{c_2}{2!}\varphi^{"}(x) + \frac{c_3}{3!}\varphi^{(3)}(x) + \cdots \end{aligned} \tag{2.22}$$

where *$\varphi(x)$* and *$\phi(x)$* represent the PDF and the CDF of a normal distribution. The PPF of each line is calculated by following the 8 steps mentioned above. In this thesis, the 6th order of Gram-Charlier expansion is used because this order is the most accurate one at the tail end [7]. In addition, accurate tail end will assure high efficiency of the power system.

2.2 Solar Forecasting and System Planning

Currently, there are different probabilistic models in solar irradiance forecasting. In this thesis, five models are demonstrated and their efficiency are analyzed.

2.2.1 Forecasting Models

These models include Beta, normal, log-normal and Weibull distributions in literature [30]–[33]. The aforementioned distributions are as follow:

1) *Beta Distribution: Beta(a,b)*

$$f_\gamma(\gamma)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(\frac{\gamma}{\gamma_{\max}})^{a-1}(1-\frac{\gamma}{\gamma_{\max}})^{b-1} \tag{2.23}$$

Where γ denotes solar irradiance and γ_{max} denotes its maximum value.

There are two ways to calculate beta parameters *a* and *b*. In *Beta(a₁,b₁)* model, parameters are calculated as:

$$a_1=\mu[\frac{\mu(1-\mu)}{\sigma^2}-1] \tag{2.24}$$

$$b_1=(1-\mu)[\frac{\mu(1-\mu)}{\sigma^2}-1] \tag{2.25}$$

where μ is the mean value and σ is the standard deviation of solar irradiance.

In the second Beta model, the parameters are calculated below:

$$b_2=(1-\mu)[\frac{\mu(1+\mu)}{\sigma^2}-1] \tag{2.26}$$

$$a_2=\frac{\mu b}{1-\mu} \tag{2.27}$$

2) *Weibull Distribution f(x|a,b)*

The PDF of Weibull distribution is shown in the following equation:

$$f_\gamma(x\,|\,a,b)=\frac{b}{a}(\frac{x}{a})^{b-1}e^{-(x/a)^b} \tag{2.28}$$

In equation (2.28), $0<x<\infty$, $a=\mu$ and $b=\mu(1-\mu)/\sigma^2-1$.

3) *Log-Normal Distribution*

The PDF of Log-normal distribution is defined as:

$$N(\ln x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)}{2\sigma^2}} \qquad (2.29)$$

where $0<x<\infty$

4) *Normal Distribution*

$$f(x\,|\,\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2.30)$$

The accuracy of five different types of distributions listed above are compared using solar irradiance data.

2.2.2 Comparison Between the Forecasting Models

In this thesis, I used solar irradiance data from National Solar Radiation Database of the Syracuse area in 2010 to analyze the accuracy of different forecasting models [34]. This hourly data is collected near Syracuse Hancock airport throughout 2010. The angle of the sun and the amount of solar irradiance is also included in this data. Figure1 shows the irradiance amount in the dataset where *x*-axis represents the days and *y*-axis represents the irradiance amount. In *x*-axis, day 1 is January 1st and day 365 means December 31st. From the figure, it is shown that solar irradiance increases dramatically during summer and decreases during winter season.

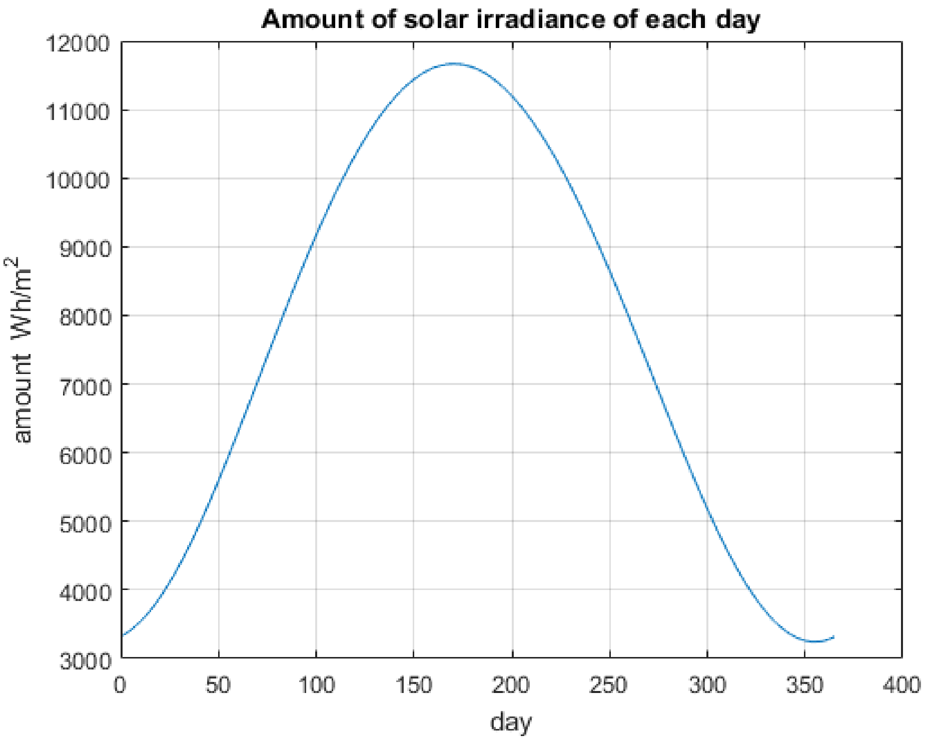

Figure 1 Amount of daily solar irradiance in 2010 [34]

Figure 2 compares the PDF of actual solar irradiance data and different distributions to see their accuracy. In my case, *Beta(a1,b1)* is the most accurate forecast, Weibull is next followed by *Beta(a2,b2).* The other two methods, i.e. Log-Normal and Normal distributions, are not accurate enough to be used.

To see the differences between forecasting methods, I calculate the RMSD and compare it to the actual data [35].

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(YA_i - YF_i)^2}{n}} \qquad (2.31)$$

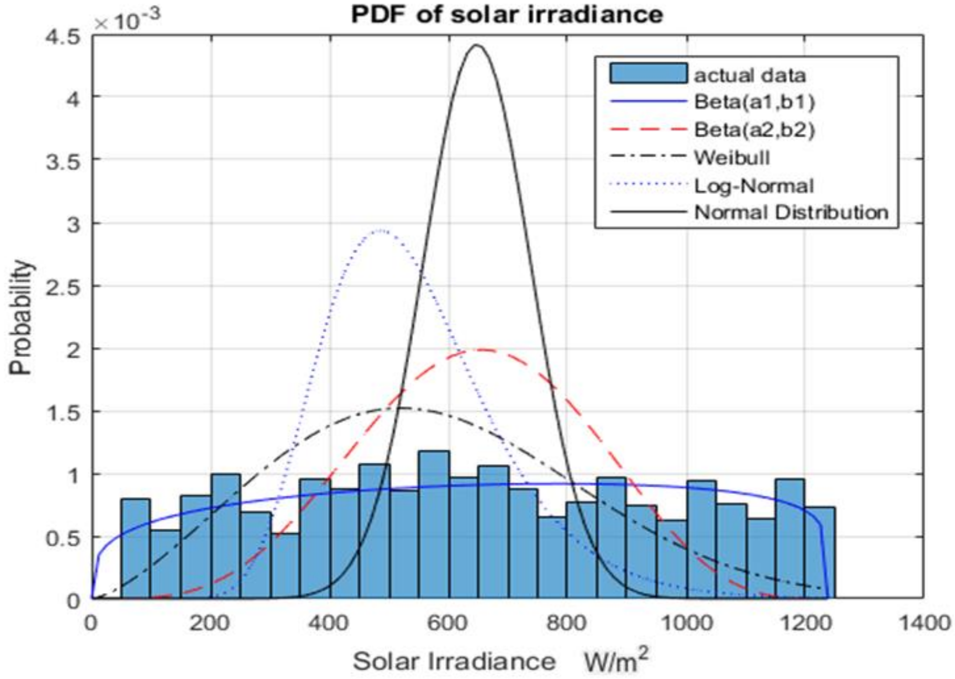

Figure 2 PDF of actual solar irradiance data and different forecasting models

In equation (2.31), *YA*$_i$ represents the *i*th point at the actual solar irradiance, *YF*$_i$ represents the *i*th point at the forecasted solar irradiance and *n* is the total number of points. In Table 1 below, the RMSD value of different forecasts are listed. The more accurate the forecast is, the smaller the RMSD value.

Table 1 RMSD value of each forecasting method

Prediction	RMSD
Beta(a1,b1)	0.00015
Beta(a2,b2)	0.00057
Weibull	0.00040
Log-Normal	0.00078
Normal	0.00110

In a PV generator,

$$P = \gamma S \eta (1 - n\Delta t) \tag{2.32}$$

Where *P* stands for the generated solar active power; *γ* represents the amount of solar irradiance; *S* is the PV modules' total area; the efficiency is represented by *η*; *Δt* is the PV cell temperature's forecast error; and *n* is the coefficient of the temperature which is provided by the solar cell manufactures. Therefore, there is a linear relationship between active power *P* and solar irradiance *γ*, which means the PDF and CDF of active power and solar irradiance are similar.

2.2.3 Impact of Solar Forecasting on the System

To see the impact of different types of forecasting, I used IEEE 30 bus system as test system for this study. The PPF of each forecast model is calculated and compared. In this thesis, I focus on the extreme operating percentage of each line. Hence, I considered apparent power *(S)* instead of considering active power *(P).*

In my case, there are six PV generators in the system, located at buses 1,2,4,7,8 and 13. Among these generators, the one on bus 2 has the largest capacity and the generator located on bus 7 has the smallest.

In this thesis, 90% MVA rating is analyzed because this means the line flow is close to its maximum capacity but it does not exceed line rating. By comparing the line's 90% MVA rating and the CDF of apparent power, the line's extreme operating percentage can be known. Also, Monte Carlo simulation is used for comparison purposes. The confidence level of apparent power (CDF>90%) is listed and RMSD value is calculated as the accuracy index of PPF. For CDF 90% and over 90% MVA Rating, it is the value of these two parameters closer to Monte Carlo the more accurate. In addition, for RMSD value, it is the smaller the value the better.

In Fig. 3-5, the CDF of power flow in three different lines are drawn and the Monte Carlo method is used as reference. These lines are No.13, No.8 and No.4.

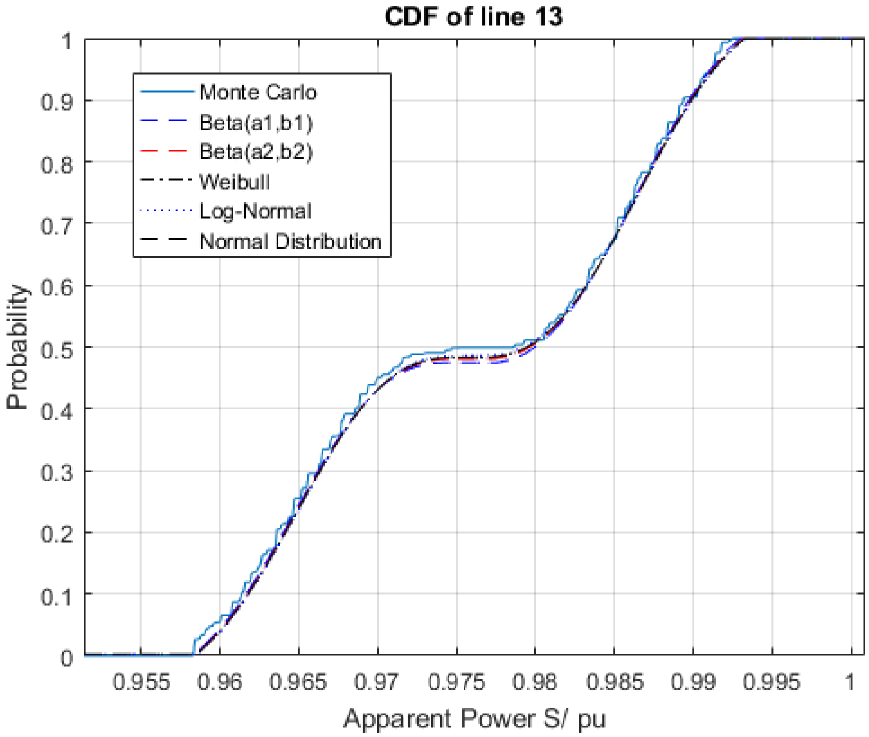

Figure 3 CDF of line 13

Line 13 is between bus 9 and bus 10, which are not connected to any PV generators. As seen in Table 2, the difference between each forecast is very limited in this line. In addition, the RMSD values are similar and the CDF over 90% does not change much. However, the probability of apparent power over 90% varies much more relatively. From Table 2, it is shown that the tail end of *Beta(a1,b1)* is the closest which means it has the closest extreme operating probability compared to other methods.

