University of Calgary

PRISM Repository **https://prism.ucalgary.ca**

The Vault Open Theses and Dissertations

2013-12-16

Forecasting Photo-Voltaic Solar Power in Electricity Systems

Zhang, Yue

Zhang, Y. (2013). Forecasting Photo-Voltaic Solar Power in Electricity Systems (Master's thesis, University of Calgary, Calgary, Canada). Retrieved from https://prism.ucalgary.ca. doi:10.11575/PRISM/26210
http://hdl.handle.net/11023/1203
Downloaded from PRISM Repository, University of Calgary

UNIVERSITY OF CALGARY

Forecasting Photo-Voltaic Solar Power in Electricity Systems

by

Yue Zhang

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

CALGARY, ALBERTA

December, 2013

© Yue Zhang 2013

Abstract

This thesis concentrates on short-term solar Photovoltaic (PV) power output forecasting at array and system levels. The analysis was conducted on three arrays around the world and one system level system in California. Array level power was found to have higher power fluctuation than system level power. Hence, the proposed array level forecasting involves a similar day-based data-preprocessing to deal with this fluctuation. The processed array level data was fed into a forecasting engine. A persistence model, an Auto Regressive Integrated Moving Average (ARIMA), a Radial Basis Function Neural Network (RBFNN) and a Least Squares Support Vector Machine (LS-SVM) model was used as a forecasting engine. This thesis also investigates the applicability of a number of established forecasting methods for system level solar power output forecasting. In particular, ARIMA, RBFNN and LS-SVM are examined and simulation results are provided.

Through simulation, the best array level forecasting accuracy is achieved by a forecasting tool which combines the proposed similar day method and persistence model. The proposed similar day method works better than similar day methods in the literature and the best array level forecasting tool generated a more accurate forecast compared to a autoregressive with exogenous input (ARX) model in the literature. Due to the lower fluctuation of system level power data, system level forecasting has a better forecasting accuracy. The best performance is achieved through ARIMA model.

Acknowledgements

Firstly, I would like to express my deepest gratitude to Professor David Wood and Professor Hamidreza Zareipour for their invaluable support and encouragement. Their patient supervision of and guidance over my M.Sc study has been priceless. I have learned valuable lessons from their academic vision and professionalism.

I would also like to express my appreciation to Professor Edwin Nowicki for his kind and patient help, as well as Angela Morton in the Electrical and Computer Science Engineering Departments. My gratitude also goes to my friends from Hamidzrea's research team who gave me help and support: Arman, Hamid, Syed, Greg, Hamed, Marc, Ali, Behnam and Jorge. Special thanks to Dr. Jan from the University of San Diego and Dr. Peder from the Technical University of Denmark for sharing their PV data.

Also acknowledged is the unique research environment at the University of Calgary and the kind services and support from the Department of Electrical and Computer Science Engineering as well as from the Energy Environment Experiential Learning Project.

My gratitude also goes to NSERC and the ENMAX Corporation who provided renewable energy research funding under the Industrial Research Chairs program.

Last, but not least, I want to thank my family for their love and encouragement. My father, Guojun, has taught me the meaning of responsibility and hard-work. My mother, Mingyun, has given me endless support. Special thanks to my grandfather and my grandmother, who provided a free and happy childhood with their great love.

Dedication

This thesis is dedicated to my father, Guojun Zhang, my mother Mingyun Song, my grandfather, Xingqi Song, and my grandmother Guizhi Qu.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
ACF	Auto-Correlation Function
ARIMA	Auto Regressive Integrated Moving Average
DMI	Danish Meteorological Institute
ECMWF	European Centre for Medium-Range Weather Forecasts
GEM	Global Environmental Multi-scale Model
GFS	Global Forecast System
LS-SVM	Least Squares Support Vector Machine
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MBE	Mean Bias Error
MMP	Maximum Power Point
MMPT	Maximum Power Point Tracking

MOS	Model Output Statistics
NAM	North American Model
NDFD	National Digital Forecast Database
nMAE	Normalized Mean Absolute Error
NOAA	National Oceanic and Atmospheric Administrations
nRMSE	Normalized Root-Mean-Square Error
NWP	Numerical Weather Prediction
WRF	Weather and Research Forecasting
PACF	Partially Auto-Correlation Function
PV	Photovoltaic
RAMS	Regional Atmospheric Modeling System
RBFNN	Radial Basis Function Neural Network
KKT	Karush-Kuhn-Kucker

Variables	Definition
δ	Solar declination angle
θ_A	Solar altitude angle
θ_γ	Solar azimuth angle
θ_Z	Solar zenith angle
D	day of a year
ω	Hour angle
θ_φ	Longitude
θ_ϕ	Latitude
H_A	Altitude
θ_α	Surface azimuth angle
θ_β	Surface inclination angle
θ_S	Angle between the sun and the normal to the surface

T_E	Equation of time
I_{global}	Global solar radiation
I_{direct}	Direct solar radiation
$I_{diffuse}$	Diffuse solar radiation
$I_{reflected}$	Reflected solar radiation
I_E	Extraterrestrial solar radiation
η_{direct}	Atmospheric transmittance for direct solar radiation
$\eta_{diffuse}$	Atmospheric transmittance for diffuse solar radiation
$\eta_{reflected}$	Atmospheric transmittance for reflected solar radiation
τ_0	Climate factor 0
τ_1	Climate factor 1
τ_2	Climate factor 2
I	Terminal current of the cell

V	Terminal voltage of the cell
I_{PH}	Photo-current
I_0	Diode saturation current
q	Electron charge
k_B	Boltzmann's constant
T_C	Cell temperature
I_{SC}	Short-circuit current
V_{OC}	Open-circuit voltage
P_m	Power at maximum power point
I_m	Current at maximum power point
V_m	Voltage at maximum power point
R_{SH}	Series resistor
R_S	Parallel resistor

I_{01}	Current through diode 1
I_{02}	Current through diode 2
z_t	A stationary stochastic process
c	Constant parameter for ARIMA model
ϕ_i	Auto regressive parameters for ARIMA model
θj	Moving average parameters for ARIMA model
ϵ_t	Independently and identically distributed normal random variables
σ_ϵ^2	Variance of random variables ϵ_t
B	Backward shift operator
v_t	d^{th} order differenced process of z_t
$\phi_p(B)$	Non-seasonal auto regressive operator
p	Order of non-seasonal auto regressive operator
$\theta_q(B)$	Non-seasonal moving average operator

q	Order of non-seasonal moving average operator
d	Non-seasonal difference order
$\Phi_P(B^s)$	Seasonal auto regressive operator
P	Order of seasonal auto regressive operator
$\Theta_Q(B^s)$	Seasonal moving average operator
Q	Order of seasonal moving average operator
D	Seasonal difference order
$\psi(x)$	Non-linear mapping function of LS-SVM
$K(x, x_k)$	Kernel function of LS-SVM
$f(x)$	Non-linear regression function of LS-SVM
ω	Weight vector of non-linear regression function
b	Bias of non-linear regression function
$\min J(\omega, b, e)$	Optimization problem in LS-SVM

ζ	Adjustment parameter
e_k^2	Quadratic loss function
L	Lagrange function of LS-SVM
α_k	Lagrange multipliers
$H_j(x)$	Gaussian basis function
μ_j	Centre of the Gaussian basis function
σ_j	Smoothness parameter of the Gaussian basis function
P_t	Original hourly power data
P_t^D	The intra-day fluctuation for time t at day D
I_R	Recorded global horizontal irradiance
I_F	Forecast global horizontal irradiance
T_R	Recorded ambient temperature
T_F	Forecast ambient temperature

WS_R	Recorded wind speed
WS_F	Forecast wind speed
DNI_R	Recorded direct normal irradiance
DNI_F	Forecast direct normal irradiance
SA	Solar altitude
F_F	Forecast fog
LC_F	Forecast low cloud cover
MC_F	Forecast medium cloud cover
HC_F	Forecast high cloud cover
TC_F	Forecast total cloud cover
$\mathrm{ED}_{PR}(i, d)$	Euclidean distance of recorded PV power output for the forecast day d and previous day i
PR_i^h	Recorded power at hour h for day i
PR_d^h	Recorded power at hour h for day d

$\mathrm{ED}_V(i,d)$	Euclidean distance of candidate output-related variable V for the forecast day d and previous day i
V_i^h	Forecast value of output-related variable V at time h for day i
V_d^h	Forecast value of output-related variable V at time h for day d
EVA_V	Evaluation index for output-related variable V
ED_{PR}^N	Normalized value of ED_{PR}
ED_V^N	Normalized value of ED_V
S	Number of testing days
$\mathrm{ED}_B^N(i,d)$	Bivariate euclidean distance between day d and day i
W_{GHI}	Weight for global horizontal irradiance in bivariate euclidean distance
W_V	Weight for output-related variable V in bivariate euclidean distance
EVA_B	Evaluation index for bivariate euclidean distance
$\mathrm{ED}_{HD}^N(i,d)$	Hybrid euclidean distance between day d and day i

F_{GHI}	Weight for global horizontal irradiance in hybrid euclidean distance
F_V	Weight for output-related variable V in hybrid euclidean distance
$P_F(t,d)$	Forecasting value for the forecasting day d at hour t
$P_R(t,i)$	Recorded power output at day i at hour t
PM_j	A persistence model using j training days
P_C	The capacity of the PV site
$\mathrm{ED}_1(i,j)$	Euclidean distance between day i and day j calculated through Similar Day Method 1
$\mathrm{ED}_2(i,j)$	Euclidean distance between day i and day j calculated through Similar Day Method 2
IMP_{Mean}	The accuracy improvement based on average forecasting error
IMP_{Std}	The improvement based on standard deviation of forecasting error

Chapter 1

Introduction

1.1 Research Motivation

Driven by a drop in price and an increase in efficiency, photovoltaic (PV) electricity generation is growing rapidly. The average annual growth rate has been 40% over the past decade, and this growth is expected to continue in the future [1]. According to the European Photovoltaic Industry Association [2], PV is the third most important renewable energy source and there were 31.1 GW of newly installed PV capacity in 2012. The global annual market is expected to reach 84 GW in 2017.

The price of PV modules has decreased consistently over the past three decades. The price has fallen 19.3% with every doubling of installed PV capacity [3]. Although other cost components of PV systems, such as inverters and installation, are not falling as rapidly as the module costs, the overall cost of generating electricity from PV systems is expected to drop. The costs are expected to be competitive with retail electricity prices by 2020 and competitive with wholesale electricity prices by 2030 [4]. This decrease in price will encourage more people and utilities to install PV and will contribute to further PV development.

The best cell efficiency reported is 37.70% for three-junction PV cells in a laboratory environment [5]. Although the efficiency improvement rate of PV cells at the research state is relatively slow, the efficiency of commercial modules is expected to improve: the efficiency of typical commercial flat-plate modules is forecast to increase from 16% in 2010 to 40% in 2050 [4]. The efficiency improvement of PV cells or modules will help reduce the area required for each PV installation and consequentially cut down on the total plant costs.

The growth of PV-based electricity generation has advantages for the environment compared to conventional fossil fuel-based technologies. However, the inherent variability of PV

systems is a challenge at high solar power penetration levels. One approach to dealing with variability is to utilize accurate PV output forecasting. Using such forecasting, off-grid PV users can optimize the capacity of the required energy storage system [6]. Similarly, distributed grid-connected PV users can optimize their energy usage schedules and centralized grid-connected PV plants can improve their strategy when bidding in the electricity market by employing those forecasts. System operators can make a better arrangement of the reserves. The output of a single PV system is more vulnerable to local weather changes. However, the impact of local weather changes on aggregated output for the PV power of a region will be smaller. Thus, this thesis is focused on investigating the application of forecasting technology at array and system (aggregate) levels.

1.2 Literature Review

In this section, a review of irradiance prediction, PV power output prediction and the economic value of forecast technology is provided.

1.2.1 Literature Review of Solar Irradiance Forecasting

Solar irradiance is the dominant variable for PV power output. Forecast irradiance can guide PV output prediction. In the following, three types of irradiance forecasting methods are reviewed, including, numerical weather prediction (NWP) methods, image-based methods and statistical methods.

In [7], the forecasting accuracy of one-day-ahead global horizontal irradiance from the European Center for Medium-Range Weather Forecasts (ECMWF) model was assessed. ECMWF publishes the global horizontal irradiance forecast at 0:00 Coordinated Universal Time (UTC) and 12:00 UTC and provides a forecast with a spatial resolution of 25 km and a time step of three hours up to ten-day-ahead. The accuracy was validated by more than 200 ground measured global horizontal irradiance centres in Germany where the normal-

ized Root-Mean-Square Error (nRMSE) was around 40%. This model tends to over-predict irradiance which is especially obvious for cloudy conditions at noon.

In [8], the forecasting accuracy of a Weather and Research Forecasting (WRF) model was examined. The WRF model is derived from the Global Forecast System (GFS) model. There are several versions of WRF models, which involve different physical parameterizations. Validated by three ground-measuring centers in Spain, the nRMSE ranged from 29% to 37%.

In [9], the accuracy of 48-hour-ahead forecasts of global horizontal irradiance from Environment Canada's Global Environmental Multi-scale Model (GEM) was assessed. This model runs four times a day (0:00 UTC, 6:00 UTC, 12:00 UTC and 18:00 UTC) and generates forecasts up to 48 hours in advance with a spatial resolution of 15 km and a time step of 7.5 minutes. The forecasting result was validated through ten ground stations over North America and the nRMSE ranged from 16.7% to 43.6%. The lowest error occur at Desert Rock, NV where most of the days are under sunny weather; and the highest error occur at Penn State, PA which has a complex weather distribution.

In [10], the forecasting accuracy of the US National Digital Forecast Database (NDFD) was examined. This model converts the three or six-hour cloud index forecast to hourly global horizontal irradiance forecasting through a local correlation function. The nRMSE was 38% for one-day-ahead forecasting.

In [11], the forecasting accuracy of global horizontal irradiance from three NWP models, namely, the North American Model (NAM), the GFS, and the ECMWF was evaluated for the USA. The NAM model runs four times a day (0:00 UTC, 6:00 UTC, 12:00 UTC and 18:00 UTC) and generates up to 36-hour forecasts with a spatial resolution of 12 km and a time step of one hour. Additionally, 84-hour-ahead forecasting is also available from the NAM with a time step of three hours. The GFS model also runs four times a day (0:00 UTC, 6:00 UTC, 12:00 UTC and 18:00 UTC) and generates up to 180-hour forecasts with a

spatial resolution 50 km and a time step of three hours. Through evaluation, the ECMWF forecasting was more accurate than the GFS and the NAM. The mean bias error (MBE) is less than 50 W/m^2 under clear conditions and more than 200 W/m^2 under cloudy conditions.

In [12], the forecasting accuracy of global horizontal irradiance from the NDFD , the ECMWF and the WRF models was evaluated for the USA. Measuring the nRMSE of one-day-ahead forecasting, the WRF ranged from 18% to 50%, the ECMWF ranged from 18% to 40% and the NDFD ranged from 18% to 41%. Based on the results for these three sites, the prediction error of the WRF model was higher than the others.

The NWP is the most accurate irradiance forecasting method when the horizons are longer than five hours [13]. The nRMSE of different NWP models generally varied from 15% to 45%, and the variation is mainly from location difference, not model difference. Because of highly unstable cloud movement and low spatial resolution of the NWP models, locations with a lot of cloudy weather will experience high forecasting error regardless of the type of NWP model.

In [14], an image-based short-term irradiance forecasting was proposed. The prediction was done via statistical analysis of images taken by satellite with 2.5 km $\times$ 2.5 km spatial resolution and 30-minute temporal resolution. Through cloud movement analysis, cloud and irradiance was forecast, with a nRMSE of 18% to 19%.

In [15], another satellite image based irradiance forecast was proposed. The pixel-specific cloud motion was predicted from the analysis of two consecutive satellite images. The accuracy was validated by seven ground stations across the USA. The Root-Mean-Square Error (RMSE) ranged from 125 W/m^2 to 188 W/m^2 for day-ahead irradiance forecasting. The mean observed irradiance for these seven sites ranged from 323 W/m^2 to 498 W/m^2.

In [16], irradiance nowcasting was proposed to achieve a high temporal and spatial resolution through the analysis of sky cover images. The sky cover images were taken by a sky imager every 30 seconds and were used to predict cloud movement. This method can

correctly predict 70% of cloud conditions within 2 km for the current time.

In summary, image-based irradiance forecasting is done by analyzing cloud movement from two consecutive images. The satellite image-based methods are used within a six-hour horizon and ground-taken image-based methods are used within one-hour horizon. Within six hours, image-based irradiance forecasting has good accuracy. However, due to the high cloud cover variability, extending the forecasting horizon will significantly decrease the forecast accuracy [16].

In [17], four groups of empirical models were reviewed, including sunshine-based models, cloud-based models, temperature-based models and other meteorological parameter-based models. Empirical models use astronomical, geographical, geometrical, physical and meteorological factors available from weather stations to calculate the forecast irradiance. The forecasting accuracy for daily global solar radiation was evaluated for Yazd, Iran, using data from 2004 to 2008. The RMSE of the sunshine-based model was 0.5385 MJ/m^2, the RMSE of the cloud-based model was 1.152 MJ/m^2, RMSE of the temperature-based model was 0.7103 MJ/m^2 and the RMSE of the meteorological parameters-based model was 0.8542 MJ/m^2.