Table 2 RMSD value and tail end results for line 13

Prediction	RMSD	CDF 90%*	Over 90% MVA Rating**
Monte Carlo	N/A	0.9901	17.10%
Beta(a1,b1)	0.0219	0.9898	19.90%
Beta(a2,b2)	0.0230	0.9899	20.26%
Weibull	0.0234	0.9899	20.36%
Log-Normal	0.0239	0.9899	20.37%
Normal	0.0234	0.99	20.53%

*CDF 90% - cumulative density function over 90%

**Over 90% MVA Rating – the probability of apparent power over 90% of line MVA Rating

The error percentage can be calculated by using the following equation:

*Error= (MC- Forecast)/MC*100%* (2.33)

In equation (2.33), *MC* represents the Monte Carlo value and *Forecast* represents the forecast value.

From analyzing the maximum and minimum error percentage of MVA rating, Table 2 shows that *Beta(a1,b1)* has the smallest error percentage of 16.37%, while normal distribution has an error of 20.06%. As mentioned above, the buses connected to line 13 are not connected to any PV generators. Additionally, these buses are not connected to any other bus that is connected to PV generators.

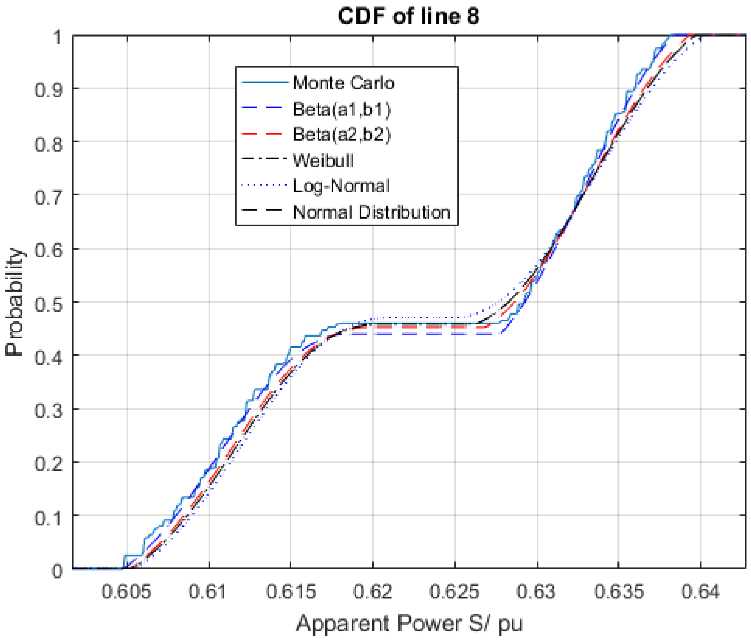

Figure 4 CDF of line 8

Table 3 RMSD value and tail end results for line 8

Prediction	RMSD	CDF 90%	Over 90% MVA Rating
Monte Carlo	N/A	0.6355	5.76%
Beta(a1,b1)	0.0453	0.6361	5.89%
Beta(a2,b2)	0.0477	0.6366	4.42%
Weibull	0.0526	0.6368	3.89%
Log-Normal	0.0603	0.6371	3.15%
Normal	0.0536	0.6368	3.92%

Line 8 is connecting bus 5 and 7. In my system, bus No.7 is connected to a small generator and bus No.5 is connected to two other buses, which are connected to PV generators. Compared to line 13, the differences between each forecast are much more significant and the extreme operating probability errors are rather high. However, the difference between each tail end is not that significant. Listed in Table 3, the CDF over 90% and RMSD value changed much more than in Table 2. This means that Line 8 is much more sensitive to accurate and inaccurate forecast than Line 13. This is because Line 8 is connected to a small PV generator while Line 13 is not connected to any PV generators at all.

From Table 3, *Beta(a1,b1)* has the least error percentage in MVA rating which is 2.2% while Log-Normal has the largest which is 45.31%. For this line, the error of accurate and inaccurate line's extreme operating probability can vary significantly.

The third line No.4 is from bus 2 to bus 5. In this line, bus 2 is connected to a large PV generator and bus 5 is connected to two other buses, which is connected to PV generators. The difference between accurate and inaccurate forecast is much more dramatic than the other two lines. In Table 4, there are significant differences in the CDF over 90%. *Beta(a1,b1)* is still the closest to the actual data which has the error probability of 19.65% while the maximum model Log-Normal has 56.33%.

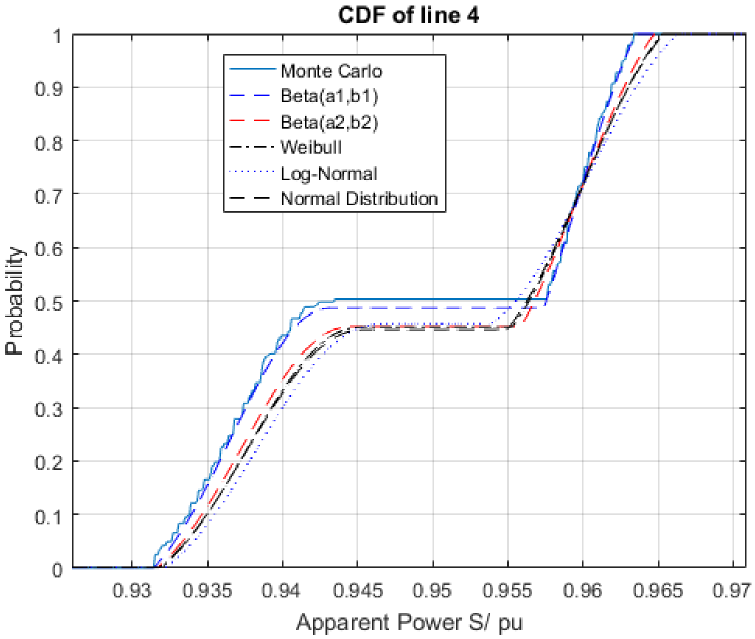

Figure 5 CDF of line 4

Table 4 RMSD value and tail end results for line 4

Prediction	RMSD	CDF 90%	Over 90% MVA Rating
Monte Carlo	N/A	0.9619	10.74%
Beta(a1,b1)	0.0753	0.9621	12.85%
Beta(a2,b2)	0.0819	0.9628	15.12%
Weibull	0.0842	0.9630	15.74%
Log-Normal	0.0847	0.9636	16.79%
Normal	0.0863	0.9631	15.83%

In my system, line 13 has the least connection to PV generators while line 4 has the most. Due to the connection with PV generators, lines behave differently with different forecast models. Therefore, in my case, the PPF of line 4 varies the most while line 13 varies the least.

In Syracuse area, *Beta(a1,b1)* is the most accurate forecast and normal or Log-Normal are the least. This is shown from the tail end of PPF and the result matching the forecasted data.

2.2.4 System Planning Analysis

In this thesis, the PPF of five different types of forecast models were calculated and compared. *Beta(a1,b1)* provides the most accurate forecast and normal distribution or Log-Normal distribution provides the least accurate one. In PPF, transmission lines demonstrated a diverse performance due to their location. Some lines are much more sensitive to forecast accuracy and some are less sensitive. This is dependent to the location of PV system with respect to transmission lines. The closer the transmission line to a large PV generator, the more sensitive the PPF is to PV forecasting accuracy.

From the percentage of MVA rating error point of view, accurate forecast can cause less than 20% error in line’s extreme operating probability while inaccurate forecast can cause more than

40%. This indicates for Lines connected to large PV generator, an accurate forecast is very important or it can hugely affect system planning. If the apparent power of a system is estimated with 40% error, a line could take too much load which cause overload. In addition, if an overloading line has been take out, it might even cause blackout. On the other hand, less sensitive lines are not so dependent on accurate forecast.

CHAPTER 3
PHOTOVOLTAIC GENERATOR OUTPUT FORECASTING

Chapter 2 discussed the importance of PV forecasting, while in this chapter, I forecasted the PV output by applying different machine learning methods. In this chapter, the challenges in PV power output forecasting are demonstrated and some forecasting methods are listed. In addition, I used different forecasting methods based on three different sizes of data and the most accurate method for each data size is found.

3.1 Challenges in PV Power Output Forecasting

From the given data of the Syracuse area, four major challenges in PV forecasting are demonstrated and solved.

3.1.1 Impact of Weather Parameters

PV power output majorly depends on the solar irradiance. Equation (2.32) also proved an estimated relationship between solar irradiance and power output. However, this equation is not accurate enough to be used. In addition, other weather parameters such as temperature and cloud cover can also be very important to PV generation output. Additionally, compared to load forecasting, solar forecasting is much more difficult due to uncertainty of weather conditions. This makes it important to analyze the relationship between weather parameters and power output.

In Fig. 6, I listed three random days in August and it is shown that while power output is highly dependent on solar irradiance, it is not the only factor. This means other weather parameters such as temperature will also affect power output.

In Fig. 7, take August 4th as an example, typically when solar irradiance is high, power output should be high, but this does not happen in hour 13 (1:00pm). In addition, during hour 8 to 10 (8am to 10am), solar irradiance is very low while power output is around 5 to 10 kw. This might be caused by other weather parameters such as temperature.

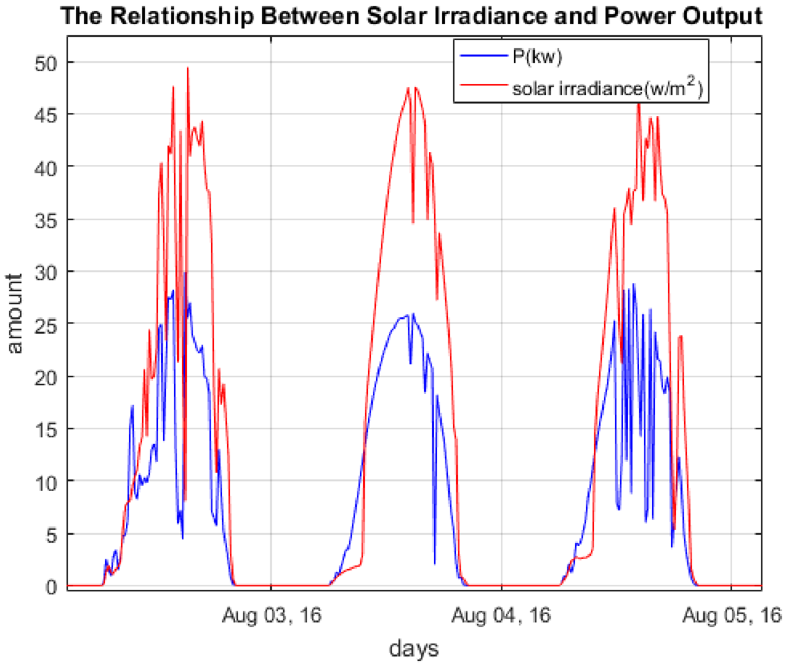

Figure 6 Relationship between solar irradiance and power output of three days

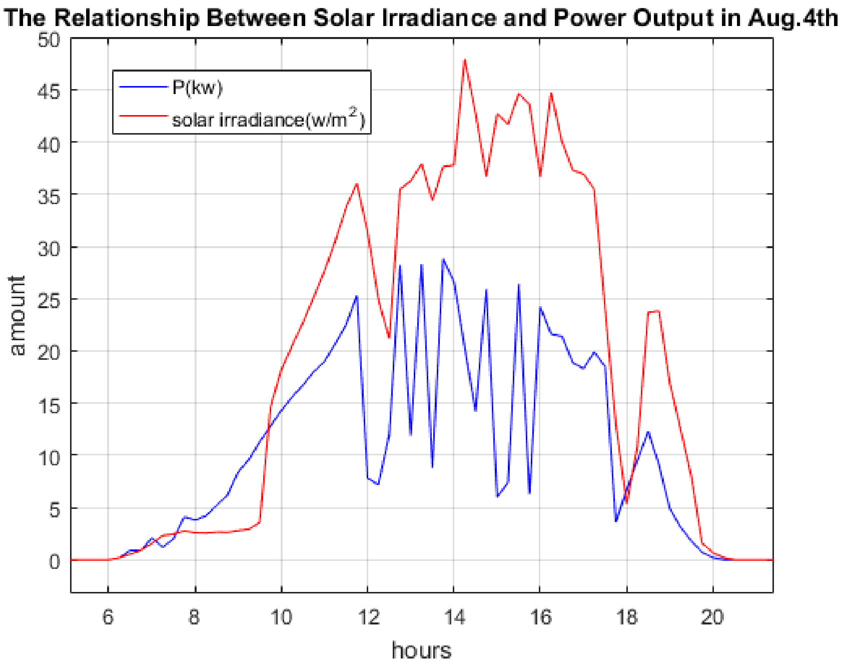

Figure 7 Relationship between solar irradiance and power output of Aug. 4th

To overcome this difficulty, I classified the data into different categories. Based on my data, 10 weather parameters are analyzed which include solar irradiance, wind speed, wind direction, cloud cover, humidity, precipitant, air pressure, temperature, visibility, and weather code. I found that the relationships could be divided into three types based on the correlation: positive, negative, and small correlation. For positive correlation, which is the correlation coefficient (CC) greater than 0.2 and smaller than 1, if the weather parameter increases, then the power output will also increase. With the CC closer to 1, this relationship becomes more significant. For negative correlation, which is the CC smaller than -0.2 and greater than -1. The relationship for negative correlation is the opposite of positive correlation, with the CC closer to -1 the more obvious this negative linear relationship goes. For small correlation, i.e. for CC between -0.2 to 0.2, the weather parameter and power output do not demonstrate a relationship.

1) Positive Correlation

Three parameters have the positive correlations with power output, they are: solar irradiance, temperature, and visibility shown in Fig. 8-10.

From Fig. 8, it is shown that solar irradiance and power output have close to a positive linear relationship. Typically, when solar irradiance is high, power output is also high. Low solar irradiance results in low power output.

In Fig. 9-10, when the temperature or visibility decreases, power output is likely to decrease; when they increase, power output will also tend to increase. Therefore, this positive CC relationship is observed in these three weather parameters.

The reason for this positive correlation is because high solar irradiance usually results in high temperature, and high visibility will ensure a large amount of solar irradiance is received by the solar panel. Since PV power output is highly dependent on solar irradiance, this relationship of temperature and visibility between power output can be explained.

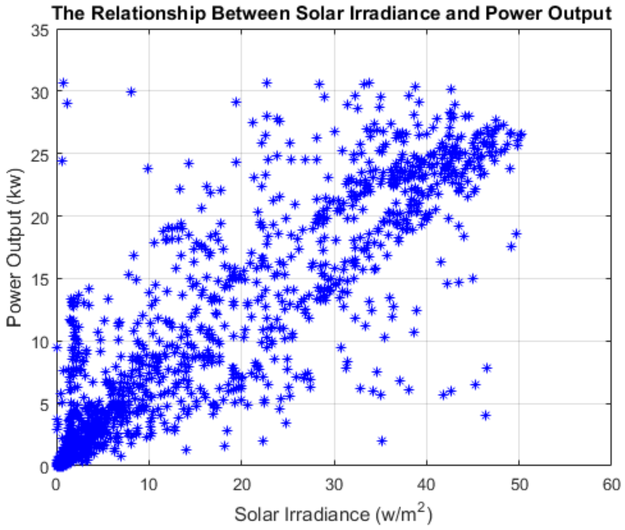

Figure 8 Relationship between solar irradiance and power output

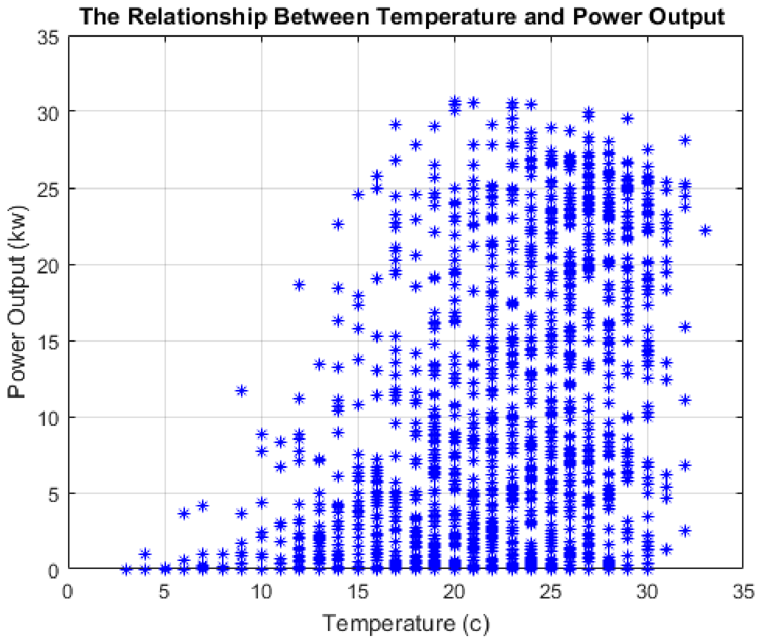

Figure 9 Relationship between temperature and power output

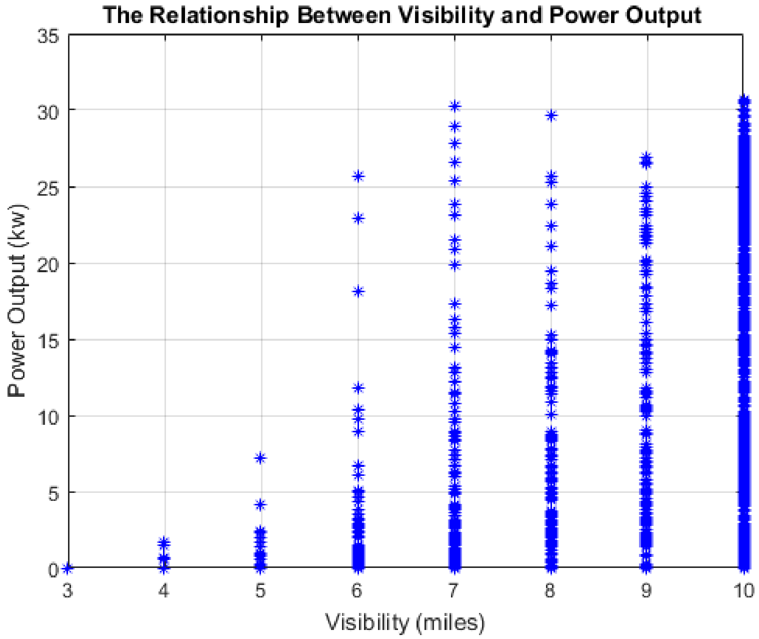

Figure 10 Relationship between visibility and power output

The CC between these parameters and power output are listed in Table 5:

Table 5 Coefficient values for positive correlation weather parameters

	Solar Irradiance	Temperature	Visibility
CC	0.86	0.41	0.46

From Table 5, the conclusion that solar irradiance is more related to power output than visibility can be drawn. In addition, visibility is more related to power output than temperature.

2) Negative Correlation

There are also three parameters having negative correlations and they are cloud cover, humidity, and weather code shown in Fig. 11-14.

In Fig. 11, it is shown that as the humidity increases, the power output tends to decrease. In addition, if humidity decreases, power output increases. Usually, high humidity happens in cloudy and rainy weather conditions, which have low solar irradiance. This explains why humidity has a negative correlation coefficient relationship with power output.

In Fig. 12, the weather condition of my data is divided into four categories: 1-sunny, 2-cloudy, 3-overcast, and 4-rainy. As seen from Figure 12, when the weather condition is overcast and rainy, the power output is low. This is due to the fact that solar irradiance is usually low during overcast and rainy conditions. Since power output is strongly dependent on solar irradiance, the power output is also low during these weather conditions.

Fig. 13 shows the relationship between cloud cover and power output. Due to the large amount of zero power output, which occur at night, this relationship does not seem so clear: no matter what is the cloud cover at night, the power output will always be zero. Therefore, I plot the relationship between these two during daytime. From Fig. 14, it is shown that this relationship is more obvious when cloud cover is low, power output is high and when cloud cover is high, power output is low. For cloud cover, high cloud cover will reject solar light reaching the solar panel, which will result in low power output.