In [18], different neural network-based irradiance forecasting models were reviewed for hourly, daily and month irradiance forecasting. These models included a feed-forward neural network, a radial basis function neural network (RBFNN), a recurrent neural network, a neuro-fuzzy neural network, a wavelet neural network. In terms of forecasting accuracy of monthly solar radiation, the Mean Absolute Percentage Error (MAPE) varied from 0.3% to 16.4%.

In [19], univariate and multivariate models for hourly irradiance forecasting were evaluated. The RMSE forecasting error of univariate models, including ARIMA (39.44 W/m^2), a feed-forward neural network (30.14 W/m^2), a neuro-fuzzy neural network (31.93 W/m^2), a recurrent neural network (32.40 W/m^2), and an RBFNN (31.09 W/m^2) was slightly larger than four multivariate models, which included a feed-forward neural network (31.18 W/m^2),

a neuro-fuzzy neural network (30.97 W/m^2), a recurrent neural network (33.94 W/m^2), and an RBFNN (32.61 W/m^2). The largest error occurred for the univariate ARIMA model; other types of neural network models had a similar forecasting accuracy regardless of input variables.

In [20], a hybrid model that combines ARIMA with a feed-forward neural network is proposed for hourly irradiance forecasting. ARIMA was used to predict irradiance in Summer and Spring and a feed-forward neural network model was used to predict irradiance in Autumn and Winter. The forecasting result, in terms of nRMSE, ranged from 15.1% to 17.7% for this hybrid model, while, the nRMSE ranged from 15.4% to 17.7% when using the ARIMA model alone. When using a feed-forward neural network model alone, the forecasting error ranged from 14.9% to 19.6%. The hybrid model had better accuracy because ARIMA is better at predicting irradiance on sunny days and feed-forward neural network models are better at predicting irradiance on cloudy days.

In [21], a least squares support vector machine (LS-SVM) forecasting model was used to generate one to three hours ahead irradiance forecasts for three locations in the USA. The one-hour-ahead forecasting Mean Absolute Error (MAE) was 34.23% for Seattle, 33.77% for Denver and 62.86% for Miami. Thus, even using the same model, the forecasting accuracy for different locations is significantly different.

In summary, aside from NWP and image-based methods, artificial intelligence methods, empirical methods and time series models were also frequently used in irradiance forecasting. These models are especially useful for locations with no irradiance measurement equipment and limited forecast meteorological information from local meteorological centres. In general, non-linear models are good at forecasting irradiance on cloudy days and linear models are good at forecasting irradiance on sunny days [22], but there is no universal "best" irradiance forecasting model. A forecasting model should be designed to match the local weather in order to achieve satisfactory forecasting accuracy.

1.2.2 Literature Review of Solar Power Forecasting

In [23], the forecasting accuracy of autoregressive and autoregressive with exogenous input (ARX) models were examined. These two models were used to generate up to 36-hour-ahead power output forecasts for 21 PV systems in Denmark. ARX performed better than AR and the nRMSE of ARX was about 8% for a six-hour-ahead forecast. When the horizon is longer than two hours, the importance of exogenous input (forecast irradiance from NWP models) is significant.

In [24], a physical model was developed to predict daily power production for a 250 kW PV system in China (30.1°N and 131.0°E). By modelling the solar radiation and PV cells performance, the daily forecasting error ranged from 5.24% to 13.14%.

In [25], the forecasting accuracy of a physical model and a feed-forward neural network model was compared for a 1 MW PV system. The physical model included a solar radiation model and a PV cell model. The feed-forward neural network model uses temperature, cloud cover, irradiance, and position of the sun as inputs. For a one-year period, the nRMSE was 12.45% for the physical model, and 10.5% for the feed-forward neural network model.

In [26], a particle swarm optimization algorithm was used to train a feed-forward neural network model for PV power output forecasting. This model has two hidden layers and uses day, time, cloud cover index, air temperature, wind speed, air humidity, UV index, precipitation and air pressure as inputs. The result shows that this particle swarm optimization algorithm is better than the classical training methods for the feed-forward neural network.

In [27], a recurrent neural network using element structure was used to predict PV power output for a 4080 W PV system in Denmark for up to 24 hours. The inputs were clear sky irradiance and forecast weather type index for the forecast days. For a half-year testing period, the MAPE for the recurrent neural network model was 16.47% and the MAPE for the feed-forward neural network was 30.72%.

In [28], a similar recurrent neural network using element structure was used to predict 24-

hour-ahead PV power output for a 1 kW system in Thanyaburi. Inputs were the calculated clear sky irradiance and forecast weather type indices for the forecast days. Validated by four testing days, the MAPE was 16.83%.

In [29], a neuro-fuzzy forecasting model was proposed for daily energy production prediction. Input variables including day, irradiance, air temperature, wind speed, air humidity and air pressure were preprocessed by a fuzzy filter and fed into the feed-forward neural network model. The forecasting accuracy was evaluated by three PV plants and the average error was around 5%.

In [30], a hybrid model which combines an RBFNN with a weather classification method was proposed to predict day-ahead PV power output for an 18 kW system in Wuhan, China. According to the irradiance, total cloud and low cloud cover, the dataset was classified into three groups and used to train three sub-models, including a sun prediction model, a cloud prediction model and a rain prediction model. Inputs for each model were the time, past daily energy production, forecast daily irradiance, wind speed, temperature and humidity. Each model was evaluated over four testing days. The MAPE of the sun prediction model ranged from 8.29% to 10.8%, the MAPE of the cloud prediction model ranged from 6.36% to 15.08% and the MAPE of the rain prediction model ranged from 24.16% to 54.44%.

In [31], another hybrid model which combines a support vector machine and a weather classification method was proposed for a 20 kW PV station in China. The training set was classified into four groups according to the weather type, and was used to train four sub-SVM models. According to the weather types of the forecast day, the sub SVM model was selected to do the prediction. Inputs were the power output and the high, low and average forecast temperature for the forecast day. The one-day-ahead forecasting with 15-minute intervals had an average nRMSE of 10.5%. For each sub-models, the nRMSE was 9.12% for the cloudy model, 12.6% for the foggy model, 12.4% for the rainy model and 7.85% for the sunny model.

In [32], a feed-forward neural network model was used to predict power output for a 15 kW system in Ashland, Oregon. This one-day-ahead forecasting was done with 30-minute intervals. Input variables were power and temperature. In this method, the inputs are chosen by a similar day method and the most similar day is chosen by comparing the euclidean distance of high, low and average temperature between the forecast day and the historical day. If the weather type of the historical days is different from the day type of the forecast day, these days will not be selected regardless of the euclidean distance. The forecasting error (MAPE) was 18.89% for rainy days and 10.06% for sunny days.

In [33], a hybrid model combining a weighted support vector machine and similar day method was used to forecast power output for a 500 kW plant at Xuzhou, China. Five days selected through the similar day method were used to train the weighted support vector machine model and days with similarity with the forecast day were given a higher weight. The similarity was calculated through multiplying season similarity, weather similarity and temperature similarity. Inputs were irradiance and temperature. For a ten day testing period, the nRMSE for a one-hour-ahead forecast was 4.36%.

In [34], a combination of wavelet transform and an RBFNN was introduced to generate a one-hour-ahead PV output forecast for a 15 kW PV plant in Ashland, USA. Inputs of the network were the decomposed past PV power output, irradiance and temperature. The MAPE varied from 4.24% to 13.81% depending on the season.

In [35], a feed-forward neural network was used to make 10-minute-ahead and 20-minute-ahead forecasts for a 40 kW PV system in Hong Kong. This network uses solar elevation, azimuth angle, temperature and irradiance as inputs. The forecasting accuracy is measured through the correlation coefficient between forecast power and recorded power. This coefficient was 0.94% for 10-minute-ahead forecasting and 0.88% for 20-minute-ahead forecasting.

In [36], the accuracy of five univariate forecasting models were compared for a 1 MW PV plant in Merced, California, including a persistence model, an ARIMA model, a k-

nearest neighbours model, a feed-forward neural network and a feed-forward neural network optimized by Genetic Algorithms(GA/ANN). The nRMSE of one-hour-ahead forecasting was 19.27% for the persistence model, 18.95% for ARIMA, 20.90% for k-nearest neighbours, 15.82% for the feed-forward neural network and 13.07% for GA/ANN. Through comparison, GA/ANN had better forecasting accuracy in this location. Moreover, when extending the horizon from one hour to two hours, the accuracy significantly dropped and the forecasting error of different models became similar.

In [37], two-day-ahead forecasting accuracy of a persistence model, an ARIMA model, a k-nearest neighbours model, a feed-forward neural network, a recurrent neural network, a time delay neural network, an adaptive neuro-fuzzy inference systems and an RBFNN were compared for a 36 kW PV plant. Examined by 20% of the whole year data, nRMSE was 21.18% for the persistence model, 17.36% for ARIMA, 17.07% for k-nearest neighbours, 13.89% for the feed-forward neural network, 13.79% for the recurrent neural network, 15.12% for the RBFNN, 14.76% for the time delay neural network, and 14.09% for the adaptive neuro-fuzzy inference systems. Overall, the feed-forward neural network performed better than the others.

In the literature, various forecasting models have been applied to array level solar power forecasting, but the reported accuracy varies significantly from one location to another regardless of the forecasting models. This difference could be attributed to the fact that solar power time series at array level are highly influenced by local weather phenomena. Thus, it is necessary to investigate alternative modeling mechanisms capable of detecting historical patterns despite the lack of continuity in the data.

In addition, current studies all focus on array level forecasting, and no study has investigated the applicability of established forecasting methods at the system level. The system level forecasting could be useful for system operators to make a better reserves arrangement. Moreover, with the expansion of PV systems within the electricity market, this forecast may

help all participants to optimize their bidding strategies.

1.2.3 Literature Review for Economic Value of Forecasting

In [38], the economic value of direct normal irradiance forecasting was evaluated for a 50 MW solar thermal plant. When participating in the Spanish day-ahead market, an accurate direct normal irradiance forecast can help to reduce the penalty when the actual production differ from the scheduled production. In this study, the forecasting accuracy of a two-day persistence model and a site-specific Model Output Statistics (MOS) model was evaluated. From 2007 to 2009, the nRMSE of the persistence model ranged from 73% to 81% and the nRMSE of the MOS model ranged from 56% to 77%. If the solar thermal power plant schedules the output according to the persistence model, the average annual penalty would be 460,662€. If the scheduling is based on the MOS model, the average annual penalty will be 241,600€. In summary, an increase of 1% in forecasting accuracy will save 0.7% in penalties.

In [39], the value of irradiance forecasting was evaluated for another 11 MW solar thermal power plant in Spain. This study evaluated the revenue of operations guided by two different forecasting models, namely, aerosol-based forecasting and ECMWF forecasting. The nRMSE of direct horizontal irradiance forecasting is 25.1% for the aerosol-based model and 18.5% for the ECMWF. Under three cloudy days, revenue for the aerosol-based model guided strategy was 46,900€ while revenue for the ECMWF guided strategy was 53,500€ which was very close to ideal revenue (54,500€). In summary, improving irradiance forecasting accuracy could help solar power plants gain more revenue and reduce penalties.

1.3 Research Objectives

The first objective of this thesis is to investigate PV power forecasting at a single array level. Analysis of solar power data shows that solar power time series are highly non-stationary

and continuity of production patterns are highly disturbed by local weather changes. Thus, it is a challenge for forecasting models to identify production patterns and associate them to input variables. In this thesis, an effort is made to design a forecasting tool which can search historical data and identify power output patterns that could be used to improve forecasting accuracy.

The second objective of this thesis is to investigate the applicability of established forecasting methods in predicting short-term variations of the aggregated output of solar systems distributed across a system. Because of the geographical diversity of solar arrays across a wide area, aggregated solar power output is less prone to the effects of local weather change. In this thesis, solar power forecasts at the aggregate system level are generated and their accuracy is compared with that of array level power forecasts.

1.4 Thesis Outline

Following is the outline for this thesis:

1. Chapter 2 provides a detailed description of sunlight, PV cells, PV system and forecasting methods.

2. Chapter 3 describes the short-term PV power forecasting tool at array level.

3. Chapter 4 describes the short-term PV power forecasting tool at system level.

4. Chapter 5 summarizes the main contributions of this thesis, and proposes some possible future research directions.

Chapter 2

Background Review

2.1 Introduction

This chapter presents necessary background information. In Section 2.2, the features of sunlight are introduced. In Section 2.3, the structure and characteristics of PV cells are described. In Section 2.4, the structure and typical examples of different PV systems are introduced. In Section 2.5, the forecasting models used in this thesis, including ARIMA, LS-SVM and RBFNN, are described.

2.2 Sunlight Patterns and Clear Sky Radiation Model

Sunlight has a direct influence on the assessment and production of PV systems. Thus, it is necessary to understand basic properties of sunlight. In the following, components of sunlight and the pattern of sunlight changes are described. This pattern is used to build a clear sky radiation model.

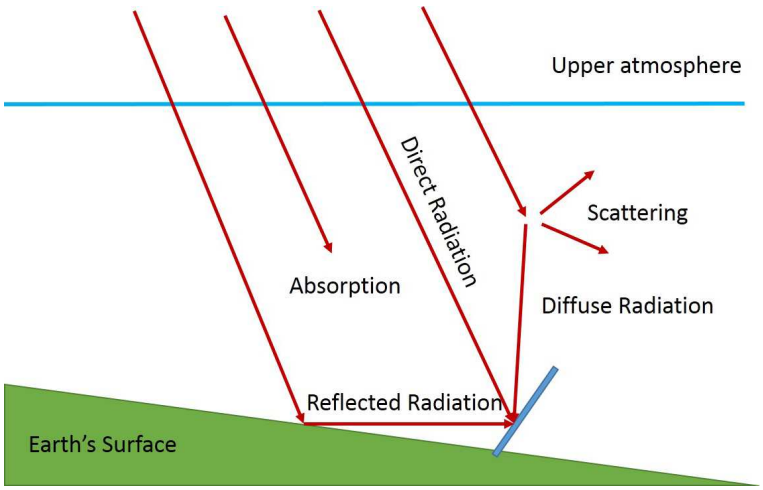

Figure 2.1: The path of solar radiation [41]

2.2.1 Sunlight Patterns

Sunlight is emitted from the sun in all directions. When sunlight reaches the upper atmosphere, its strength has decreased to 1367 W/m^2 (measured as irradiance), which is defined as solar constant [40]. When sunlight enters the atmosphere, some of it is scattered and some is absorbed by the air molecules, dust and clouds as shown in Figure 2.1 [41]. Sunlight that reaches the surface of PV modules without absorption and scattering is defined as direct solar radiation. Sunlight reflected by the ground is defined as reflected solar radiation. Sunlight scattered by the air molecules, dust and etc. is defined as diffuse solar radiation. These three types of radiation make up global solar radiation which is frequently used in solar power forecasting. The strength of global solar radiation depends on the length of travel through the atmosphere. Obviously, longer distance will lead to weaker global solar radiation. Since the sunlight needs to travel a longer path to the polar regions than the tropics, polar regions will receive less solar radiation. Air mass is the term for measuring this length. It equals 0 when radiation reaches the upper atmosphere and equals 1 when the sun is straight overhead. Air mass equals to 1.5 is the standard condition to rate the capacity of PV system. The maximum irradiance that could be received on the surface of the earth can be estimated by air mass using [42]:

$$I_{Max} = 1367 \times 0.7^{AM^{0.678}} \tag{2.1}$$

where, I_{Max} is the strength of irradiance and AM is the value of air mass.

Typically, summer days have more sunlight hours than winter days during a day, and sunlight becomes stronger during the morning and becomes weaker during the afternoon. These patterns result from the relative position changes of the Sun and the Earth.