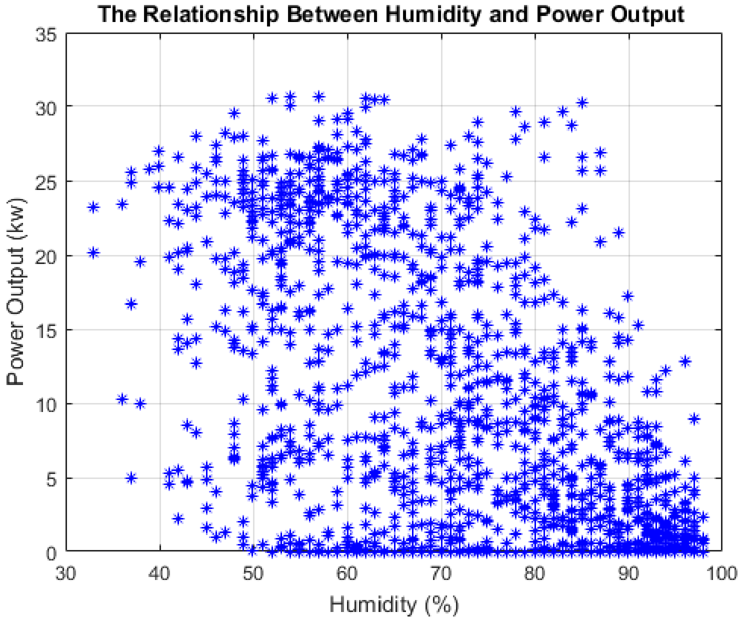

Figure 11 Relationship between humidity and power output

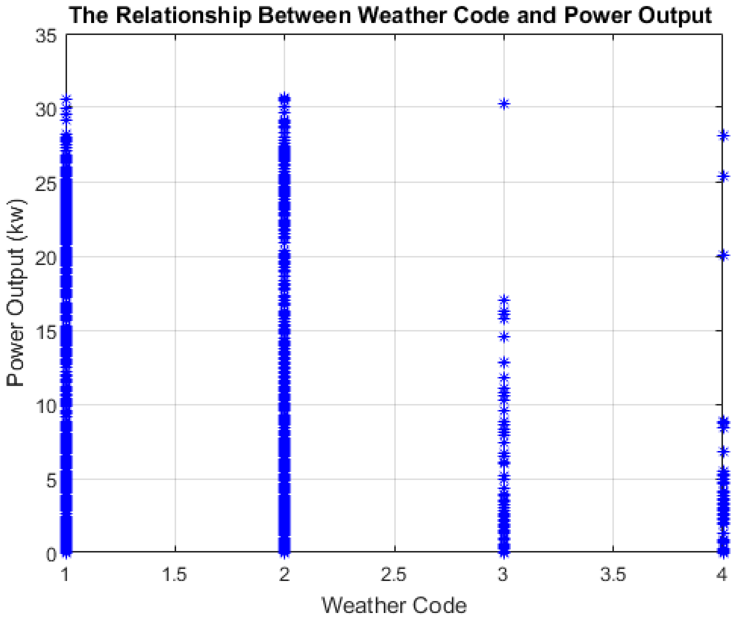

Figure 12 Relationship between weather code and power output

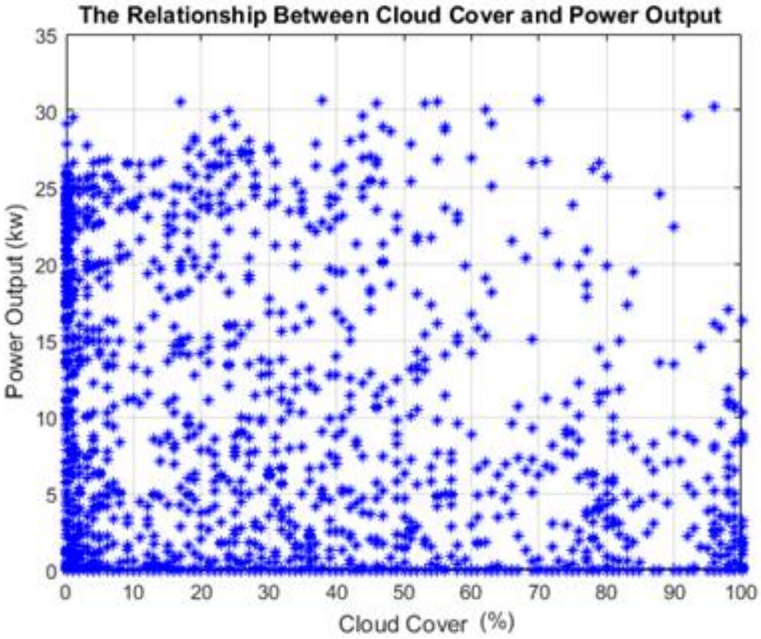

Figure 13 Relationship between cloud cover and power output

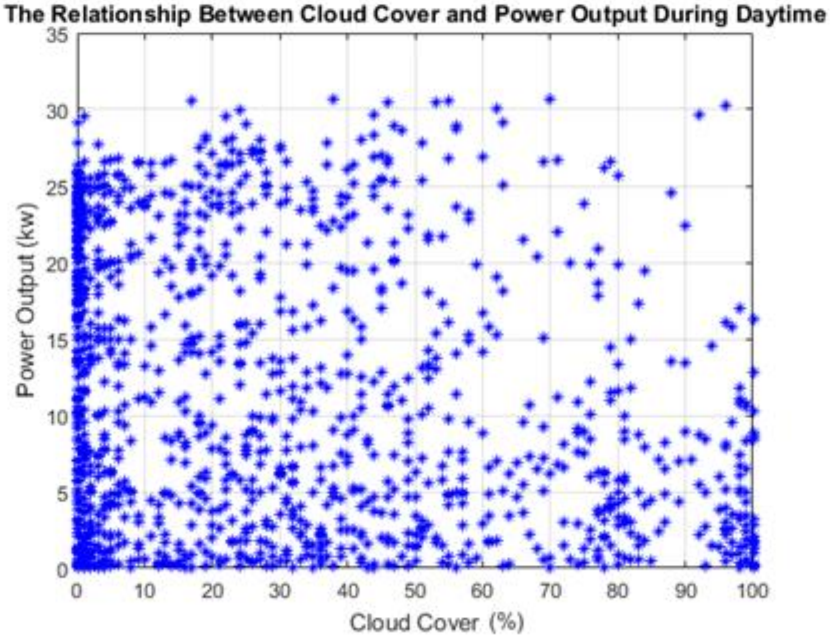

Figure 14 Relationship between cloud cover and power output during daytime

The CC between these negative CC parameters and power output are listed in Table 6:

Table 6 Coefficient values for negative correlation weather parameters

	Humidity	Weather Code	Cloud Cover
CC	-0.53	-0.24	-0.23

This time, if the CC is closer to -1, it means they have more negative correlation effect. As shown from Table 6, humidity has the most negative effect to power output, and the weather code comes the next.

3) Small Correlation

For small correlation, these parameters have very little correlation with the power output, and there are four parameters with the following relationship: time of the day, air pressure, wind speed, and wind direction shown in Fig. 15-18. However, a small correlation does not mean these parameters do not have any effect on solar power output. For example, in Fig. 15, it is shown that daytime and power output do not have positive or negative correlations, but they have non-linear relationship. Typically, noontime (from 12pm to 3pm) has the largest power output, and during nighttime (from 9pm to 5am), there will not be any power output at all. This relationship is due to factors such as the solar irradiance and the angle of the sun. This is very important in this thesis, because when I apply forecasting methods, time of the day is one of the most important parameter I consider. If two time periods have the exact same temperature, weather code, cloud cover, and humidity but one is at noon and the other is at the midnight, they will surely have different power outputs.

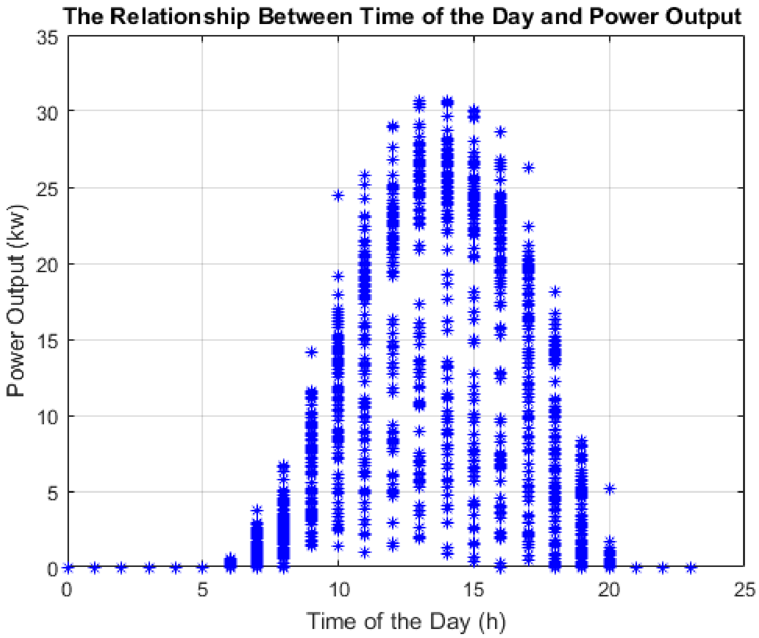

Figure 15 Relationship between time of the day and power output

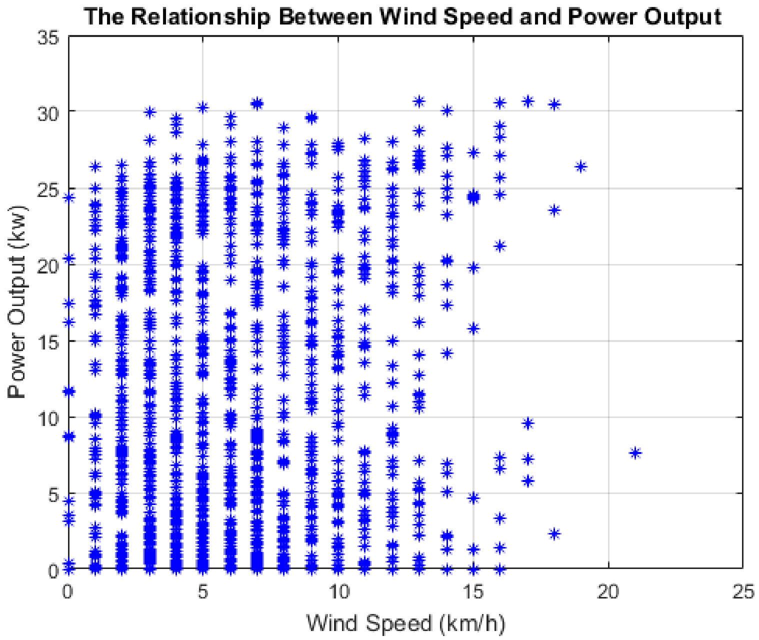

Figure 16 Relationship between wind speed and power output

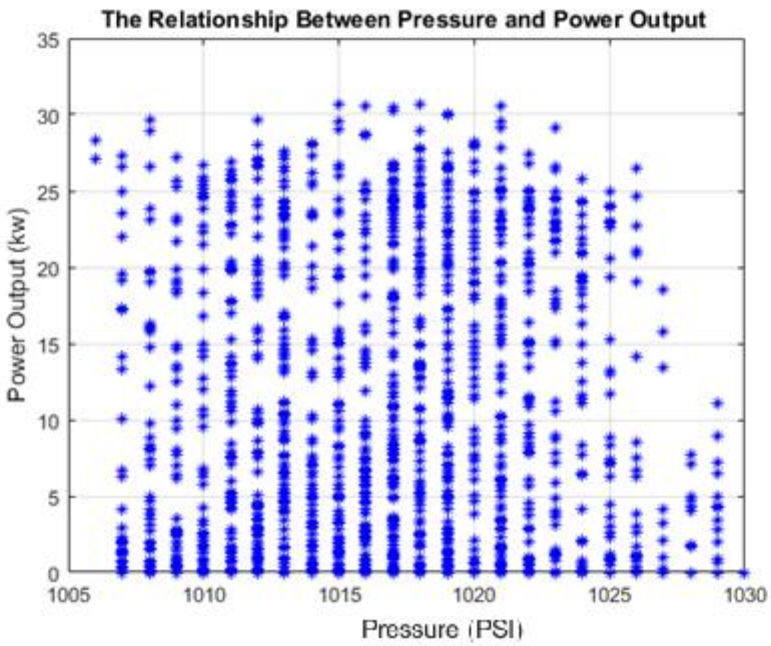

Figure 17 Relationship between pressure and power output

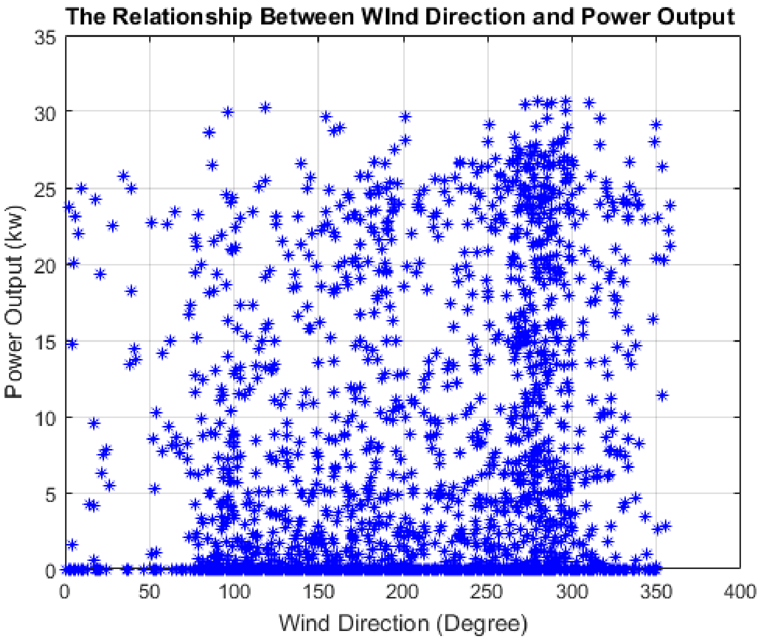

Figure 18 Relationship between wind direction and power output

As seen in Fig. 16-18, air pressure, wind speed and wind direction have very little relationship with power output. For example, power output can be high or low during either high or low wind speed. Same applies to wind direction and air pressure, which means there is not much connection between power output and the aforementioned parameters.

The CC of small correlation coefficient parameters are listed in Table 7 below:

Table 7 Coefficient values for small correlation weather parameters

	Time of the Day	Wind Speed	Pressure	Wind Degree
CC	0.16	0.11	0.03	0.19

As shown in the Table 7, the coefficients are all very small. However, this does not mean these parameters are not important. As mentioned above, time of the day is one of the most important parameters I use to forecast, due to the fact that time of the day has very small linear coefficient but high non-linear coefficient.

The higher the absolute value of CC is, the more important the parameter is. Therefore, these parameters can be sorted by their importance:

solar irradiance > humidity > visibility > temperature > weather code > cloud cover > wind degree > wind speed > pressure

Time of the day is not sorted because it has a very important non-linear relationship.

This means that power output is also dependent on weather parameters, so it is important to consider these parameters in my forecast methods.