Seasonal sunlight changes result from the apparent yearly movement of the Sun which is plotted in Figure 2.2. The angle between the apparent path of sun and the celestial equator is 23.45°. Solar declination δ represents the angle between the deviation of the line linking the centre of the Earth and the Sun with the equatorial plane. Depending on the date, solar

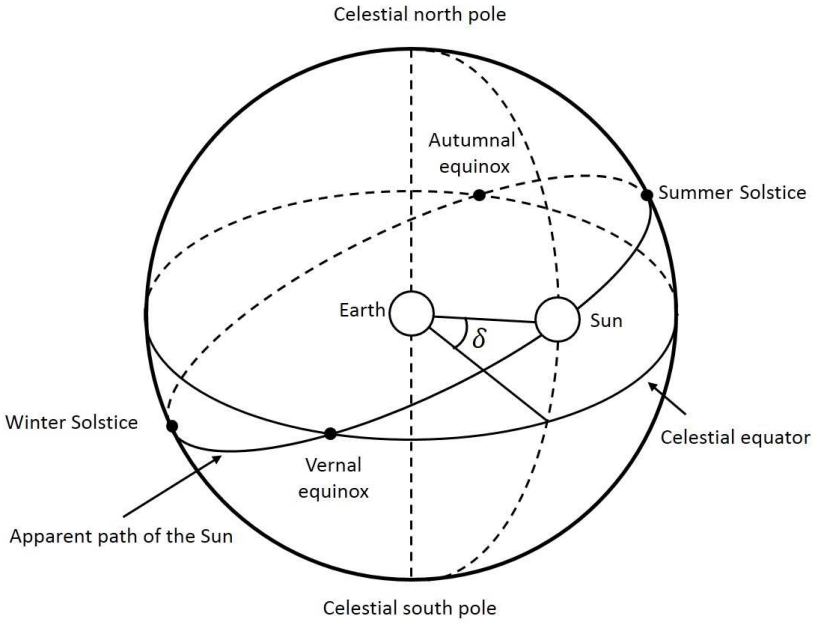

Figure 2.2: Apparent yearly movement of the Sun [41]

declination (in radians) can be estimated as:

$$\delta = \pi \frac{23.45}{180} \sin 2\pi \frac{284 + D}{365} \tag{2.2}$$

where, D is the day of a year. Within one day, this angle is assumed to be constant. At March 20 or 21(vernal equinox) and September 22 or 23 (autumnal equinox), this solar declination is zero. At June 21 or 22 (summer solstice) and December 21 or 23 (winter solstice), this solar declination is 23.45°. This differences lead to a long day length during summer and a short day length during winter.

Daily sunlight changes result from the apparent daily movement of the Sun which is plotted in Figure 2.3. Figure 2.3 (a) shows the daily rotation of the earth using the expression of a rotation of the celestial sphere. θ_ϕ represents the geographical latitude of the observation location on the Earth. ω is the hour angle that represents the instantaneous point of the sun which can be calculated by:

$$\omega = 15(T_{LC} - TZ - 12) + \theta_\varphi + T_E/4 \tag{2.3}$$

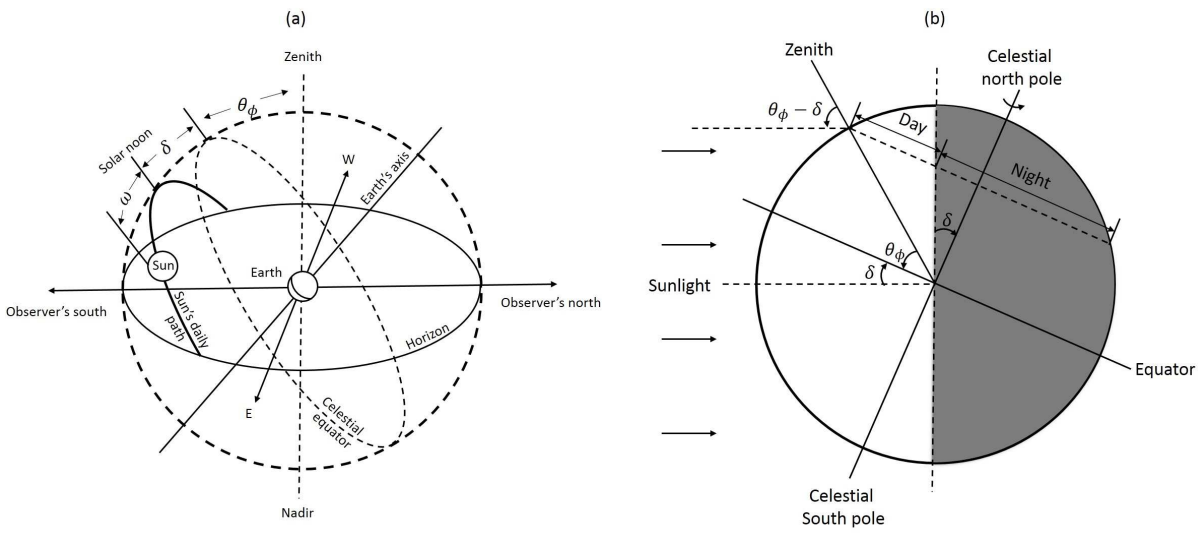

Figure 2.3: Apparent daily movement of the Sun [41]: (a) Apparent position between the Sun and the Earth in the celestial sphere, (b) Day and night results from this apparent daily movement

where, T_{LC} is the local mean time shown on the clock, TZ represents the time zone and θ_φ refers to the longitude of the PV site. T_E is the equation of time that is calculated through:

$$T_E = 229.1831(0.000075+0.001868\cos\theta_\varsigma-0.032077\sin\theta_\varsigma-0.014615\cos 2\theta_\varsigma-0.040849\sin 2\theta_\varsigma) \tag{2.4}$$

where, $\theta_\varsigma = \dfrac{360}{364}(D-1)$. This apparent daily movement of the Sun in the daily path will lead the changes of day and night which is shown in Figure 2.3 (b).

2.2.2 Clear Sky Radiation Model

From the above description of sunlight, we know that global solar radiation that could reach the inclined surface on the Earth includes direct, diffuse and reflected solar radiation. If the installation information of a PV system is known, the global horizontal irradiance that could reach the surface of a PV array under clear sky conditions could be calculated through Hottel's solar radiation model [27].

Figure 2.4 illustrates the solar angles and surface orientation angles related to the PV

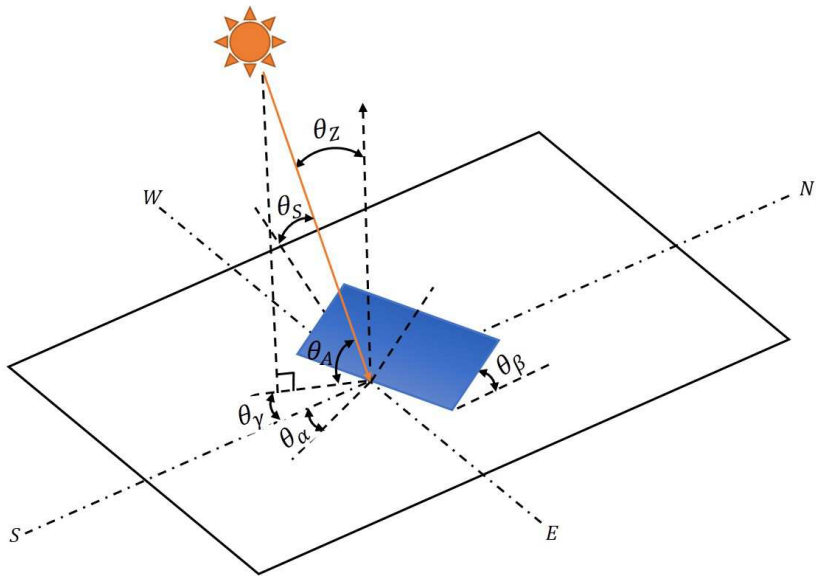

Figure 2.4: Installation angle of PV array [27]

installation. Solar zenith angle θ_Z is the angle measured between a point directly overhead and the centre of the sun. The solar altitude θ_A equal to $90 - \theta_Z$. Solar azimuth θ_γ measures the direction of the sun and it increases from east to west and reaches zero at solar noon. θ_α represents the surface azimuth angle and θ_β represents the surface inclination angle. Solar inclination angle θ_S represents the angle between the sun and the normal to the surface. This angle will be used later for irradiance calculation. The solar zenith angle θ_Z, solar azimuth angle θ_γ and solar inclination angle could be calculated through:

$$\cos\theta_Z = \sin\delta\sin\theta_\phi + \cos\delta\cos\theta_\phi\cos\omega \tag{2.5}$$

$$\cos\theta_\gamma = (\sin\theta_\alpha\sin\theta_\phi - \sin\delta)/\cos\theta_\alpha\cos\delta \tag{2.6}$$

$$\begin{aligned}\cos\theta_S = {} & \sin\delta\sin\theta_\phi\cos\theta_\beta - \sin\delta\cos\theta_\phi\sin\theta_\beta\cos\theta_\alpha + \cos\delta\cos\theta_\phi\cos\theta_\beta\cos\omega + \\ & \cos\delta\sin\theta_\phi\sin\theta_\beta\cos\theta_\alpha\cos\omega + \cos\delta\sin\theta_\alpha\sin\omega\sin\theta_\beta\end{aligned} \tag{2.7}$$

The global solar radiation that hits the surface of the array can be calculated with the following equations [43]:

$$I_{global} = I_{direct} + I_{diffused} + I_{reflected} \tag{2.8}$$

$$I_{direct} = I_E \cdot \eta_{direct} \cdot \cos\theta_S \tag{2.9}$$

$$I_{diffuse} = I_E \cdot \eta_{diffuse} \cdot \cos\theta_Z \cdot (\frac{1+\cos\theta_\beta}{2}) \tag{2.10}$$

$$I_{reflected} = I_E \cdot \rho \cdot \eta_{reflected} \cdot \cos\theta_Z \cdot (\frac{1+\cos\theta_\beta}{2}) \tag{2.11}$$

$$I_E = I_S[1 + 0.033\cos(360D/365)] \tag{2.12}$$

where, I_{global} is the global solar radiation, I_{direct} is the direct solar radiation, $I_{diffuse}$ is the diffuse solar radiation, $I_{reflected}$ is the reflected solar radiation and I_E is the extraterrestrial solar radiation. I_S is the solar constant and ρ is the average reflectance of the ground, which is 0% for total absorption and 100% for total reflection. $\eta_{direct}, \eta_{diffuse}, \eta_{reflected}$ are respectively the atmospheric transmittance for direct, diffuse and reflected solar radiation which could be calculated through [43]:

$$\eta_{direct} = \upsilon_0 + \upsilon_1 \exp(-\upsilon_2/\cos\theta_Z) \tag{2.13}$$

$$\eta_{diffuse} = 0.271 - 0.294\eta_{direct} \tag{2.14}$$

$$\eta_{reflected} = 0.271 + 0.706\eta_{direct} \tag{2.15}$$

where, υ_0, υ_1, υ_2, is the constants that can be calculated through:

$$\upsilon_0 = \tau_0[0.4237 - 0.00821(6 - H_A)^2] \tag{2.16}$$

$$\upsilon_1 = \tau_1[0.5055 - 0.00595(6.5 - H_A)^2] \tag{2.17}$$

$$\upsilon_2 = \tau_2[0.2711 - 0.01858(2.5 - H_A)^2] \tag{2.18}$$

where, H_A refers to the altitude of the location in km. τ_0, τ_1, τ_2 are the climate factors listed in Table 2.1 for four typical climate types.

In summary, for a given PV site for which installation information is known, the Hottel clear sky radiation model can estimate the hourly clear sky irradiance which could hit the

Table 2.1: Climate factors for four typical climate types

Climate Type	τ_0	τ_1	τ_2
Tropical	0.95	0.98	1.02
Mid-altitude Summer	0.97	0.99	1.02
Subarctic Summer	0.99	0.99	1.01
Mid-altitude Winter	1.03	1.01	1

surface of the PV array. This clear sky irradiance value could be used for PV resource assessment or PV output estimation [27].

2.3 Structure and Electrical Characteristics of PV Cells

The PV cell is the basic component of a PV array which converts solar radiation to electricity. In this section, the structure, mathematical model and features of a typical PV cell are introduced. This introduction aims to illustrate the relationship between solar radiation and the electricity generated by a PV array.

2.3.1 Structure of PV Cells

PV cells are wired in series to build a PV module and those PV modules are strung in series and parallel to make a PV array. Figure 2.5 (a) is a schematic of a PV array and the component encircled by a black rectangle is a PV cell. Due to the surface area limitation, the power generated by a single cell is limited. For example, one 100 cm^2 single crystalline silicon solar cell can, at most, generate 1.5 W power when exposed to full sunshine [44]. The power from one cell is thus not enough for real application. Thus, PV cells are usually wired in series to gain a higher output voltage and are wired in parallel to gain a higher output current. If higher power is needed, several PV models are strung together to make a PV array.

Figure 2.5 (b) is a schematic of a typical crystalline silicon PV cell [45]. The PV cell is composed of a front contact, an anti-reflection layer, a P-type silicon, a N-type silicon and

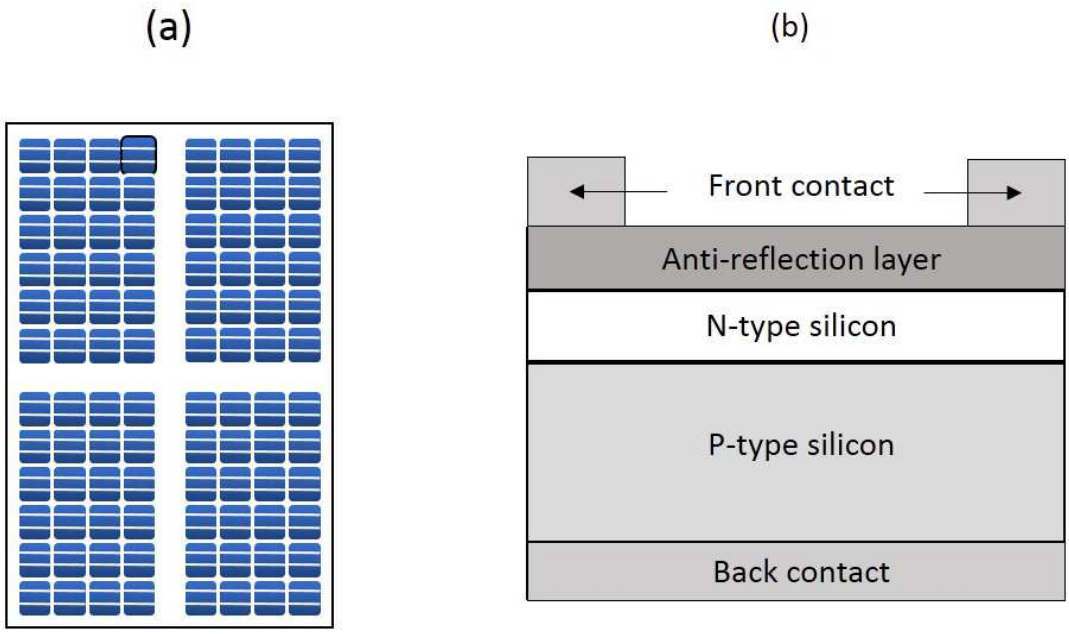

Figure 2.5: (a) PV cells, PV modules and PV arrays (b) Structure of a crystalline silicon solar cell [45]

a back contact. The front contact, which is made of a good conductor, is used to collect electrons. The back contact is made of metal and serves as a conductor and the covers for the back surface. The anti-reflection layer is made of a combination of glasses with different refractive index and thickness. This layer helps the PV cells receive more sunlight and reduce reflection. N-type silicon is a doped layer that contains one more valence electron than normal silicon; P-type silicon is a doped layer that contains one less valence electron than normal silicon. Only four electrons are needed to bond the silicon atoms, so N-type silicon tends to donate valence electrons and P-type silicon tends to adopt valence electrons. The connection of the N-type and P-type silicon forms a P-N junction which contains an electric field and resists the movement of electrons from N-type silicon to P-type silicon. The thick P-type silicon layer absorbs most of the sunlight and generates most of the power [46]. When sunlight with sufficient energy hits the silicon, the elections will be forced to move from front contact to the load and return through the back contact [44].

2.3.2 Electrical Characteristics of PV cells

The electrical characteristics of PV cells are illustrated through the equivalent circuit of PV cells and the corresponding I-V characteristics in this section.

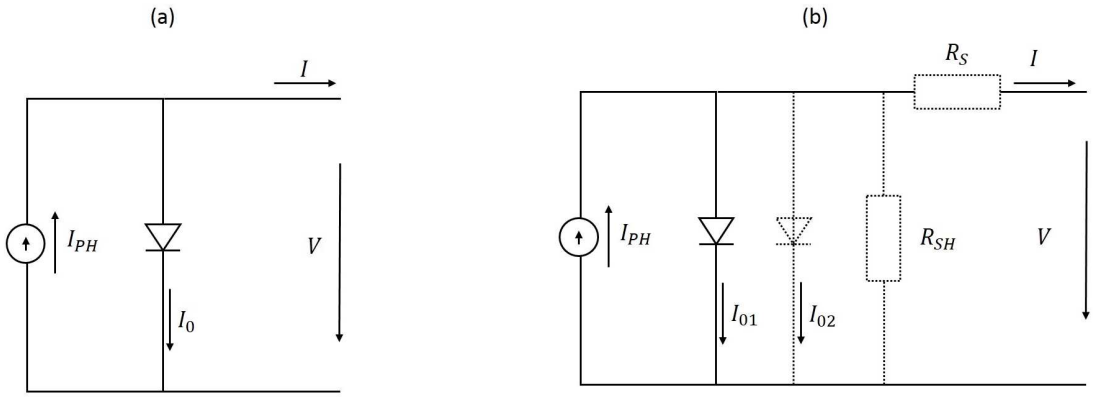

Figure 2.6: General circuit diagram for: (a) single diode PV cell model (b) two diodes PV cell model

The equivalent circuit of a single diode PV cell model can be modeled with a parallel combination of a current source and a rectifying diode as shown in Figure 2.6 (a). The I-V characteristic of the single diode PV cell model can be expressed as:

$$I = I_{PH} - I_0 \left[\exp \frac{qV}{k_B T_C} - 1 \right] \tag{2.19}$$

where, q is an electron charge ($q = 1.6 \times 10^{-19} C$), k_B is the Boltzmann's constant ($k = 1.38 \times 10^{-23} J/K$), T_C is the cell temperature and V is the terminal voltage of the cell. I_0 is the diode saturation current, and this current indicates that PV cells function as a semiconductor current rectifier or diode when there is no sunlight hitting the cells. I_{PH} is the photo-current, which is related to the strength and wavelength of the light. Usually, the applied voltage will not affect I_{PH}, except for cells like A-Si and some other thick film cells [47]. The I-V characteristic of this single diode PV cell model is plotted in Figure 2.7

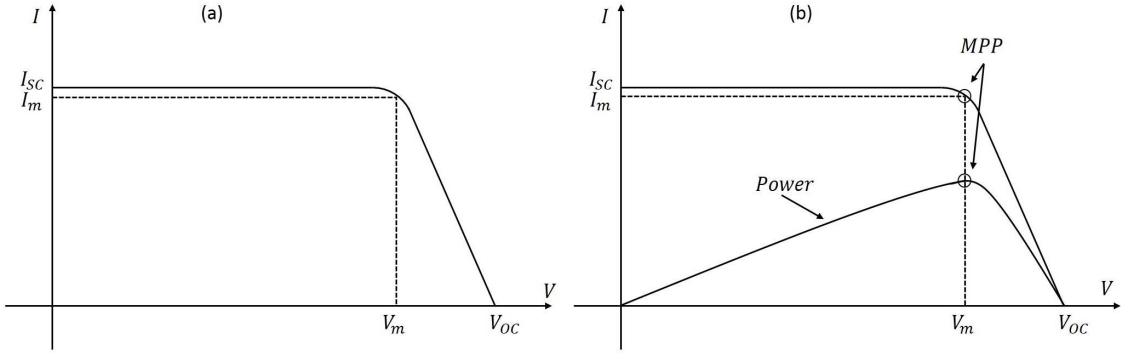

Figure 2.7: (a) I-V curve of a PV cell, (b) Power curve and maximum power point (MPP)

(a). The short-circuit current and open-circuit voltage can be calculated through:

$$I_{SC} = I_{PH} \tag{2.20}$$

$$V_{OC} = \frac{k_B T_C}{q} \ln(1 + \frac{I_{PH}}{I_0}) \tag{2.21}$$

The relationship of the short-circuit current (I_{SC}) and open-circuit voltage (V_{OC}) with radiation (G) is shown in Figure 2.8 [44]. I_{SC} is proportional to solar radiation and also related to cell temperature. I_{SC} increases at $0.05\%/K - 0.07\%/K$ for crystalline silicon solar cells, and I_{SC} increases at approximately $0.02\%/K$ for amorphous silicon solar cells. A lower temperature coefficient can improve the performance of a solar cell during hot weather. V_{OC} increases very rapidly with radiation until it reaches a saturation value. After this point, V_{OC} will grow very slowly and this slow growth usually cannot be observed due to internal and external resistances.

The power generated by the two diodes PV cell model is shown in Figure 2.7 (b). Maximum power is generated at voltage V_m and current I_m. The ratio between $I_m V_m$ and $I_{SC} V_{OC}$ is the fill factor. Typically, crystalline silicon solar cells have a fill factor of $0.7 \sim 0.8$ and amorphous silicon solar cells have a fill factor of $0.5 \sim 0.7$ [44].

The equivalent circuit of a two diodes PV cell model can be modelled by one current source, rectifying diodes, a series resistor R_{SH} and a parallel resistor R_S, as shown in Figure

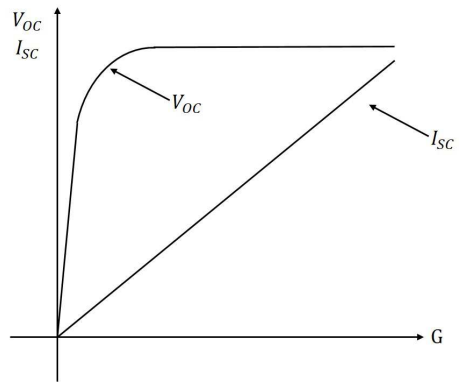

Figure 2.8: Short-circuit current and open-circuit voltage relationship with solar radiation

2.6 (b). The I-V characteristic of this model can be expressed as:

$$I = I_{PH} - I_{01}\left[\exp\frac{V + IR_S}{k_B T_C} - 1\right] - I_{02}\left[\exp\frac{V + IR_S}{2k_B T_C} - 1\right] - \frac{V + IR_S}{R_{SH}} \tag{2.22}$$

where, the series resistor R_{SH} represents the resistance of silicon wafer, contact and circuit and parallel resistor R_S represents the loss currents from the surface and the edges of a solar cell [44]. I_{01} and I_{02} represent the current through the diodes. The effect of the second diode, parallel resistor and series resistor are explained in Figure 2.9. Figure 2.9 (a) shows the I-V characteristics for three different ratios of I_{02}/I_{01}. Figure 2.9 (b) and (c) show how series and parallel resistors affect the I-V characteristics. Those influences are important since they occur in the region of the MPP.

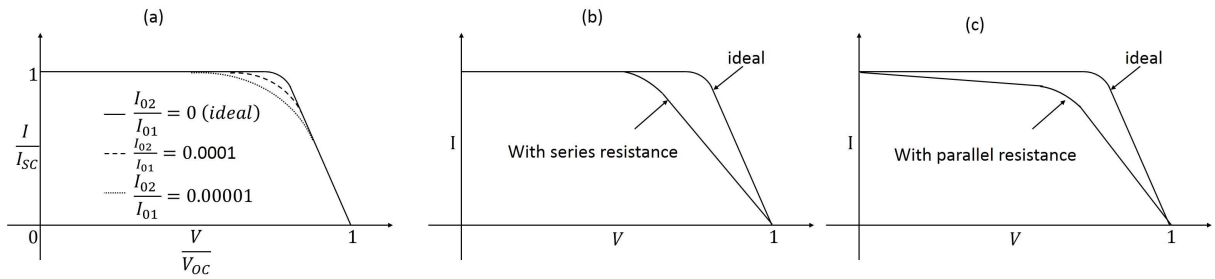

Figure 2.9: The effect of (a) two diodes, (b) series resistance and (c) parallel resistance on the I-V characteristics of the PV cell

For a single PV array, the maximum DC output can be calculated using [48]:

$$P_{Max} = \eta A \lambda [1 - 0.005(T_A + 25)] \tag{2.23}$$

where, P_{Max} is the maximum DC power output (kW), η is the conversion efficiency of PV cell (%), A is the area of the PV array (m^2), λ is the solar radiation (kW/m^2) and T_A is the ambient temperature (℃). The conversion efficiency is defined as the ratio of the maximum power output $P_m = I_m V_m$ from this cell to the solar power P_S falling on it and this efficiency is determined by the type of the PV cells. When combined with a clear sky radiation model, the maximum PV output can be estimated for any location. However, this equation is only suitable for a very small PV system without a shading problem and with all arrays at the same orientation.

2.4 PV Systems

In this section, we introduce different types of PV systems, comprising groups of PV arrays. Depending on whether a PV system is connected to the grid or not, it can be classified as off-grid or grid-connected systems.

2.4.1 Off-grid PV Systems

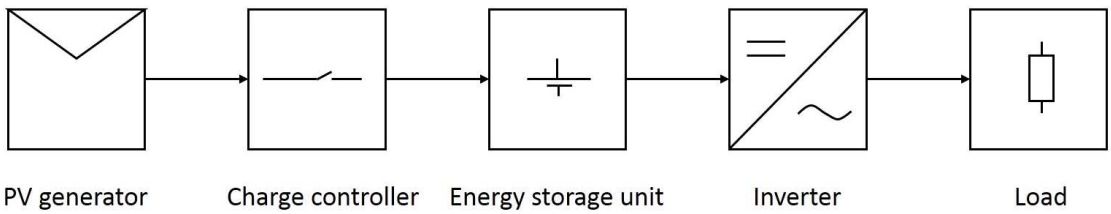

Figure 2.10: Block diagram of a residential PV system [44]

Currently, there are great variations of off-grid PV systems such as solar calculators, solar street lamps, a system that can supply power for a remote house or building and etc. [44]. Off-grid PV systems typically integrated an energy storage system to ensure power supply

when there is no radiation (e.g. at night) or very limited radiation (e.g during cloudy or rainy days) [49]. Off-grid PV systems can be used to supply power for a house or a building in remote areas. There are about two thousand million people around the world who do not have access to the power grid. Even in central Europe, there are people who do not have access to the public grid [44]. Thus, off-grid PV systems have obvious value for them. Figure 2.10 shows the typical structure of an off-grid residential PV system including a PV generator, a charge controller, a battery and an inverter. The charge controller can protect the battery against deep discharge and overcharging and can ensure the efficient operation of the battery. The battery, together with the charger controller, is the energy storage system, and this system is critical for a residential PV system when there is insufficient radiation. Inverters are used to convert the DC power to AC power in order to supply the AC applications.

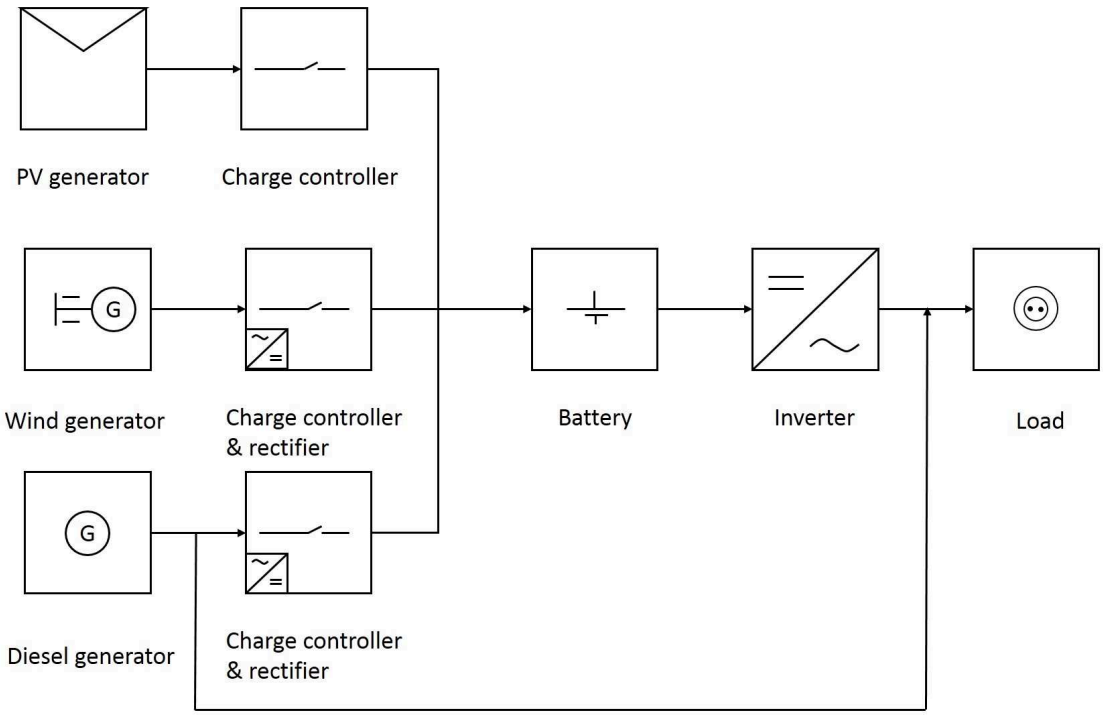

Figure 2.11: Block diagram of a hybrid PV system [44]

Because of the annual fluctuation of solar radiation, an exclusively PV power supply

system needs either a large solar generator or a large battery to maintain the power supply. Thus, hybrid systems are usually used, which are powered by different types of generators, such as wind and diesel generators. Figure 2.11 shows a hybrid PV system that integrates wind and a backup generator (e.g diesel generator). Under favourable weather conditions, all power is supplied by PV and wind generator, and the surplus power will be used to charge the battery. Under unfavourable weather conditions or at night, the power will be supplied by the battery or from the diesel generator directly. If the battery reaches deep discharge, it will be charged by the diesel generator through the charge controller and rectifier.

2.4.2 Grid-connected PV Systems

Grid-connected PV systems have a permanent connection with the electricity grid through inverters. The grid-connected PV systems can be subdivided as distributed grid-connected PV systems and central grid-connected PV power plants.

Distributed grid-connected PV systems are usually installed on the roof of a house or building. Figure 2.12 (a) shows the structure of a distributed grid-connected PV system. Compared to the off-grid PV system, there is no energy storage system. When the PV power is not enough, users can draw power from the electricity grid to supply the applications. When there is surplus power from the PV generators, the power will be fed back to the grid.

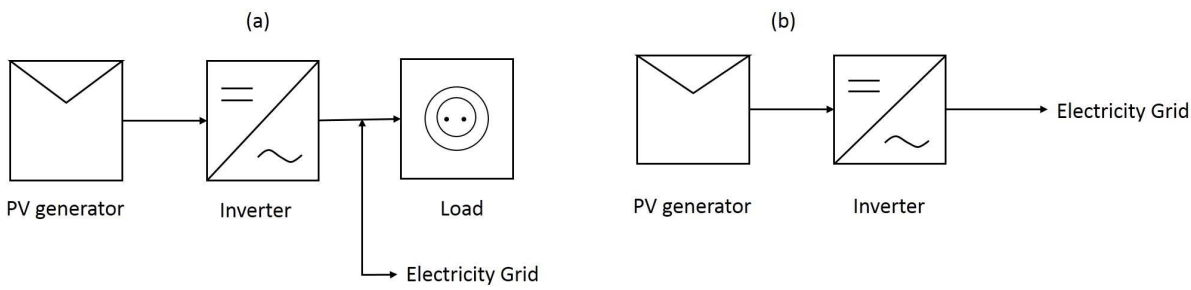

Figure 2.12: Block diagram of grid-connected PV systems [44]: (a) Distributed grid-connected PV systems, (b) Central grid-connected PV systems

Grid-connected PV plants usually have a capacity larger than 1 MW. These systems are

usually built on unused land and connected to the middle and high voltage grid. Figure 2.12 (b) shows the structure of central grid-connected PV systems. Unlike a distributed grid-connected system, they do not draw power from the grid. Depending on the size of the PV system and its configuration, the inverters are different. Table 2.2 lists the power range, availability of maximum power point tracking (MPPT) function and typical efficiency of modular, string, multi-string and central inverters [50]. Micro-inverters are easy to install on the back of the PV modules. However, it may be difficult to replace a faulty inverter. Currently, the latest micro-inverters can upload solar power data through websites [51]. String inverters are used in small PV systems. Since there is only one MPPT in a string inverter, some PV models may not work at their MPP point. Multi-string inverters apply to larger PV systems and include multiple MPPT. Central inverters have a similar structure of the string inverters, but they are used for PV systems that are larger than 10 kW and the unit cost of central inverters is usually low.

Table 2.2: Inverter types and characteristics [50]

Inverter type	Micro	String	Multi-string	Central
Power Range (kW)	0.1-0.3	0.7-11	2-17	10 -300
MPPT	Yes	Yes	Multiple	Multiple
Typical Efficiency (%)	95%	93% 97%	97%	97%

Currently, the largest PV power plant in Canada is the 97 MW Sarnia PV power plant. This plant uses 1,300,000 thin-film modules and covers 1,100 acres. The largest PV plant in the world is the 250 MW Agua-Caliente solar project in the USA, which covers 2400 acres and has an annual generation of 626.219 GWh. Greece plans to install a 10,000 MW Helios PV power plant by 2020.

2.5 Forecasting Models

There are many modeling or forecasting tools described in the literature. ARIMA is a good representative of univariate linear models. LS-SVM is a non-linear model with a convex optimization process. RBFNN is a non-linear model that does not lead to a convex optimization, so it may be get trapped in a local optimal point. These methods have been used and reported in the literature and have shown good performance [23, 30, 31]. Thus, they are chosen as representative forecasting models. However, other models could be used for the purpose of this thesis.

2.5.1 ARIMA Model

ARIMA was initially introduced by Box and Jenkins for time series forecasting [52]. Since the ARIMA model has a better ability to capture diurnal cycle characteristics than similar methods [53], this method is utilized in this study. The following is a description of the ARIMA model. ARMA(p,q) for a stationary stochastic process z_t can be written as [52]:

$$z_t = c + \sum_{i=1}^{p} \phi_i z_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta j \epsilon_{t-j} \tag{2.24}$$

where, c, ϕ_i and θj are the free parameters for this model. ϵ_t are independently and identically distributed normal random variables with mean zero and variance σ_ϵ^2. When introducing a backward shift operator ($Bz_t = z_{t-1}$), Equation (2.24) can be expressed as [52]:

$$\phi(B)z_t = c + \theta(B)\epsilon_t \tag{2.25}$$

where, $\phi_p(B) = 1 - \phi_1 B - ... - \phi_p B^p$ is the non-seasonal auto regressive operator, $\theta_q(B) = 1 - \theta_1 B - ... - \theta_q B^q$ is the non-seasonal moving average operator. The variable mean is removed by d^{th} order differenced process $v_t = (1 - B)^d z_t$. The ARMA(p,q) model for the difference process v is called the Auto Regressive Integrated Moving Average model ARIMA(p,d,q) for the process z_t. If a time series has seasonality, indexed by s, then a seasonal ARIMA$(p, d, q)(P, D, Q)_s$ model can be expressed as:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D z_t = c + \theta_q(B)\Theta_Q(B^s)\epsilon_t \tag{2.26}$$

where, $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are seasonal auto regressive and moving average operators; B^s is the seasonal backward shift operator which is defined as $B^s z_t = z_{t-s}$ and D is the seasonal difference order.