3.1.2 Dramatic Change in Power Output

The second challenge is that solar power output is intermittent in nature. This intermittency might be caused by floating clouds and rainstorms. As shown in Fig. 19, power output can be significantly different on adjacent days. In addition, in August 22nd, power output changed dramatically in adjacent hours. All these sudden changes make PV output forecasting a difficult task. To solve this problem, I added method such as feedback to my methods. For example, during KNN forecasting, if the actual power output value and the forecasted value during 11:00 has a huge difference, since it is three hours ahead forecasting, I will use data from other closest weather condition to forecast 14:00. In addition, an accurate weather forecast can be important to solve this kind of problem.

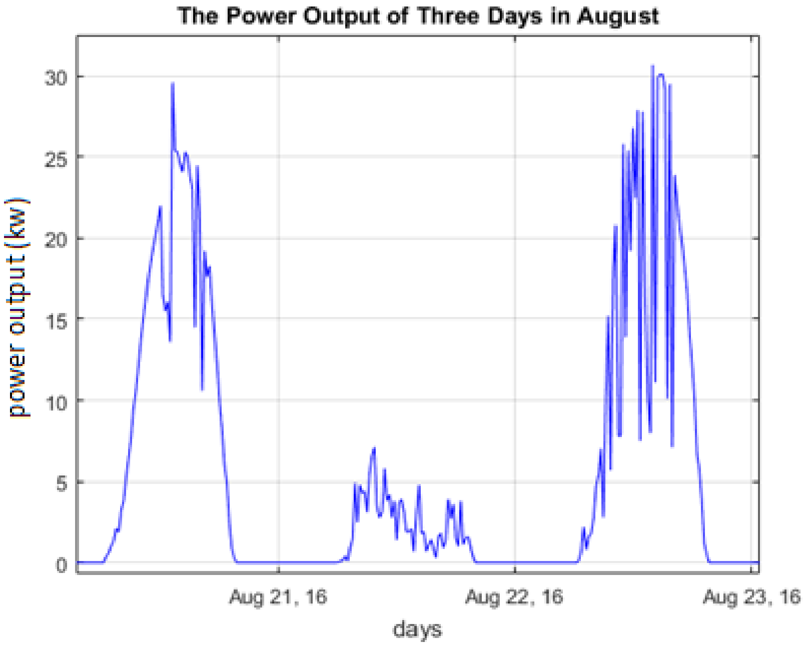

Figure 19 Power output of three days

3.1.3 Forecasting During Cloudy or Rainy Weather Conditions

For PV forecasting, it is much harder to forecast during cloudy or rainy conditions than sunny weather conditions. This is extremely critical to the Syracuse area due to the lake effect. In Syracuse, there are many rainy and snowy days, and it is estimated that it only has about 163 sunny days per year. On the other hand, snowfall will reach up to 124 inches per year[36]. To solve this, I used classification methods to divide the data into different weather conditions to have a more accurate forecast. The comparison between weather of Syracuse and the average United States is listed in Table 8. As seen from Table 8, Syracuse has much more rain and snow than the average of U.S. UV index represents the solar irradiance and it is lower than the average value of U.S.

Table 8 Comparison between Syracuse and the average value of United States

Climate	Syracuse	United States
Rainfall (in.)	43.4	39.2
Snowfall (in.)	123.8	25.8
Precipitation Days	171	102
Sunny Days	163	205
UV Index	3.2	4.3
Avg. July High Temperature	81.6	86.1
Avg. Jan. Low Temperature	16.7	22.6

3.1.4 The Size of Datasets

For PV forecasting, a data size that is too small can cause inaccuracy, while a data size that is too large might reduce calculation speed. In this thesis, I used different forecasting methods for different size of data. These methods can ensure the accuracy of a particular size of the data.

3.2 PV Forecasting Methods

In this thesis, I used five different forecasting methods to predict PV output. They are Artificial Neural Network (ANN), Linear Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Symbolic Aggregate approximation (SAX).

The widely-used ANN forecasting method is used as the baseline to prove the accuracy of my methods. The ANN method is a way to process data that acts like human brain, and its operating function is shown in Fig. 20 [8]:

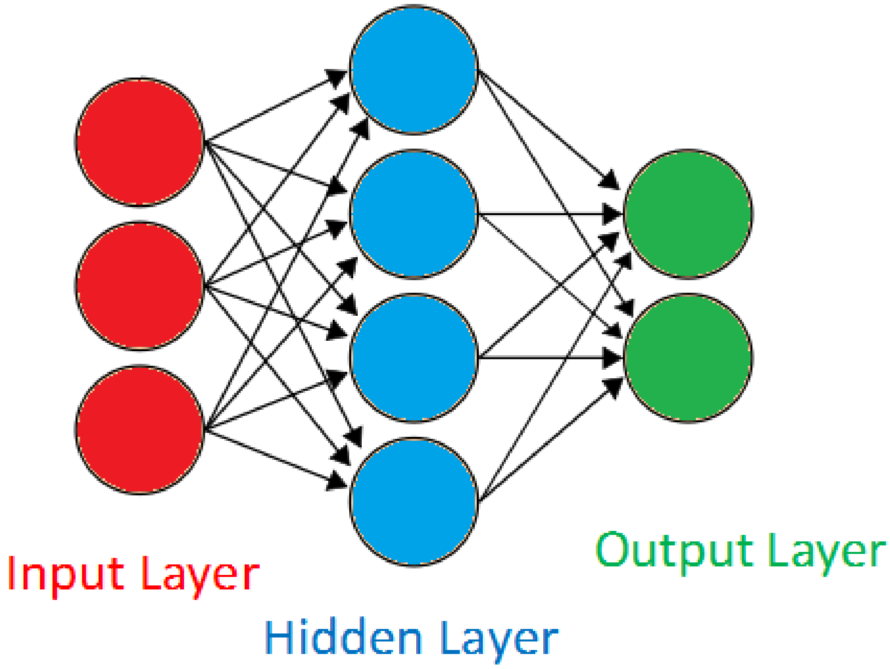

Figure 20 Function of ANN

There are three types of layers in the ANN method: the input layer, the hidden layer and the output layer. When I input the data into the input layer, it randomly assigns weights to the linkages to start the calculation. After that, by finding the linkages between input layers and hidden layers, it finds the activation function of hidden layers. In addition, by finding the linkages between hidden layers and output layers, the activation function of output layers can be determined. Activation function is the output of a node using an input. Later, the system calculates the error ratio and recalculates the linkages between hidden layers and output layers. By doing this, it decreases the value of error from hidden layers. In the next step, is the system recalculates the linkages between input layers and hidden layers to reduce the value of errors. The system repeats the process until the criterion is met, usually when the value of errors is less than a certain amount. The way the ANN method operates is difficult to express in a traditional computer program because it is self-learning and self-training. For traditional computer program, it can only learn by doing different steps in an algorithm.

The Neural Time Series forecast in Matlab is used to perform ANN forecast. Nonlinear Autoregressive with External Input (NARX) combining with Levenberg-Marquardt algorithm is used to train the data [8],[10]. In additional, I used 9 hidden neurons in my forecasting system. Figure 21 shows the ANN system I used to forecast. As shown in the figure, time delay is 3-hours and the hidden layer is 9. The system has 10 additional inputs besides the power output before the forecast time and the only output is the forecasting power output. These 10 additional inputs are the ten weather parameters mentioned in section 3.1.

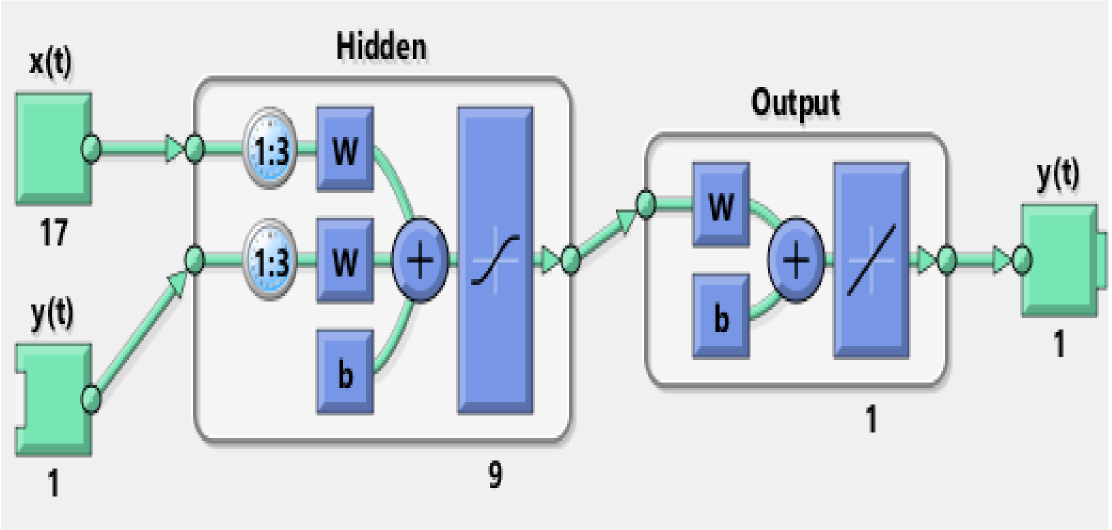

Figure 21 ANN system in 3-hours ahead forecasting

Another widely used method is SVM. SVM is a rather new method, which has a two-stage architecture. Similar to ANN, SVM also uses neural network to analyze the data. First, the space is separated into several disjointed regions using self-organizing map (SOM) architecture. Second, the best kernel function is found to construct the most suitable regions. However, this method did not work well with the dataset used for prediction. Hence, its results are not included in this thesis.

3.3 PV Forecasting Based on Different Data Sizes

In this section, different forecasting methods are compared with three different data sizes. The goal was to see which method performs best in a particular data size.

3.3.1 Small Data Size Forecasting

First, different forecasting methods are compared using a rather small size of data. The data contained only one month of hourly data from September 2016 to forecast seven days in October 2016. The goal of was to have an accurate forecast right after solar generation facilities have been installed.

There are three parts in my dataset. The first part is weather parameter, which includes the parameters introduced in section 3.1.1. In addition, the forecasted precipitant, lowest temperature, highest temperature, weather code, wind speed, and direction are also provided for the future five days. However, the farther ahead the forecast is, the less accurate the forecasted data is. The second part is solar irradiance, and its instantaneous, average, minimum, and maximum values are provided. The last part is power output, and the output of the system's PV generator is given.

After comparing different methods with the actual power output, linear regression and its related methods such as M5 Model Tree (M5P) and M5 Rule is found to work best in small data size forecasting. On the other hand, methods such as KNN and SVM are relatively less suitable. This is due to the fact that linear regression tends to find the function between power output and weather parameters, and this function is not affected as much by the size of the data. For example, in my case:

$$\textit{Power output} = 0.55 * \textit{solar irradiance} + 0.07 * \textit{humidity} + 0.09 * \textit{pressure} + 0.62 * \textit{visibility} - 101.26 \tag{3.1}$$

M5P and M5 Rule are methods related to linear regression. Instead of forecasting the whole dataset with only one function, M5P separates the data using trees and M5Rule separates the data using rules. In addition, their forecast is more accurate than linear regression due to these separations. The results are shown in Fig. 22 and 23. As shown in Fig. 22, liner regression forecast works better than the ANN method. However, the drawback of this method is that its peak value is not high enough. This is due to the fact that it tends to forecast the whole value of a day using just one function. However, most of the time, using only one function is not enough. M5P, on the other hand, does not have this problem as shown in Fig. 22. This is because M5P separates the data prior to calculating the forecast value. Sometimes, there might be high humidity, air pressure and visibility but a low power output. This makes the parameter in Equation (3.1) too low to reach the peak value during noon time. The detailed figure for this insufficient for peak value is shown in Fig. 24. As it is shown, the peak value of M5P method is closer to the actual peak value than linear regression method.

Table 9 shows the RMSE values and MAE values for different methods. The smaller the errors the better, and it matches my conclusion that M5P operates the best while dealing with a small dataset, and linear regression, M5P and M5 Rule are more accurate than traditional methods such as ANN and SVM.

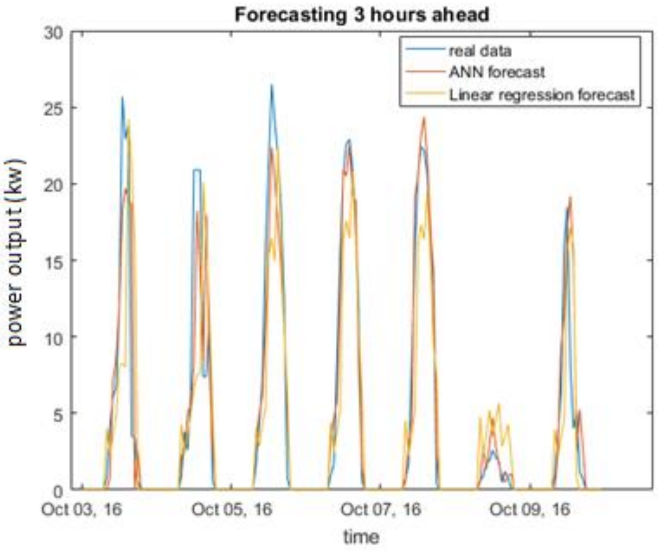

Figure 22 Comparsion between ANN and Linear Regression

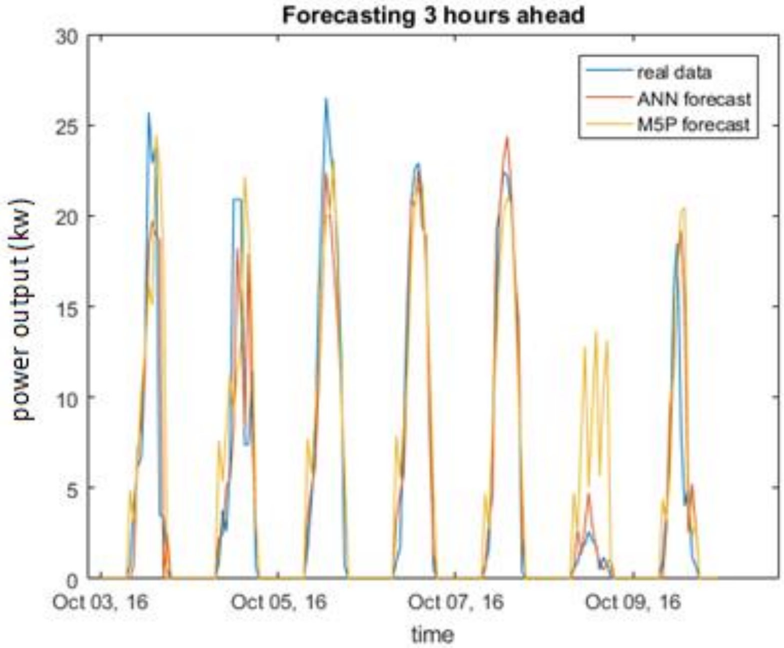

Figure 23 Comparison between ANN and M5P

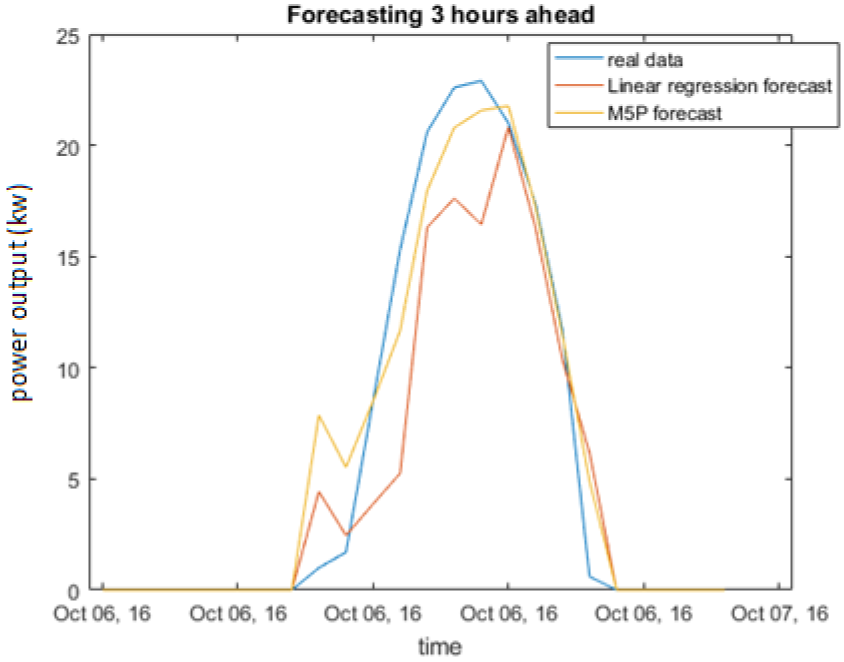

Figure 24 Comparison between peak value for Linear Regression and M5P

Table 9 Errors for small data size forecasting

	RMSE	MAE
Linear Regression	4.41	2.21
Artificial Neural Network	4.57	2.33
K-Nearest Neighbor	4.69	2.01
M5P	4.19	2.02
M5Rule	4.25	2.08
Support Vector Machine	5.29	2.62

3.3.2 Medium Data Size Forecasting

Next, the power output of a relatively larger dataset is forecasted, which contains three months of hourly data from the summer 2016 to forecast seven days in late August. This time, by having more data I found that KNN is the most suitable method. In addition, the accuracy of KNN increased dramatically compared to the small data size forecast.

Compared to methods like ANN, KNN is a lazy forecasting method, which means instead of finding the forecasting pattern, it tends to find out the closest data from the historical database. The principle of this method is to find the most similar weather parameter with the time I want to forecast and use it as forecasted power output. The steps are outlined as follows:

(1) I calculated the CC of different weather parameters with the power output, results are shown in Table I-III in chapter II.

(2) I normalized the CC by using the following equation:

$$CN_i = \frac{|CC_i|}{\sum |CC_i|} \qquad (3.2)$$

From equation (3.2), *CC$_i$* represents the CC of the *ith* parameter.

(3) I calculate the weather difference of current time period *i* and find the closest weather pattern to that time period *j* using the following equations (3.3) and (3.4):

$$D_{aj} = \left| a_i - a_j \right| \bullet CN_a \quad (j=1,2, \ldots , i\text{-}1, i\text{-}2, \ldots n) \tag{3.3}$$

Where D_{aj} represents the difference in parameter *a* and a_j represents the value of parameter *a* at time *j*.

$$D_j = \sum D_a \tag{3.4}$$

Here, D_j stands for the weather difference at time *j*.

(4) The *jth* time found must have same weather pattern and time of the day as time *i* or else the second largest D_j is found and the method is continued until the condition is satisfied.

(5) Lastly, I found the D_j from step 4 and took it as my reference time. At 3-hours ahead forecasting, I used the power output of 3-hours ahead of time *j* as the forecasted output.

The advantage of the KNN method is that it is fast and requires very few memories. The disadvantage is that it needs a rather large dataset to forecast. The larger the dataset, the more accurate the forecast is. This conclusion matches my result: *KNN in a medium data size is much more accurate than in a small data size.*

Our results are shown in Fig. 25, where I forecast August 25-31 using three months of summer data. Fig. 26 also shows the PV forecast for one day.

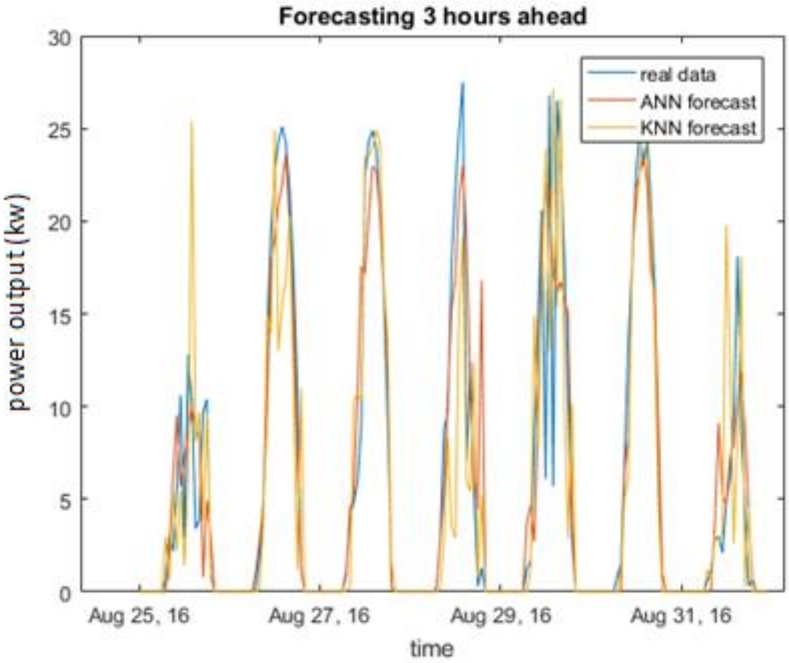

Figure 25 Comparison between ANN and KNN

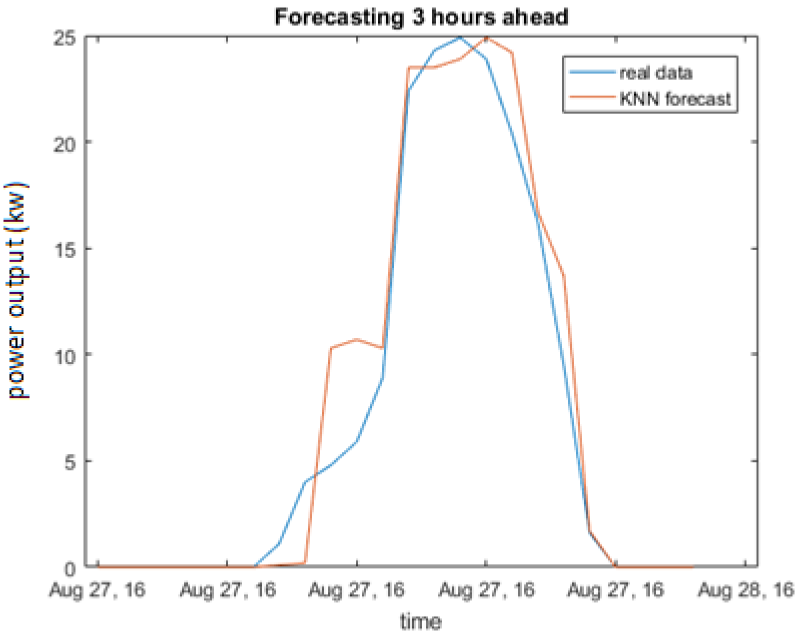

Figure 26 Results for one day using KNN

As is shown in the Fig. 25 and 26, KNN can successfully forecast the power output most of the time. It can also perform accurately during rainy or cloudy weather conditions, which is when there is dramatic change in power output. However, there are some errors due to lack of historical data. For example, in August 25, around 14:00, the actual data should be very low, however, the forecast value is relatively high. As mentioned previously, the larger the data, the more accurate the forecast is. The errors of KNN method are shown in Table 10.

Table 10 Errors for medium data size forecasting

	RMSE	MAE
Artificial Neural Network	4.47	2.25
K-Nearest Neighbor	3.85	1.98

The errors of other methods did not improve much, therefore those method are not listed in the table. In addition, as I compare Table 9 and 10, the RMSE value for KNN improved from 4.69 to 3.85, and the MAE value also improved from 2.01 to 1.98. After analyzing, I found out it is better to use KNN for medium data size forecasting.

3.3.3 Large Data Size Forecasting

In this section, a large dataset was used to find out the best method for PV forecasting. Instead of using hourly data, three months of minutely data was used to forecast seven days in late August. I found out that although SAX loses accuracy to some extent, it is much faster in calculation speed.

The goal of SAX was to reduce calculation time and dimensionality of the data. Here are the following steps for this method:

1) I normalized the data into a normal distribution by using the following equation:

$$b = (a - \mu)/\sigma \qquad (3.5)$$

Where a stands for the original data, μ stands for the average value of the data and σ stands for the derivate value of the data and b stands for the normalized data I want to get.

2) I transformed the data into a Piecewise Aggregate Approximation (PAA) string using the following equation [18]:

$$t_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} s_j \qquad (3.6)$$

In equation (3.6), t is the time series element, n is the length of the data, and transform the length of the data into w dimensional space. Figure 27 shows the PAA string of power output. As shown in the figure, the power output data can be successfully represented by a rather small dimensional PAA string.