This seasonal ARIMA$(p, d, q)(P, D, Q)_s$ model will be utilized in this study using the ARIMA toolbox from Matlab. The model's building process includes three steps, namely: model identification, model estimation and model diagnostic checking.

2.5.2 LS-SVM Model

LS-SVM was proposed by Suykens and Vandewalle as one of the supervised learning methods that can be used for regression [54]. The original SVM was proposed by Vapnik and his colleagues as a classifier [55]. LS-SVM changes the inequality constraints into equality constraints and defines the loss function as an experienced loss function of the training set. Therefore, a quadratic programming problem becomes a linear programming problem.

Based on a given training set in the form of:

$$\{(x_i, y_i), i = 1, 2, \cdots, N\} \tag{2.27}$$

where $y_i \in R$ is the target object, $x_i \in R^n$ represents the n attributes of the target and N is the number of training samples. LS-SVM tries to use a non-linear mapping function $\phi(x)$ to map the training set from input space to a higher feature space using Kernel function $K(x, x_k)$ and build an optimal linear regression function in the new space. The non-linear regression function $f(x)$ is in the form of:

$$f(x) = \omega^T \psi(x) + b \tag{2.28}$$

where $x \in R^n$, $y \in R$, ω is the weight vector and b is the bias.

The optimization problem $\min J(\omega, b, e)$ can be expressed as [54]:

$$\min J(\omega, b, e) = \frac{1}{2}\omega^T\omega + \frac{\zeta}{2}\sum_{k=1}^{N} e_k^2 \tag{2.29}$$

such that:

$$y_k = \omega^T \cdot \psi(x_k + b + e_k) \qquad k = 1, \ldots, N \tag{2.30}$$

where ζ is the adjustment parameter and e_k^2 is the quadratic loss function defined as:

$$e_k^2 = (y_k - f(x_k))^2 \qquad k = 1, \ldots, N \tag{2.31}$$

The Lagrange function L for this problem is:

$$L = J - \sum_{k=1}^{N} \alpha_k[\omega^T \cdot \psi(x_k + b + e_k) - y_k] \tag{2.32}$$

where α_k is the Lagrange multiplier. According to Karush-Kuhn-Kucker (KKT) condition:

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = 0 \Rightarrow \omega = \sum\limits_{k=1}^{N} \alpha_k \cdot \psi(x_k) \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \sum\limits_{k=1}^{N} \alpha_k = 0 \\ \dfrac{\partial L}{\partial e_k} = 0 \Rightarrow \alpha_k = \zeta \cdot e_k \\ \dfrac{\partial L}{\partial \alpha_k} = 0 \Rightarrow \omega^T \cdot \psi(x_k + b + e_k) - y_k = 0 \end{cases} \tag{2.33}$$

The following equation can be obtained by eliminating ω and e_k:

$$\begin{bmatrix} 0 & \delta \\ \delta^T & K + \zeta^{-1} \cdot I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{2.34}$$

where $\delta = [1, \ldots, 1]$, I is an identity matrix, $Y = [y_1, \ldots, y_k]$, $\alpha = [\alpha_1, \ldots, \alpha_N]$, $K \in R^{N \times N}$ and $K_{ij} = \psi^T(x_i) \cdot \psi(x_j)$.

Therefore, the optimum parameters α and b can be calculated through Equation (2.34) and the predicted output $y(x)$ can be obtained :

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k)) + b \tag{2.35}$$

where $K(x, x_k) = \psi^T(x) \cdot \psi(x_k)$ is the kernel function and figure 2.13 [56] is the visual expression of Equation (2.35).

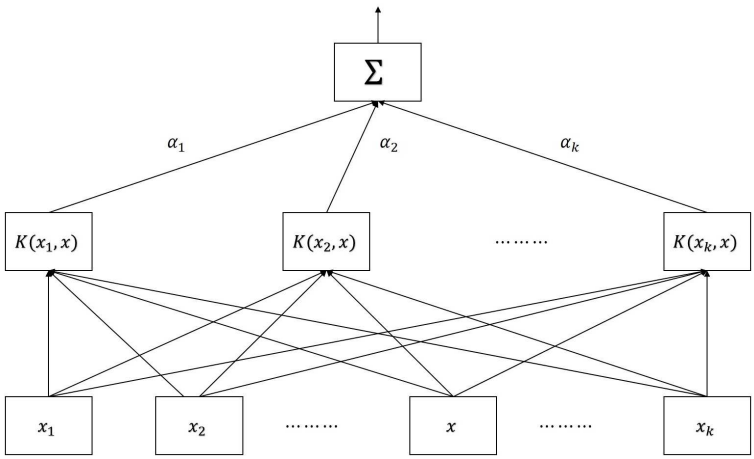

Figure 2.13: Structure of the LS-SVM [56]

2.5.3 RBFNN Model

RBFNN was first formulated by Broomhead and Lowe [57]. The basic form of RBFNN mapping is:

$$\hat{y} = \sum_{j=1}^{M} w_j H_j(x) + w_0 \tag{2.36}$$

where $\hat{y}$ is the forecast value and x is the input vector with a dimension of k, M is the number of hidden neurons, w_0 is the bias term and $H_j(x)$ is the Gaussian basis function for hidden neuron j:

$$H_j(x) = \exp\left(-\frac{||x - \mu_j||^2}{2\sigma_j^2}\right) \tag{2.37}$$

where, μ_j is the centre of the Gaussian basis function and σ_j is smoothness parameter of the Gaussian basis function.

Figure 2.14 is the schematic of RBFNN structure [58], which is similar to a three layer neural network that uses RBFs as activation functions, but it is actually a unique neural network. Unlike a feed forward neural network, whose weights are trained by the BP method, RBFNN is trained by two stage methods. The first stage training is done by unsupervised training: the input dataset x_n alone is used to determine the radial basis function parameters and the first layer weights. When the first stage training is done, the weights of the first

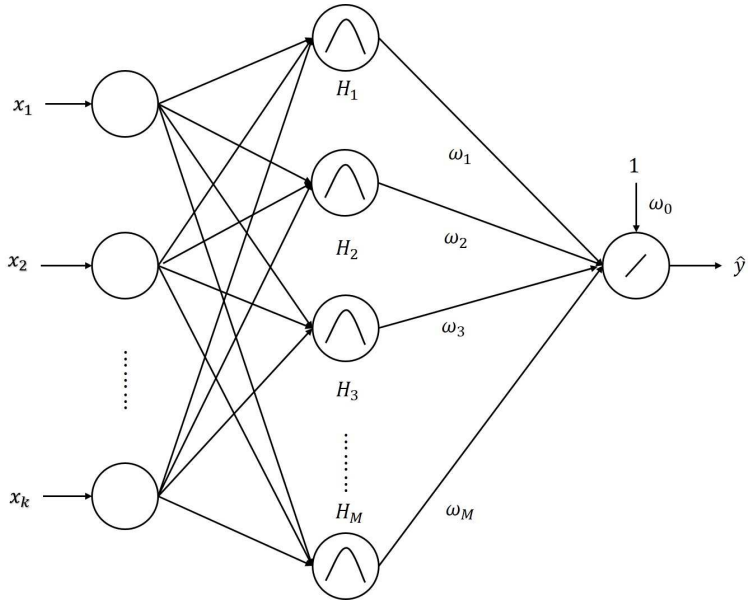

Figure 2.14: Structure of the RBFNN [58]

layers and parameters of the radial basis function are fixed. The second stage training is supervised training: both input and target data are used to train the weights for the second layers with the following minimization problem:

$$H^T H W^T = H^T T \tag{2.38}$$

where T is the vector of the $[t_i]$ and t_i is the target value of $\hat{y}$ and the minimization function is:

$$\min \sum_{i=1}^{N} (\hat{y}_i - t_i)^2 \tag{2.39}$$

2.6 Summary

In this chapter, some background information is given for sunlight, PV cells, PV systems and forecasting models. Through the model of sunlight and PV cells, the maximum PV output can be estimated for a small PV system at a certain location. However, this estimation is only feasible for a very small PV system and does not cover shading issues. In practice,

the output forecast needs forecasting models. Three common forecasting models, ARIMA, LS-SVM and RBFNN, were explained in detail in this chapter. Those models will be further used in Chapters 3 and 4, and the aim of this description is to introduce modeling and the training process.

Chapter 3

Array Level Short-term PV Output Forecasting

3.1 Introduction

In this chapter, the proposed array level forecasting tool is introduced and developed. This tool aims at 24-hour-ahead hourly forecasting. Three PV sites used in this study are introduced in Section 3.2. An analysis of weather and day length influence on the output pattern and intra-day power fluctuation is presented in Section 3.3. At all sites, meteorological and other variables were measured and forecast along with PV output. The relationship between these output-related variables and PV power output is analysed in Section 3.4. The proposed array level forecasting tool is introduced in Section 3.5. This tool includes a data-preprocessing engine and a forecasting engine. The data-preprocessing engine is built using a similar day method and the forecasting engine is built by three forecasting models: LS-SVM. RBFNN and ARIMA. Comparison with a simple persistence model is also provided. The numerical results generated by the proposed forecasting tool are presented and compared with the literature in Section 3.6.

3.2 Site and Data Description

Data acquiring costs many effect in the early stage since this research aims to select array level PV system that have a good quality and quantity of recorded data as well as forecast data. Moreover, this research also aims to select sites that have different weather distribution and different forecast data source. The datasets used in this chapter were collected from three different PV sites and are highlighted in Figure 3.1. The first site (Site 1) is located at San Diego, USA (32.86°N, 117.25°W), the second site (Site 2) is located at Braedstrup, Denmark

(55.97°N, 9.61°E) and the third site (Site 3) is located at Catania, Italy (37.41°N, 15.04°E). Sites 1 and 3 have a small latitude difference but a large longitude difference; while sites 2 and 3 have a relatively large latitude difference but a small longitude difference.

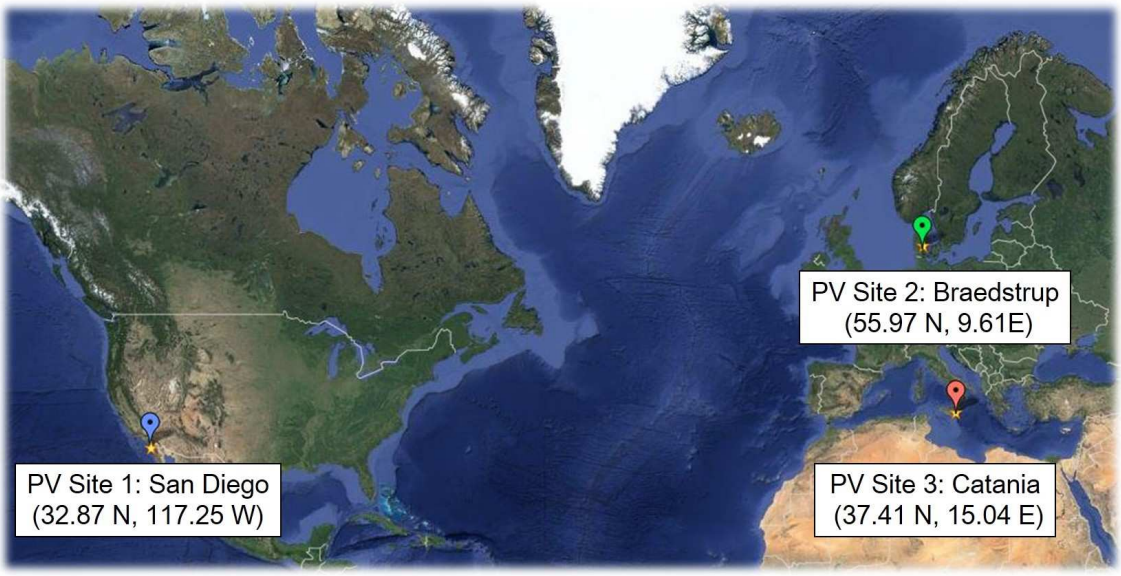

Figure 3.1: Location of three PV sites: San Diego, Braedstrup, and Catania

The geographical difference causes a difference in day length. As shown in Figure 3.2, day length changes at Braedstrup are larger than at Catania and San Diego. The day length at Braedstrup increases from 7 hours in January to around 18 hours in June, while the day length in San Diego and Catania ranges from around 10 hours to around 14 hours. Recalling the physical model in the Background Chapter, the irradiance is also stronger in June than that in January and September.

The geographical difference also results in different local weather. Figure 3.3 shows the historical daily weather over a one-year period. San Diego and Catania are located closer to the equator and have no snowy days during winter. Braedstrup has more rain and fog compared to the other two locations. Moreover, these three locations have limited sunny weather(e.g less than 50 days). San Diego and Catania have a significant number of partly cloudy days [59].

The recorded period, resolution and providers of the data are given in Table 3.1 $\sim$ 3.3.

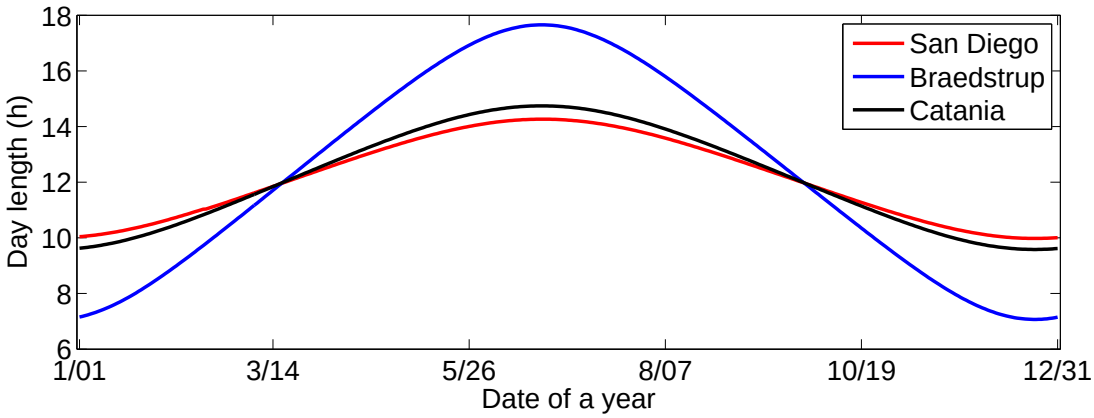

Figure 3.2: Day length changing over a year at San Diego, Braedstrup and Catania

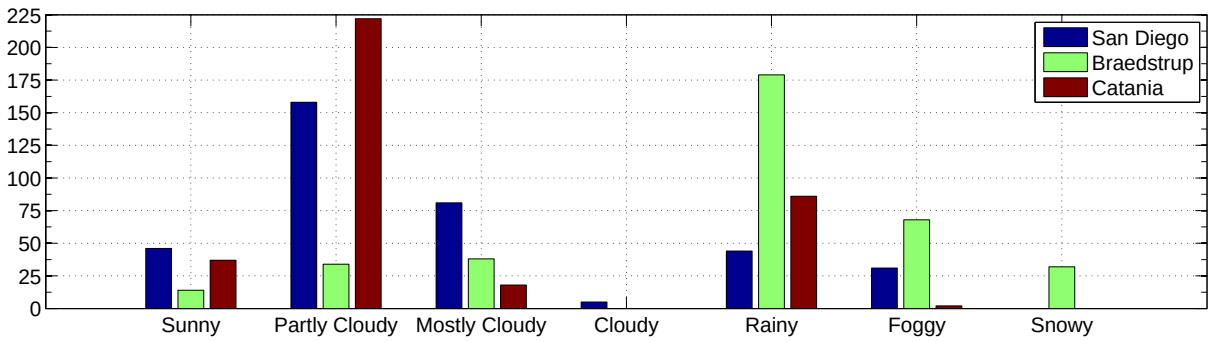

Figure 3.3: Daily weather distribution over a year for the three PV sites: San Diego, Braedstrup, and Catania

The first site is a 49.2 kW PV system. This system has 240 Kyocera KD 205 GXLP PV modules and was installed at a tilt of 10°and azimuth of 180°. Available data for this site are summarized in Table 3.1. Recorded global horizontal irradiance was measured at Hubbs Hall (300 m away from the PV site) using a LICOR Li-200SZ silicon-186 pyrometer sampled at 1 second [60]. Other recorded data were measured at this site for the period from July 1, 2010 to December 31, 2011, with a resolution of 15 minutes [61]. Forecast global horizontal irradiance at one-hour intervals was provided by the National Oceanic and Atmospheric Administrations (NOAA) using the Weather Research and Forecasting North American Mesoscale (WRF-NAM) model [11].

The second site has 21 PV systems built as a project of Sol300 in Denmark [23]. Those PV systems are made by BP 585 modules and BP GCI 1200 inverters. The rated power varies

Table 3.1: Data summary for Site 1 (San Diego)

Data	Resolution	Available Period	Source
Recorded Power Output	15 min	01-07-2010 to 31-12-2011	Birch Aquarium
Recorded Ambient Temperature	15 min	01-07-2010 to 31-12-2011	Birch Aquarium
Recorded Cell Temperature	15 min	01-07-2010 to 31-12-2011	Birch Aquarium
Recorded Wind Speed	15 min	01-07-2010 to 31-12-2011	Birch Aquarium
Recorded Global Horizontal Irradiance	1 sec	01-01-2011 to 30-06-2011	Hubbs Hall
Forecast Global Horizontal Irradiance	60 min	01-01-2011 to 31-12-2011	NOAA

from 1020 W to 4080 W, the azimuth angle varies from 100°to 230°, and the tilt angle varies from 15°to 45°. Available data for Site 2 are summarized in Table 3.2. Recorded power was measured from these 21 PV systems at 15-minute intervals. Forecast data, including global horizontal irradiance, high cloud cover, medium cloud cover, low cloud cover, total cloud cover, fog, ambient temperature and wind speed were provided by the Danish Meteorological Institute (DMI) [62]. The forecast global horizontal irradiance is recorded at three-hour intervals, and other forecast data is recorded at four-hour intervals.

Table 3.2: Data summary for Site 2 (Braedstrup)

Data	Resolution	Data Period	Source
Recorded Power Output	15 min	01-01-2006 to 31-12-2006	Braedstrup
Forecast Global Horizontal Irradiance	3 hour	01-01-2006 to 31-12-2006	DMI
Forecast Ambient Temperature	4 hour	01-01-2006 to 31-12-2006	DMI
Forecast High Cloud Cover	4 hour	01-01-2006 to 31-12-2006	DMI
Forecast Low Cloud Cover	4 hour	01-01-2006 to 31-12-2006	DMI
Forecast Medium Cloud Cover	4 hour	01-01-2006 to 31-12-2006	DMI
Forecast Total Cloud Cover	4 hour	01-01-2006 to 31-12-2006	DMI
Forecast Fog	4 hour	01-01-2006 to 31-12-2006	DMI
Forecast Wind Speed	4 hour	01-01-2006 to 31-12-2006	DMI

The third site is a 5.21 kW system, and available data for this site are summarized in Table 3.3. Recorded data comprises solar altitude, global horizontal irradiance, direct normal irradiance, ambient temperature, and power output. Forecast data, including solar altitude, global horizontal irradiance, direct normal irradiance, total cloud cover and ambient temperature were provided by the Regional Atmospheric Modeling System (RAMS). These data were all recorded at one-hour intervals for 2010.

Table 3.3: Data summary for Site 3 (Catania)

Data	Resolution	Data Period	Source
Recorded Power Output	60 min	01-01-2010 to 31-12-2010	Catania
Recorded Solar Altitude	60 min	01-01-2010 to 31-12-2010	Catania
Recorded Global Horizontal Irradiance	60 min	01-01-2010 to 31-12-2010	Catania
Recorded Direct Normal Irradiance	60 min	01-01-2010 to 31-12-2010	Catania
Recorded Ambient Temperature	60 min	01-01-2010 to 31-12-2010	Catania
Forecast Global Horizontal Irradiance	60 min	01-01-2010 to 31-12-2010	NWP RAMS
Forecast Direct Normal Irradiance	60 min	01-01-2010 to 31-12-2010	NWP RAMS
Forecast Ambient Temperature	60 min	01-01-2010 to 31-12-2010	NWP RAMS
Forecast Solar Altitude	60 min	01-01-2010 to 31-12-2010	NWP RAMS
Forecast Total Cloud Cover	60 min	01-01-2010 to 31-12-2010	NWP RAMS

3.3 Analysis of PV Power Data

In this section, the influence of weather and day length on power output and intra-day power fluctuations of PV power output data are analyzed. These characteristics are challenging for PV forecasting and are also entry points for the proposed forecasting tool.

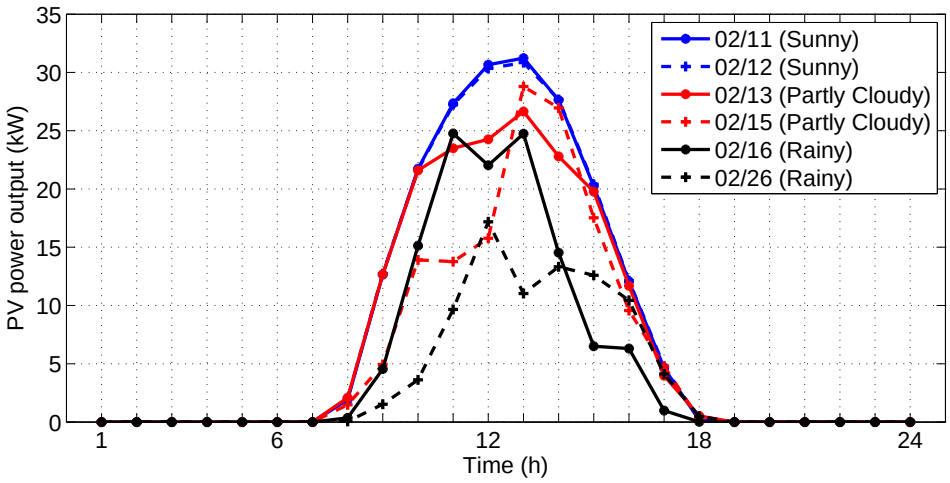

Figure 3.4: Power output under different weather condition at San Diego

3.3.1 Influence of Weather and Day Length on the PV Output

PV power output fluctuates with daily weather [30,33]. Different weather will impose different levels of influence on the output pattern. Figure 3.4 shows 6 days of PV power output at San Diego in February. For sunny weather, the PV power output did not change significantly

(e.g two blue lines in this figure). Under partly cloudy weather conditions, the PV power output on February 13 and February 15 are different (e.g two red lines in this figure). On February 15, the PV morning output was lower than the PV morning output on February 13. The two black lines represent the PV power output under rainy weather conditions. Compared to the power output under partly cloudy weather conditions, the output was even lower on rainy days. Since the rainfall of February 26 (34.04 mm [59]) is larger than that on February 16 (10.92 mm [59]), PV power produced on February 26 is lower than that of February 16. In summary, different weather affects the power at different levels. Generally, rainy and snowy weather decrease the power output the most, followed by cloudy weather and sunny weather [27].

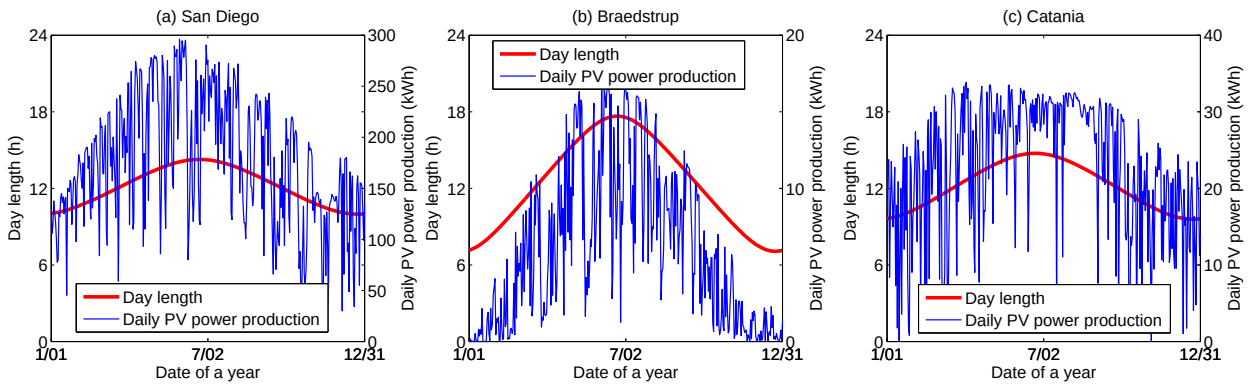

Figure 3.5: Relationship between daily PV power production and day length at (a) San Diego; (b) Braedstrup; (c) Catania

The strength and duration of irradiance has an annual pattern which is reflected in the corresponding PV power output [31]. Day length can reflect the changes of irradiance. High latitude locations have larger day length changes. The daily PV power production and corresponding day length for the three PV sites are plotted in Figure 3.5 (a) - (c). These three figures show that the daily PV power production follows the trend of day length. For example, the day length at Braedstrup changes more severely than the other two locations and the corresponding daily PV power production at Braedstrup changes more as well. Thus, the power difference is more obvious for a certain period (e.g one month) in Braedstrup even

if all days in this period have sunny weather.

3.3.2 Characteristic of Intra-Day Power Fluctuation

The intra-day power fluctuation is mainly caused by daily weather changes. In this section, these fluctuations are analysed for hourly power date [63]. As analysed in Section 3.3.1, PV power is influenced by local weather, and days with different weather usually show a different output pattern. As shown in Table 3.4, the weather is usually variable. The weather varies significantly over a one-week period during February for all three PV sites. For example, the weather was different each day at San Diego for February 12 to February 17. The corresponding normalized power output (Power is normalized by maximum output) at these three sites is plotted in Figure 3.6. Because of the unstable weather, power output at these three sites shows a daily fluctuation, aside from February 11 to February 12 at San Diego which were two sunny days. The other days present a different output pattern, even for days with similar weather. Power at San Diego had less fluctuation compared to the other two sites.

Table 3.4: One week daily weather report for San Diego, Braedstrup and Catania

Date	11-Feb	12-Feb	13-Feb	14-Feb	15-Feb	16-Feb	17-Feb
San Diego	Sunny	Sunny	Partly Cloudy	Foggy	Partly Cloudy	Rainy	Partly Cloudy
Breadstrup	Foggy	Foggy	Foggy	Snowy	Rainy	Rainy	Rainy
Catania	Partly Cloudy	Partly Cloudy	Rainy	Rainy	Partly Cloudy	Rainy	Foggy

In order to quantify the overall intra-day fluctuation, a differenced power data P_t^D (intra-day fluctuation series) was generated in the form of:

$$P_t^D = P_{t+24} - P_t \tag{3.1}$$

where, P_t is the original power data. P_t^D represents the intra-day fluctuation for time t at day D. Figure 3.7 shows the intra-day fluctuation from February 11 to 17 at these three sites. For example, the first 24 hours represents the hourly power output difference between

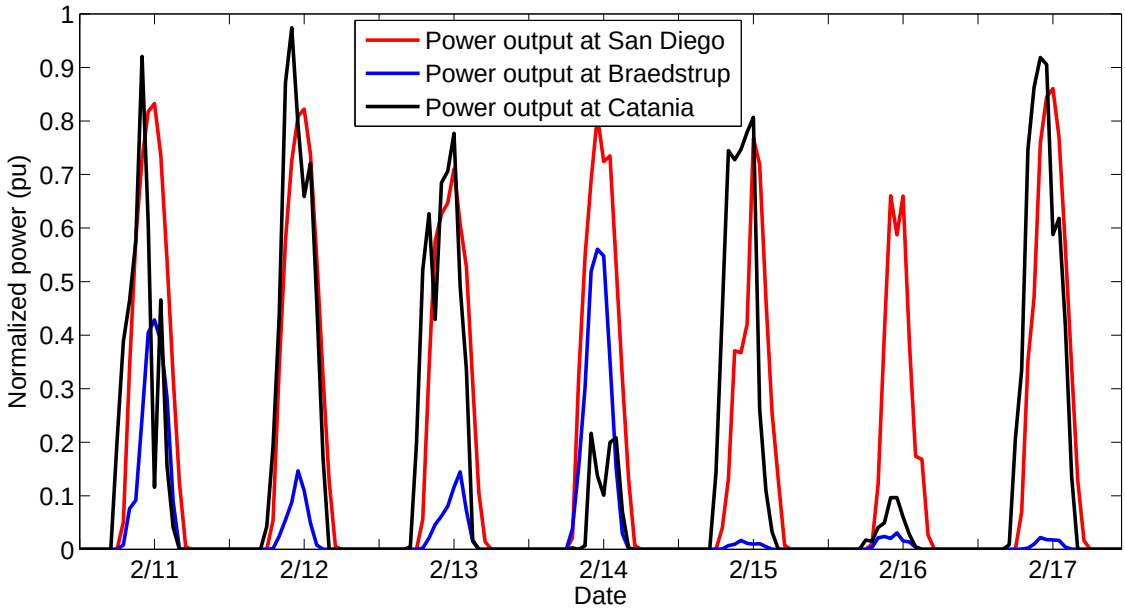

Figure 3.6: Power output at San Diego, Braestrup and Catania from February 11 to 17

February 12 and 11. Since the power output was very similar at San Diego during these two days, the corresponding intra-day fluctuation series was very low, as shown in the red line from time 1 to 24 in Figure 3.7. The power output was different at Braedstrup and Catania during these two days, so the corresponding intra-day fluctuation was large, shown in the blue and black lines from time 1 to 24 in Figure 3.7.

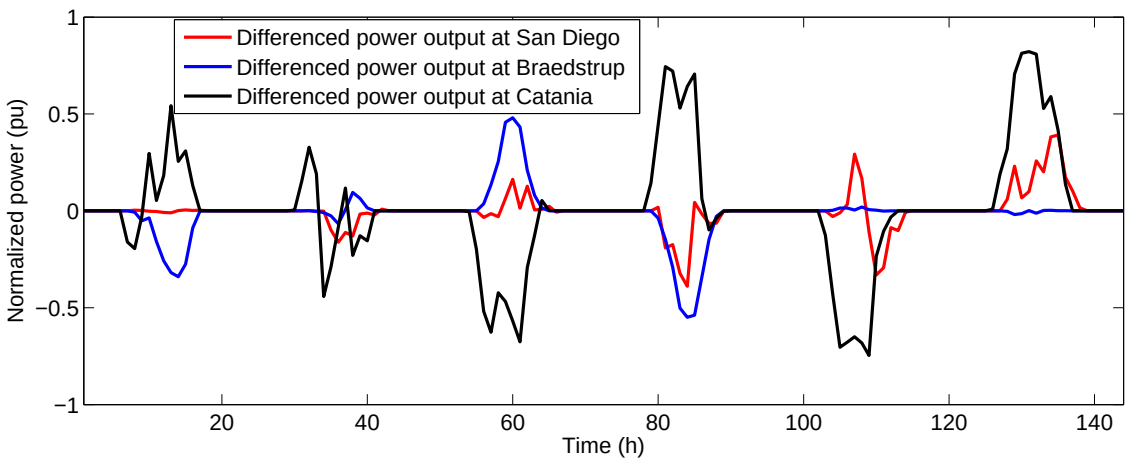

Figure 3.7: Intra-day fluctuation series for days from February 11 to 17 at San Diego, Braestrup and Catania

A short time window is not enough to represent the overall intra-day fluctuation level

of different PV sites. Thus, the intra-day fluctuation series P_t^D was generated for a one-year period in Figure 3.8. The fluctuation were significant for all these three sites. The fluctuation were sometimes up to 100%. Figure 3.8 (a) shows that the daily power difference at San Diego was generally the same over the year, the maximum intra-day fluctuation was 82.2%, and the standard deviation of the intra-day fluctuation series was 12.4%. Figure 3.8 (b) shows that the daily power difference at Braedstrup was higher during the middle of the year, the maximum intra-day fluctuation was 83.0%, and the standard deviation of the intra-day fluctuation series was 14.0%; Figure 3.8 (c) shows that the daily power difference at Catania was lower at the middle of the year, the maximum intra-day fluctuation was 90.1%, and the standard deviation of the intra-day fluctuation series was 16.2%. Compared to the other two sites, Catania had more changes in daily power.

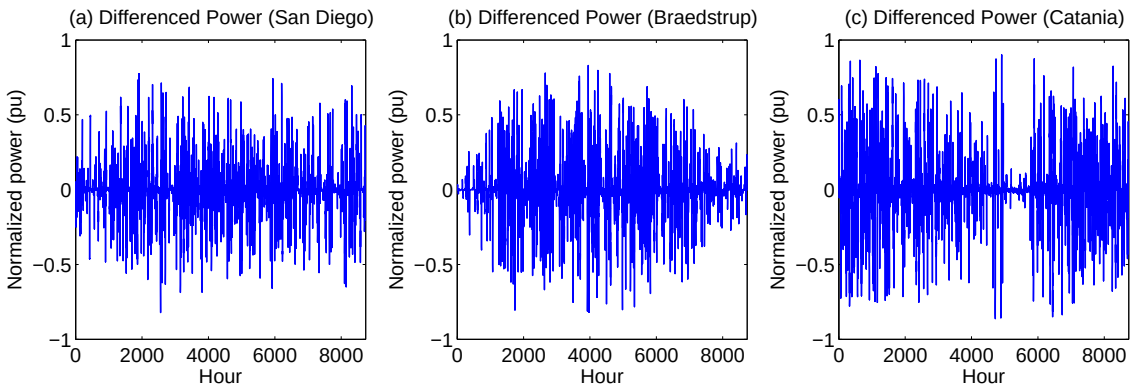

Figure 3.8: Intra-day fluctuation series over a year at: (a) San Diego (b) Braedstrup (c) Catania

3.3.3 Remarks

In summary, this section analysed the influence of weather and day length on the power output and intra-day power fluctuation of array level PV power output. These characteristics are challenging for array level forecasting. Weather changes are the inherent reason for the intra-day fluctuation of solar power output. A location with a high intra-day fluctuation will be hard to predict. As well, locations with a high latitude have larger daily irradiance

or day length changes. Thus, when choosing the training days for a certain model, high latitude locations should not use large training sets, as the absolute power output difference of training days is bigger. The analysis of these characteristics also shows a direction for building an array level forecasting tool. This tool needs to have the ability to incorporate daily power fluctuations.