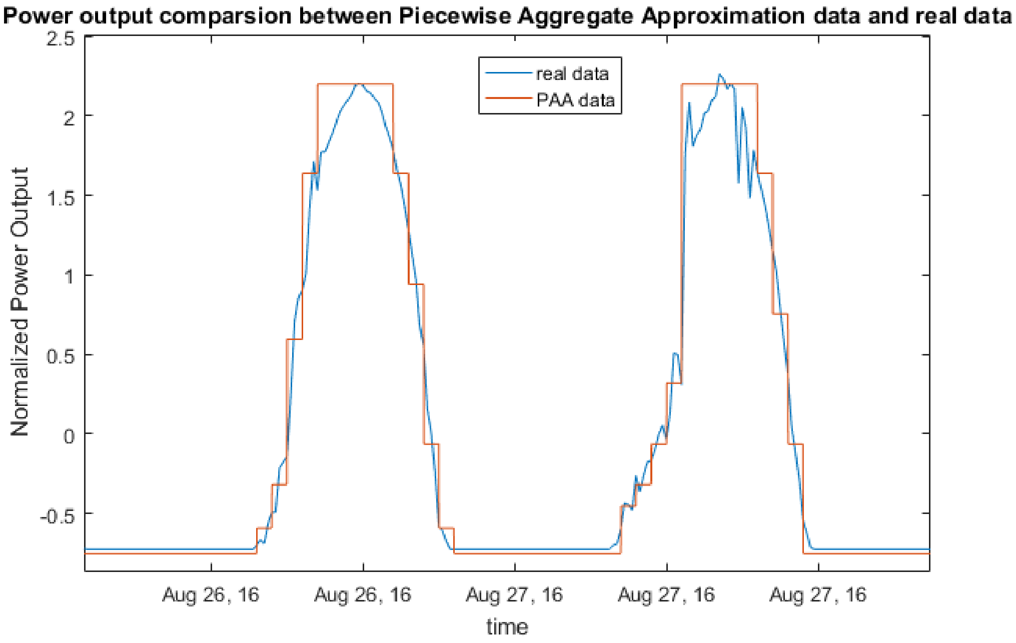

Figure 27 PAA string of power output

3) I use methods, such as linear regression or KNN, to forecast the string.

4) Lastly, I compared the data with my actual data and calculated the error.

In this paper, I used SAX combined with the KNN method to forecast during large dataset, and the results are shown in Fig. 28-30. In Fig. 28, only the KNN method is used to forecast the power output, and due to the large amount of data, it takes relatively a long processing time to forecast.

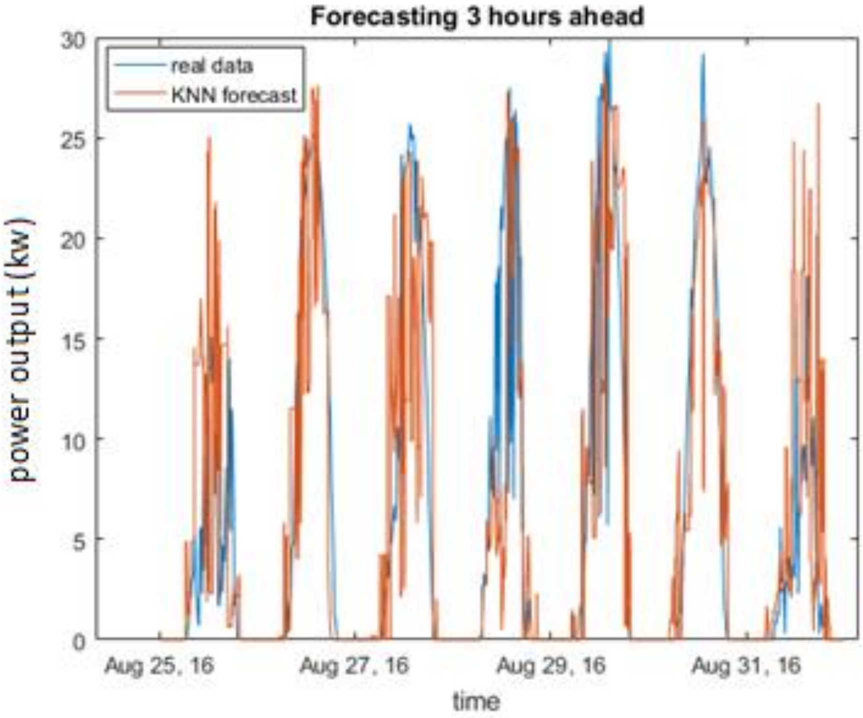

Figure 28 Forecasting large dataset using KNN

Fig. 29 shows the result of combining KNN with SAX. As I can see, compared to Fig. 28, the dimension of Fig. 29 has successfully reduced. In addition, the calculating speed has also increased a lot.

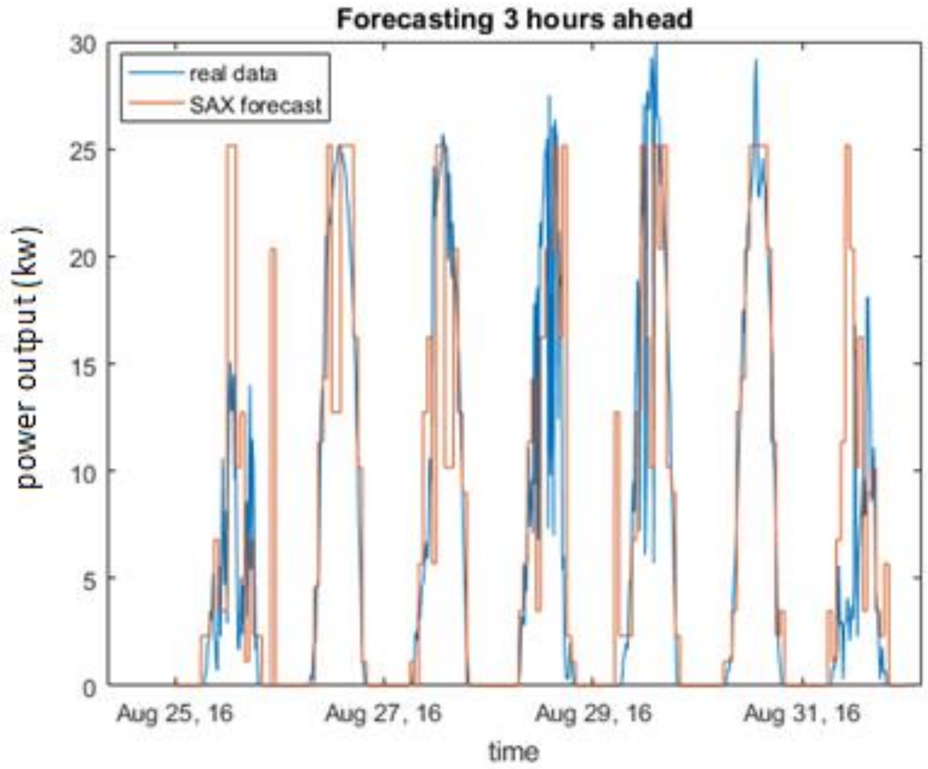

Figure 29 Forecasting large dataset using SAX+KNN

Figure 30 shows the results for only August 26th and 27th. As shown in the figure, the SAX+KNN method can successfully forecast the power output most of the time. There might be some error caused by two reasons: First, small transforming error when converted to the PAA string; Second, the KNN method did not find the accurate weather condition closest to the forecasted one. However, compared to improvement in calculation speed and data dimensionality, this error can be neglected. In addition, the comparison between the two methods are shown in Table 11. The error of KNN+SAX is slightly larger than only KNN, however, the calculation speed of this method increased by about 65 times. In addition, when I converted the data into a PAA string, I made *n=60,* which shown in dimensionality, is 60 times less than KNN+SAX. The error of the KNN method when using large dataset is slightly higher than the medium dataset in section

3.3.2. This is because when I converted the hourly data to minutely data, this conversion might cause some error, which will result in higher error during forecast.

I also used other forecasting methods that also used SAX to forecast the power output, such as the SAX+SVM. However, the results are not as accurate as the SAX+KNN method. Hence, I did not include these methods in this thesis.

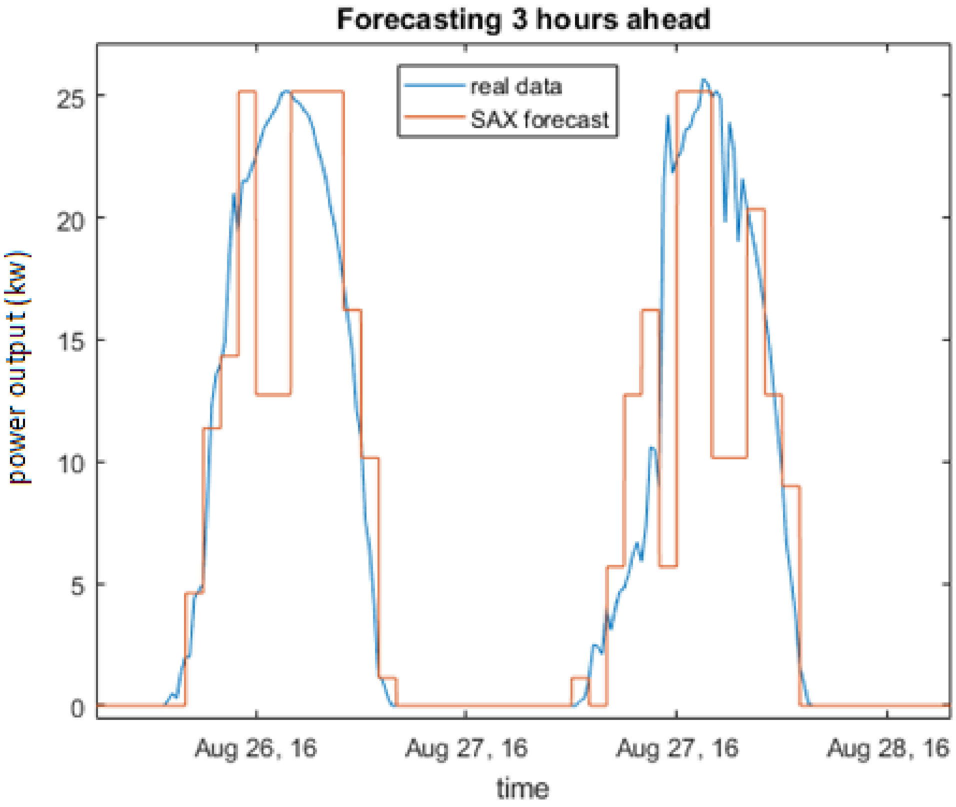

Figure 30 Results for two days using SAX+KNN

Table 11 The comparison of large data size forecasting methods

	MSE	RMSE	Calculation Time (s)	dimensionality
KNN	2.11	3.98	345.96	139680
KNN+SAX	2.19	4.05	5.36	2328

3.3.4 Conclusion

This section presents three different forecasting methods, which include M5P, KNN and KNN+SAX, during small, medium and large data sizes. During small data size, M5P can successfully forecast PV power output. The error of this method is the smallest of all. During medium data size, KNN is the most accurate method. In addition, the accuracy of this method increases dramatically as the data size increases. When I have very large data, KNN alone can take too much time to calculate the forecasting power output, thus the SAX+KNN method is the most suitable one. This method can reduce the speed dramatically according to the value of *n* I set while performing the SAX method.

The forecast for clear sky is much easier than rainy weather conditions and linear regression and M5P do not work accurately during rainy conditions. KNN and SAX+KNN, on the other hand, are much more accurate if I have the historical data for these weather conditions. Therefore, the KNN or SAX+KNN method if the data size is large enough is strongly recommended. In addition, the widely used SVM is not as accurate for my case. ANN performs better than SVM but not as well as the three methods mentioned above.

CHAPTER 4
CONCLUSION & FUTURE WORK

The focus of this study was to determine the relationship between the accuracy of PV forecasting and system planning, and to find the best performing forecasting method while having different data sizes. First, I showed the importance of accurate forecasting by calculating the probabilistic power flow. I found out that Beta distribution can more accurately represent the PDF of PV output. In addition, the transmission lines' sensitivity to accurate and inaccurate forecast depends on the location of a PV system and an accurate forecast is important to sensitive lines. Next, PV power output was forecasted by machine learning methods. This time, I showed that for small data sizes, M5P is the most accurate method. For medium data sizes, KNN is more suitable. In addition, for large data sizes, KNN+SAX can provide an accurate forecast and reduce calculation speed and dimensionality. The future work contains forecast the PV output during the winter season in the Syracuse area. In addition, I tend to see as the dataset grows even larger, how KNN and KNN+SAX will perform. I will make a comparison between probabilistic vs. conventional power flow with forecasted data, which would provide more insights on PV implementation impact on the grid.

GLOSSARY

ANN	Artificial Neural Network	2
BP	Back-propagation	2
CDF	Cumulative Distribution Function	2
CC	Correlation coefficient	22
FNN	Feed-forward neural network	2
KNN	K-Nearest Neighbor	3
MAE	Mean-absolute-error	4
M5P	M5 Model tree	37
NREL	National Renewable Energy Laboratory	1
NARX	Nonlinear Autoregressive with External Input	35
PV	Photovoltaic	1
PLF	Probabilistic Load Flow	2
PDF	Probabilistic Density Function	2
PAA	Piecewise Aggregate Approximation	43
PPF	Probabilistic power flow	1
RBF	Radial Basis Function	3

RMSE	Root-mean-square-error	4
SLF	Stochastic Load Flow	2
SOM	Self-organized map	2
SVM	Support vector machine	2
SAX	Symbolic Aggregate approximation	3

APPENDIX

Here, I listed the IEEE 30 Bus system I used in this thesis.

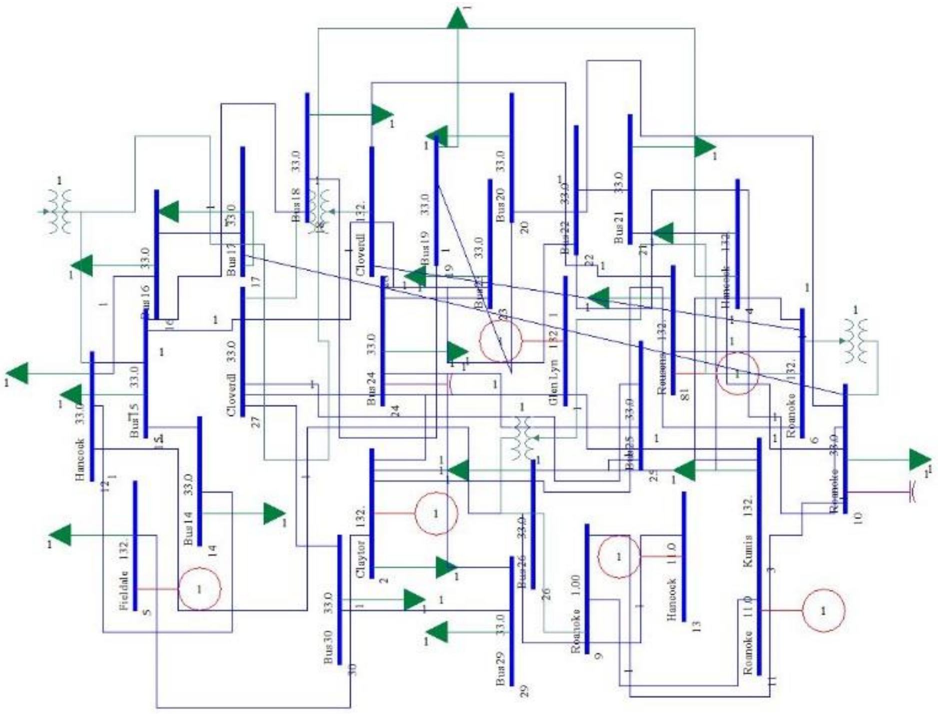

Figure 31 IEEE 30 bus system

Table 12 Generation locations

Bus	Generation Input (MW)
1	131
2	45
3	25
4	30
5	25
6	30

Table 13 The relationship between lines and buses

Line	From Bus	To Bus	Line	From Bus	To Bus
1	1	2	20	12	14
2	1	3	21	12	15
3	2	4	22	12	16
4	2	5	23	14	15
5	2	6	24	15	18
6	3	4	25	15	23
7	4	6	26	16	17
8	5	7	27	18	19
9	6	7	28	19	20
10	6	8	29	21	22
11	6	28	30	22	24
12	8	28	31	23	24
13	9	10	32	24	25
14	9	11	33	25	26
15	10	17	34	25	27
16	10	20	35	27	29
17	10	21	36	27	30
18	10	22	37	29	30
19	12	13			

REFERENCE

[1] *2013 Renewable Energy Data Book*. 2014.

[2] *NREL: Energy Analysis - Renewable Electricity Futures Study*. 2016.

[3] B. Borkowska, “Probabilistic load flow,” *IEEE Trans. Power Appar. Syst.*, no. 3, pp. 752–759, 1974.

[4] G. T. Heydt and B. M. Katz, “A stochastic model in simultaneous interchange capacity calculations,” *IEEE Trans. Power Appar. Syst.*, vol. 94, no. 2, pp. 350–359, 1975.

[5] R. N. Allan, B. Borkowska, and C. H. Grigg, “Probabilistic analysis of power flows,” in *Proceedings of the Institution of Electrical Engineers*, 1974, vol. 121, pp. 1551–1556.

[6] J. M. Morales and J. Perez-Ruiz, “Point estimate schemes to solve the probabilistic power flow,” *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1594–1601, 2007.

[7] M. Fan, V. Vittal, G. T. Heydt, and R. Ayyanar, “Probabilistic power flow studies for transmission systems with photovoltaic generation using cumulants,” *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 2251–2261, 2012.

[8] A. Mellit and A. M. Pavan, “A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy,” *Sol. Energy*, vol. 84, no. 5, pp. 807–821, 2010.

[9] M. Ding, L. Wang, and R. Bi, “An ANN-based approach for forecasting the power output of photovoltaic system,” *Procedia Environ. Sci.*, vol. 11, pp. 1308–1315, 2011.

[10] C. Chen, S. Duan, T. Cai, and B. Liu, “Online 24-h solar power forecasting based on weather type classification using artificial neural network,” *Sol. Energy*, vol. 85, no. 11, pp. 2856–2870, 2011.

[11] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, “Predicting solar generation from weather forecasts using machine learning,” in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*, 2011, pp. 528–533.

[12] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, “Forecasting power output of photovoltaic systems based on weather classification and support vector machines,” *IEEE Trans. Ind. Appl.*, vol. 48, no. 3, pp. 1064–1069, 2012.

[13] R. Xu, H. Chen, and X. Sun, “Short-term photovoltaic power forecasting with weighted support vector machine,” in *Automation and Logistics (ICAL), 2012 IEEE International Conference on*, 2012, pp. 248–253.

[14] H. Mori and A. Takahashi, “A data mining method for selecting input variables for forecasting model of global solar radiation,” in *Transmission and Distribution Conference and Exposition (T&D), 2012 IEEE PES*, 2012, pp. 1–6.

[15] M.-C. Kang, J.-M. Sohn, J. Park, S.-K. Lee, and Y.-T. Yoon, “Development of algorithm for day ahead PV generation forecasting using data mining method,” in *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on*, 2011, pp. 1–4.

[16] P. Mandal, S. T. S. Madhira, J. Meng, and R. L. Pineda, “Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques,” *Procedia Comput. Sci.*, vol. 12, pp. 332–337, 2012.

[17] J. Lonardi and P. Patel, “Finding motifs in time series,” in *Proc. of the 2nd Workshop on Temporal Data Mining*, 2002, pp. 53–68.

[18] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003, pp. 2–11.

[19] M. J. E. Alam, K. M. Muttaqi, and D. Sutanto, “A sax-based advanced computational tool for assessment of clustered rooftop solar pv impacts on lv and mv networks in smart grid,” *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 577–585, 2013.

[20] Y. Huang, J. Lu, C. Liu, X. Xu, W. Wang, and X. Zhou, “Comparative study of power forecasting methods for PV stations,” in *Power System Technology (POWERCON), 2010 International Conference on*, 2010, pp. 1–6.

[21] F. Nomiyama, J. Asai, T. Murakami, and J. Murata, “A study on global solar radiation forecasting using weather forecast data,” in *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on*, 2011, pp. 1–4.

[22] P. Bacher, H. Madsen, and H. A. Nielsen, “Online short-term solar power forecasting,” *Sol. Energy*, vol. 83, no. 10, pp. 1772–1783, 2009.

[23] I. Abadi and A. Soeprijanto, “Extreme learning machine approach to estimate hourly solar radiation on horizontal surface (PV) in Surabaya-East java,” in *Information Technology, Computer and Electrical Engineering (ICITACEE), 2014 1st International Conference on*, 2014, pp. 372–376.

[24] Y. Zhang, M. Beaudin, R. Taheri, H. Zareipour, and D. Wood, “Day-ahead power output forecasting for small-scale solar photovoltaic electricity generators,” *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2253–2262, 2015.

[25] C. V. Silva, L. Lim, D. Stevens, and D. Nakafuji, “Probabilistic Models for One-Day Ahead Solar Irradiance Forecasting in Renewable Energy Applications,” in *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*, 2015, pp. 1163–1168.

[26] P. Zhang and S. T. Lee, “Probabilistic load flow computation using the method of combined cumulants and Gram-Charlier expansion,” *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 676–682, 2004.

[27] R. N. Allan and M. R. G. Al-Shakarchi, “Probabilistic techniques in ac load-flow analysis,” in *Proceedings of the Institution of Electrical Engineers*, 1977, vol. 124, pp. 154–160.

[28] G. Casella and R. L. Berger, *Statistical inference*, vol. 2. Duxbury Pacific Grove, CA, 2002.

[29] H. Cramér, *Mathematical methods of statistics*. JSTOR, 1947.

[30] S. H. Karaki, R. B. Chedid, and R. Ramadan, “Probabilistic performance assessment of autonomous solar-wind energy conversion systems,” *IEEE Trans. Energy Convers.*, vol. 14, no. 3, pp. 766–772, 1999.

[31] Z. M. Salameh, B. S. Borowy, and A. R. Amin, “Photovoltaic module-site matching based on the capacity factors,” *IEEE Trans. Energy Convers.*, vol. 10, no. 2, pp. 326–332, 1995.

[32] I. Abouzahr and R. Ramakumar, “Loss of power supply probability of stand-alone photovoltaic systems: a closed form solution approach,” *IEEE Trans. Energy Convers.*, vol. 6, no. 1, pp. 1–11, 1991.

[33] Y. M. Atwa, E. F. El-Saadany, M. M. A. Salama, and R. Seethapathy, “Optimal renewable resources mix for distribution system energy loss minimization,” *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 360–370, 2010.

[34] S. Wilcox, “National solar radiation database 1991-2005 update: User’s manual,” National Renewable Energy Laboratory (NREL), Golden, CO., 2007.

[35] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006.

[36] “Syracuse, New York Climate- Sperling’s Best Places.”

GUANGYUAN SHI's VITA

Major:

Electrical Engineering

Education:

M.S., May 2017, Syracuse University, USA (expected)

B.E., June 2015, North China Electric Power University, China

Publications & Posters:

Guangyuan Shi, Sara Eftekharnejad, Impact of Solar Forecasting on Power System planning, North American Power Symposium, University of Denver, 09/2016, Paper

Guangyuan Shi, Sara Eftekharnejad, Reforming the Energy Vision of New York State with Increased Renewable Generation, NEXT 2016 Conference, Syracuse, 10/2016, Poster

Guangyuan Shi, Varsha Govindraj, Nimotalahi Kareem, Demand Side Management in the Smart-Grid, 2016 ASEE St. Lawrence Section Conference, Cornell University, 04/2016, Poster

Professional Experience:

Research Assistant in SU Power System & Renewable Energy Lab, 01/2016-Present

Tangshan Douhe Power Plant, Tangshan, 09/2014