3.4 Relationship Between Output-Related Variables and PV Output

As analysed in Section 3.3.1, the weather has a significant influence on the output of PV systems. However, daily weather is a fuzzy term, and thus, PV power output under the same weather may have different patterns and PV power output under different weather may show similar patterns. In the following, output-related variables that may influence the power output are analyzed using scatter plot and correlation coefficient of determination of the correlation coefficient of the fitting of a straight line to the data, R^2 [64]. This analysis is done for data from a one-year period, i.e., 2011 for San Diego, 2006 for Braedstrup and 2010 for Catania. The data interval used in this analysis is one hour for San Diego and Catania. For Braedstrup, the analysis is done using three-hour-intervals for GHI and four-hour-intervals for other variables, because hourly data was not available for this site.

3.4.0.1 Global Horizontal Irradiance

Global horizontal irradiance was frequently used in PV output forecasting (e.g., in [23, 30]). The relationship between recorded global horizontal irradiance (I_R) and PV power is presented in Figure 3.9. The correlation coefficient is 0.8504 for San Diego and 0.9095 for Catania. Although there are errors in forecasting global horizontal irradiance, the forecast global horizontal irradiance still has a clear relationship with PV power output, as shown in Figure 3.10, where Figure 3.10 (a) and (c) shows the relationship between forecast global horizontal irradiance (I_F) and power for San Diego and Catania. Since the forecast irradiance for Braedstrup is sampled at three-hour intervals, the power output for Braedstrup is

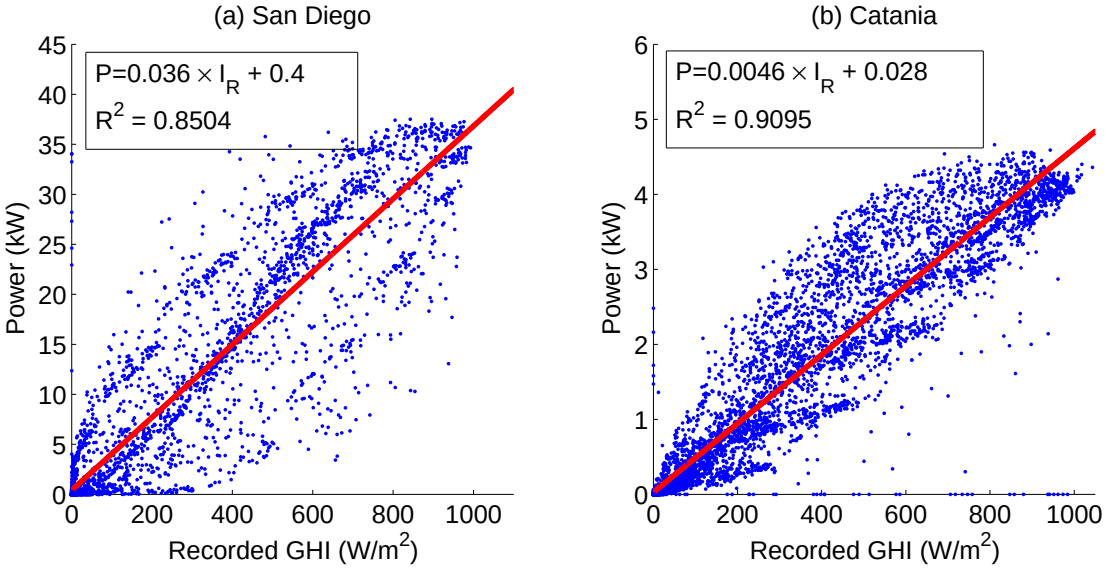

Figure 3.9: Relationship between recorded global horizontal irradiance (I_R) and PV power output at (a) San Diego; (b) Catania

re-sampled to three-hour interval and their relationship is plotted in Figure 3.10 (b). Compared to recorded global horizontal irradiance, the corresponding correlation coefficient of determination between forecast global horizontal irradiance and PV power output slightly drops from 0.8504 to 0.822 at San Diego and drops from 0.9095 to 0.843 at Catania.

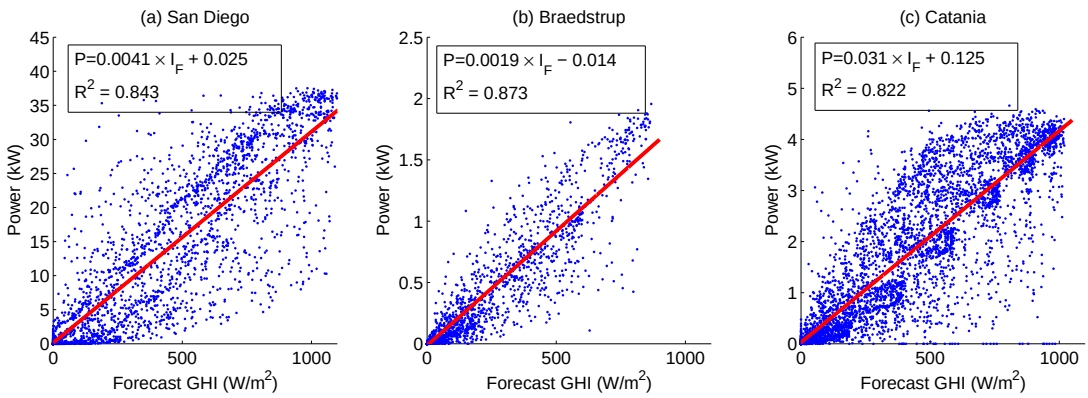

Figure 3.10: Relationship between forecast global horizontal irradiance (I_F) and PV power output at (a) San Diego; (b) Braedstrup; (c) Catania

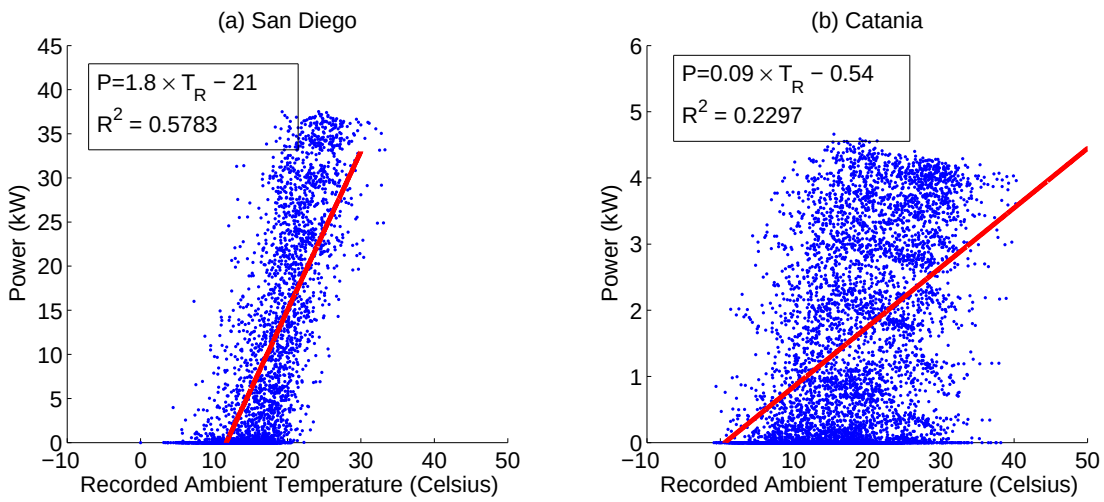

Figure 3.11: Relationship between recorded ambient temperature (T_R) and PV Power output at (a) San Diego; (b) Catania

3.4.0.2 Ambient Temperature

The power output is also influenced by temperature. As described in Chapter 2, power output drops when the temperature increases. For example, the power of the crystalline silicon solar cell drops by approximately $0.4\%/K - 0.5\%/K$ and the power of amorphous silicon solar cells drops by approximately $0.2\%/K - 0.25\%/K$ [44]. Cell temperature is rarely measured and is not a meteorological variable; thus, ambient temperature is usually used. However, higher temperature usually occur during sunny days and in summer, when the strength of solar radiation is strong. Thus, ambient temperature should be a positive indicator of power output. To further examine this variable, the linear relationship between PV power output and recorded ambient temperature (T_R) is plotted, as shown in Figure 3.11. The correlation coefficient of determination is 0.5783 between recorded ambient temperature and PV power at San Diego in Figure 3.11 (a), and the correlation coefficient of determination is 0.2297 for Catania in Figure 3.11 (b). Figure 3.12 shows the relationship between forecast ambient temperature (T_F) with PV power at Braedstrup at four-hour intervals and Catania at one-hour intervals. The corresponding R^2 is 0.2485 for Braedstrup and 0.2153 for Catania. Catania is the only location with both recorded and forecast ambient temperature in this

study. Figure 3.11 (b) and Figure 3.12 (b) show R^2 drops slightly when using forecast ambient temperature. In general, the linear relationship between PV power output and ambient temperature is weak. However, ambient temperature was widely used in PV forecasting. For example, it was used as model input in [26,65] and data preprocessing criteria in [32,33].

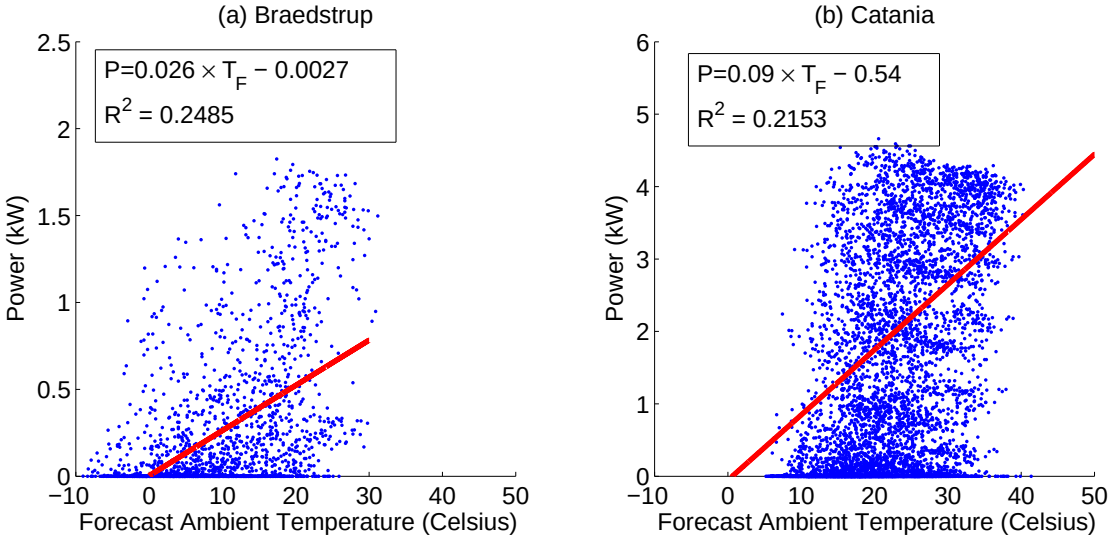

Figure 3.12: Relationship between forecast ambient temperature (T_F) and PV power output at (a) Braedstrup; (b) Catania

3.4.0.3 Wind Speed

Wind can reduce the temperature of the PV cells and improve the PV output; so it was chosen for analysis. The relationship between wind speed and PV power output is very weak. Figure 3.13 (a) shows the relationship between recorded wind speed (WS_R) and PV power output at San Diego with one-hour intervals and Figure 3.13 (b) shows the relationship between forecast wind speed (WS_F) and PV power output at Braedstrup with four-hour intervals. Both recorded and forecast wind speed has a low correlation coefficient of determination with PV power output. Thus, there is no linear relationship between wind speed and PV power. However, wind speed has also been used in previous PV forecast research [30, 65].

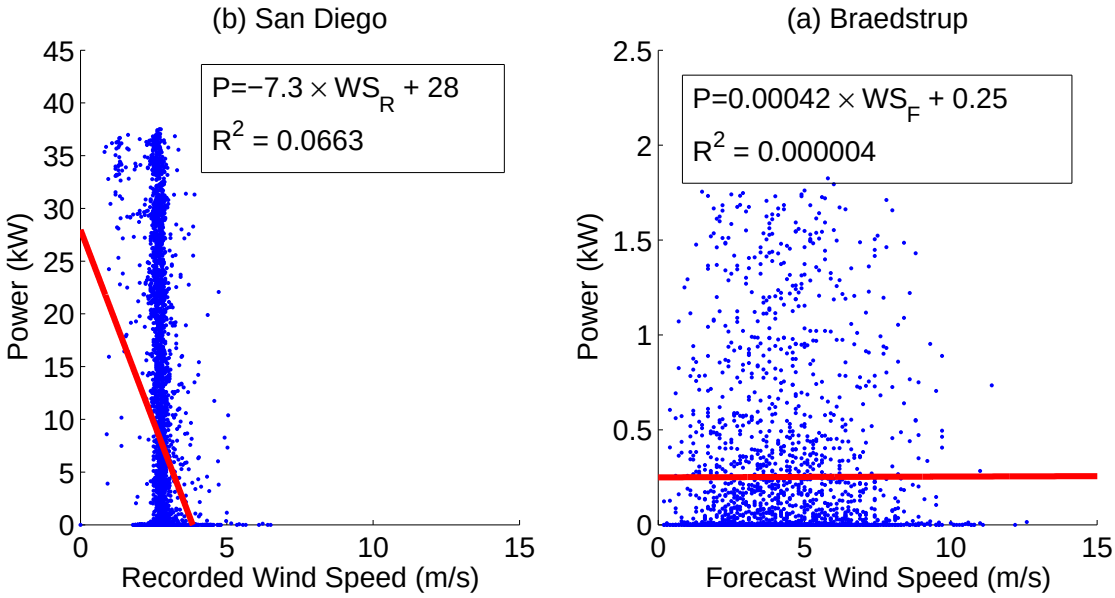

Figure 3.13: Relationship between wind speed and PV power output at (a) San Diego; (b) Braedstrup

3.4.0.4 Direct Normal Irradiance

Direct normal irradiance refers to radiation received on a unit area that is normal to the sun. This variable is only available at Catania. Figure 3.14 (a) presents the relationship between recorded direct normal irradiance (DNI_R) and PV power output. A clear linear relationship can be observed. Figure 3.14 (b) shows the relationship between forecast direct normal irradiation (DNI_F) and PV power output. Compared to recorded direct normal irradiance, the correlation coefficient of determination is lower but still significant.

3.4.0.5 Solar Altitude

Solar altitude (SA) is the angular height of the sun measured from the horizon. Compared to the above output-related variables, there is no forecasting error for solar altitude. Thus, recorded solar altitude and forecast solar altitude have the same relationship with PV power output, as shown in Figure 3.15. They have the same correlation coefficient of determination which is 0.7932.

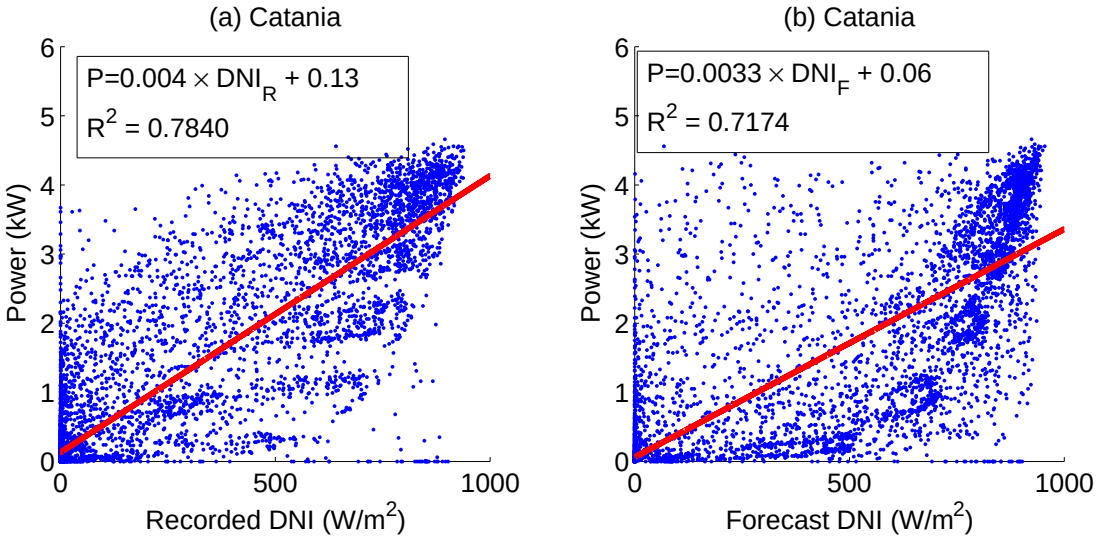

Figure 3.14: Relationship between direct normal irradiance (DNI) and PV power output at Catania: (a) Recorded DNI; (b) Forecast DNI

3.4.0.6 Fog

Fog refers to suspended water or small ice platelets in the air [66]. This variable ranges from 0 to 1, and 0 refers to the fog occurrence probability as 0% and 1 refers to the fog occurrence probability as 100%. Figure 3.16 shows there is almost no linear relationship between forecast fog (F_F) and PV power output at Braedstrup (four-hour intervals data). However, it can be observed that with fog, the power output is usually lower (e.g. the power is lower than 0.5 kW when the fog is 0.4).

3.4.0.7 Cloud

The cloud cover ranges from 0 to 1 and includes low cloud cover (LC_F), medium cloud cover (MC_F), high cloud cover (HC_F) and total cloud cover (TC_F). 0 refers to no clouds in the sky, and 1 refers to the sky being totally covered by cloud [30]. These four types of forecast cloud cover are all available at Braedstrup at four-hour intervals. Catania only has total cloud cover at one-hour intervals. In Figure 3.17, the relationship between all four types of forecast cloud cover and PV power output at Braedstrup are plotted. Similar to the relationship between fog and PV power output, the linear relationship between all

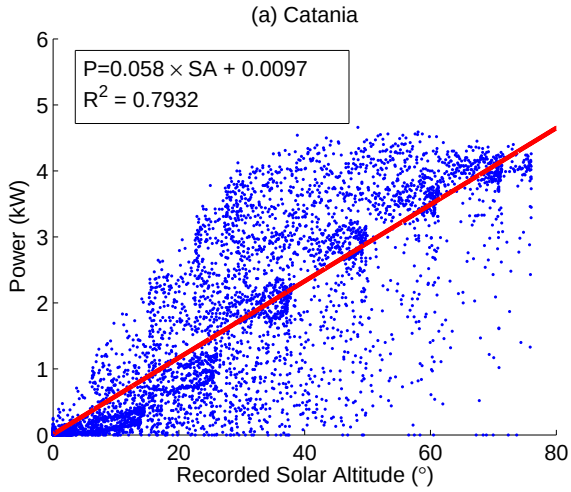

Figure 3.15: Relationship between solar altitude (SA) and PV power output at Catania

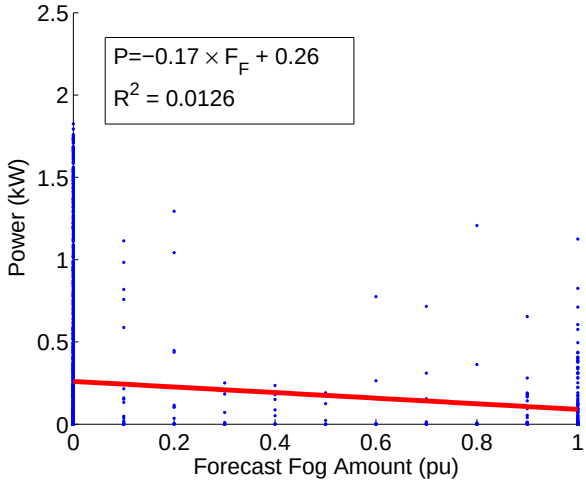

Figure 3.16: Relationship between forecast fog (F_F) and PV power output at Braedstrup

these four types of forecast cloud cover and PV power output is low. Through comparison between Figure 3.17 (a) and Figure 3.17 (d), forecast total cloud cover has a higher correlation coefficient of determination (0.0802) than forecast low cloud cover (0.0681), forecast medium cloud cover (0.0185) and forecast high cloud cover (0.0095). Catania only has forecast total cloud cover, and the relationship between it and PV power output is shown in Figure 3.18.

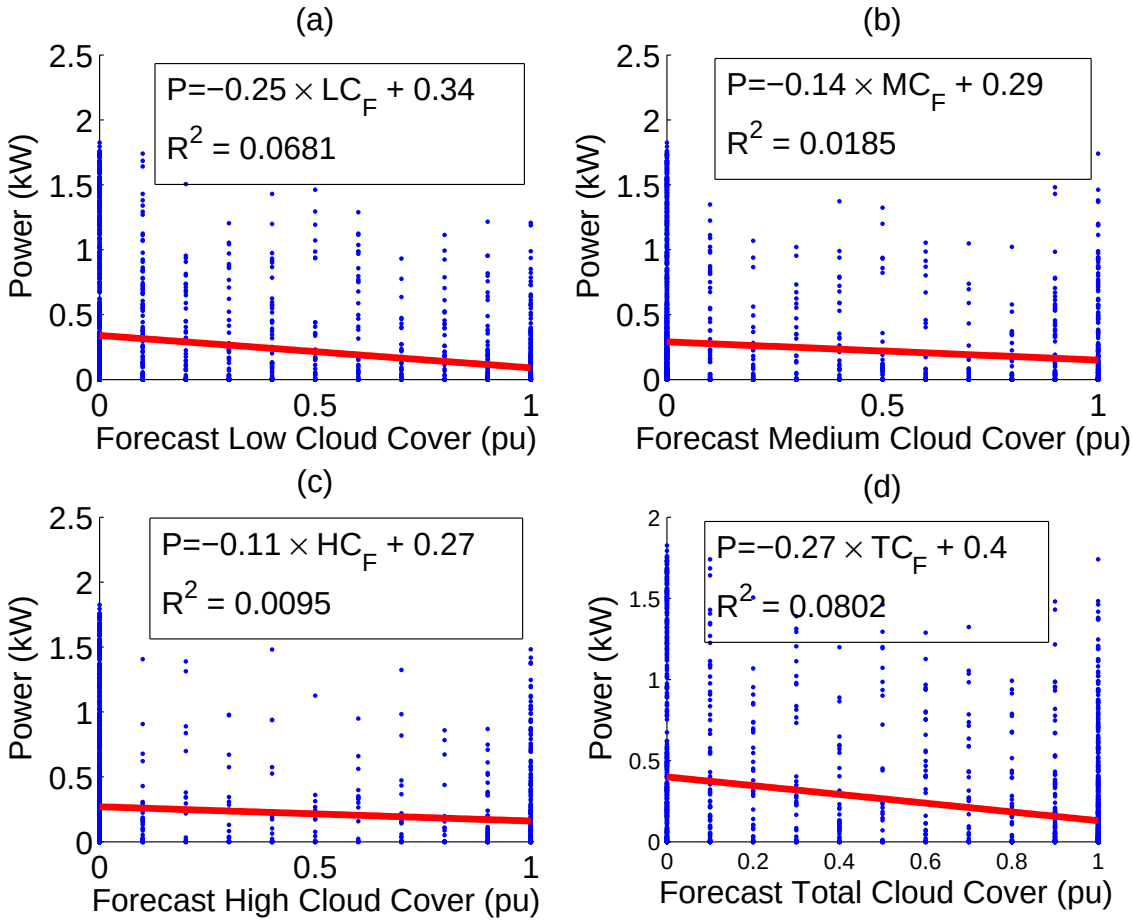

Figure 3.17: Relationship between different type of forecast cloud cover and PV power output at Braedstrup: (a) Forecast Low Cloud Cover (LC_F); (b) Forecast Medium Cover (MC_F); (c) Forecast High Cover (HC_F); (d) Forecast Total Cloud Cover (TC_F)

3.4.0.8 Remarks

Global horizontal irradiance has the closest relationship with PV power output of all the output-related variables, including ambient temperature, wind speed, direct normal irradiance, solar altitude, fog and cloud. Since the correlation coefficients of determination between ambient temperature, direct normal irradiance and solar altitude are high, these three variables are reasonable inputs for a forecasting model. The relationship between wind speed, fog, cloud and PV power output were not strong. However, they have a theoretical relationship with PV output. Besides, wind speed and cloud have been used in previous studies. Thus, they are also selected as candidate variables for further analysis.

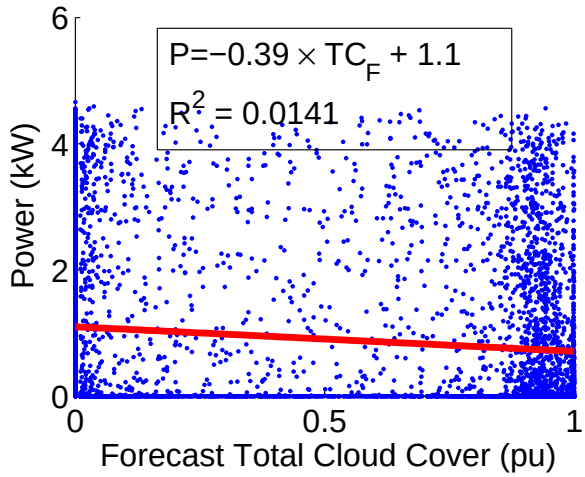

Figure 3.18: Relationship between forecast total cloud cover (TC_F) and PV power output at Catania

3.5 Forecasting Tool for Array Level PV Output Forecasting

3.5.1 Overview of the Forecasting Tool

In this section, the framework of the proposed array level forecasting tool is introduced and developed. Figure 3.19 shows the time line of this forecasting tool. This tool runs at 24:00 on day $d-1$, where day d is the forecast day. The goal is to predict the hourly power output for day d from 1:00 to 24:00.

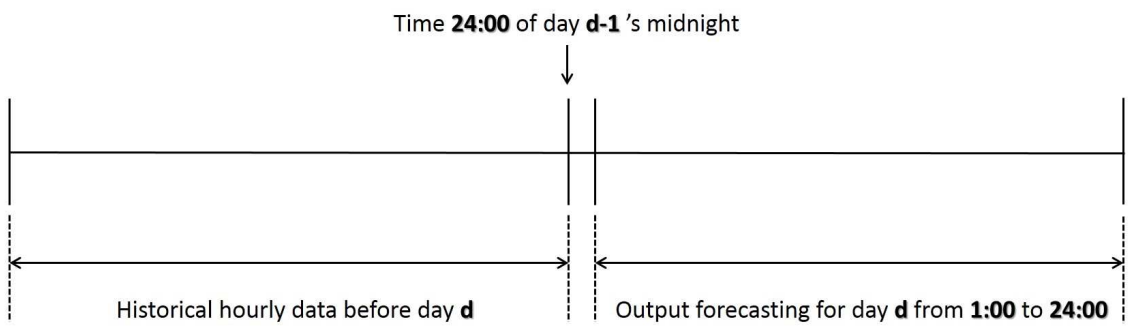

Figure 3.19: Time line of forecasting process

This forecasting tool includes two components: the Data-Preprocessing Engine and the Forecasting Engine, as shown in Figure 3.20. The data-preprocessing engine is based on the

similar day method. Through similar days searching, a group of days that have the highest similarity with the forecast day will be selected. The historical power from the selected similar days are fed into the Forecasting Engine to do an autoregressive forecasting. The external inputs (e.g. irradiance), are used in the similar day process, not in the forecasting engine. In this study, LS-SVM, RBFNN, ARIMA and a persistence model were chosen to build the Forecasting Engine.

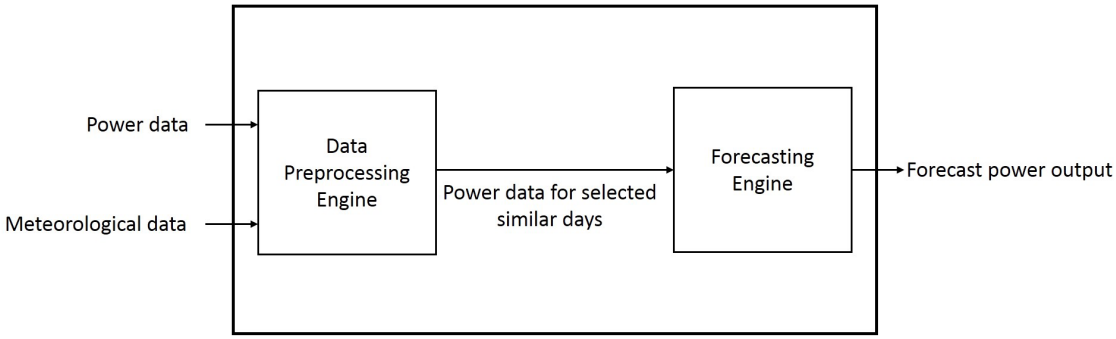

Figure 3.20: Framework of array level forecasting tool

3.5.2 Data-Preprocessing Engine

The Data-Preprocessing Engine is based on a similar day method which is described below. In this engine, an euclidean distance, which is a frequently used similar day method, is utilized to measure the similarity [32,67]. In the first stage, the euclidean distance of recorded power output is calculated between forecast day d and previous $d-1$ days. Note that the model building stage has access to historical recorded power data. This distance represents the actual output difference between day d and previous historical days. In the second stage, the euclidean distance of each output-related variable was calculated between forecast day d and previous $d-1$ days. Since these output-related variables are forecast, this distance is the forecast distance. The comparison of forecast distance and actual distance shows which output-related variable is more suitable for building forecast distance. In the third stage, these output-related variables are given a weight and used to build the hybrid euclidean

distance formula for similar days selection.

3.5.2.1 Stage 1: Euclidean Distance of Recorded PV Power Output

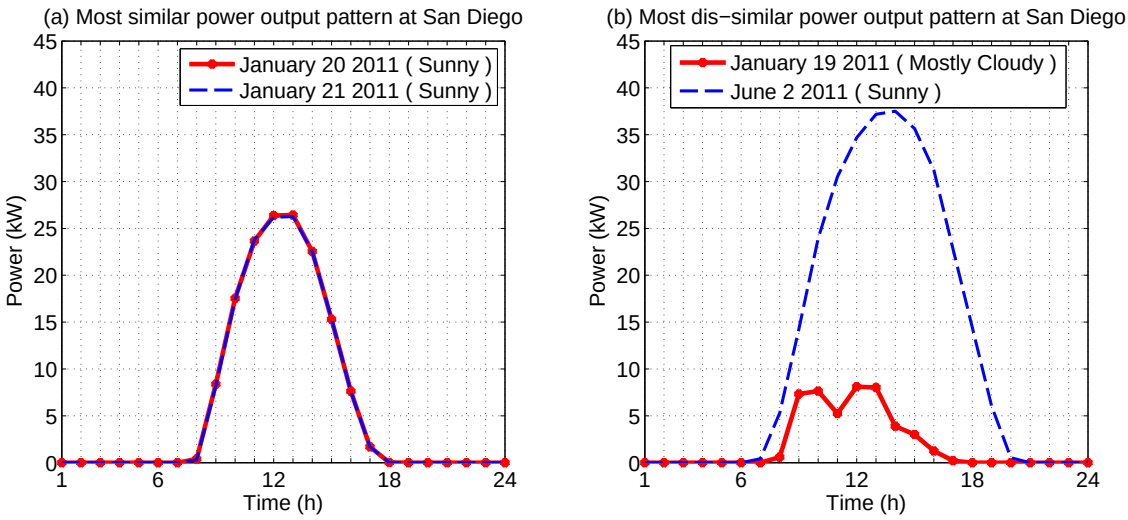

Figure 3.21: Days having the most similar output pattern and most dissimilar output pattern at San Diego (Euclidean distance between January 20 and 21 is 0.07 kW and euclidean distance between January 19 and June 2 is 16.30 kW)

The euclidean distance of recorded PV power output for the forecast day d and previous day i is defined as:

$$\mathrm{ED}_{PR}(i,d) = \sqrt{\sum_{h=1}^{24}(PR_i^h - PR_d^h)^2} \tag{3.2}$$

where, PR_i^h is the recorded power at hour h for day i, PR_d^h is the recorded power at hour h for day d and $i \in [1, d-1]$. For example, for San Diego, January 20 and 21, 2011 have the smallest distance and, January 19 and June 2 have the largest distance, as presented in Figure 3.21. The power output on January 20 was very close to the power output on January 21, whereas, the power output on January 19 was significantly different from the power output on June 2. Table 3.5 summarizes the weather and day length for those four days. The day length for January 20 and 21 differed by only 75 seconds but the day length for January 19 and June 2 has approximately a four-hour difference. Additionally, January 20 and 21 both had sunny weather; January 19 and June 2 had different weather.

Table 3.5: Weather and day length information for the most similar days and dis-similar days

Day Index	Date	Weather	Day Length (hh:mm:ss)
20	20-01-2011	Sunny	10:18:55
21	21-01-2011	Sunny	10:20:10
19	19-01-2011	Mostly Cloudy	10:17:42
153	02-06-2011	Sunny	14:07:57

3.5.2.2 Stage 2: Euclidean Distance of Each Candidate Output-Related Variable

The euclidean distance for a candidate output-related variable V is defined as:

$$\mathrm{ED}_V(i,d) = \sqrt{\sum_{h=1}^{N}(V_i^h - V_d^h)^2} \tag{3.3}$$

where, V_i^h is the forecast value of output-related variable V at time h for day i, V_d^h is the forecast value of output-related variable V at time h for day d and $i \in [1, d-1]$. To quantify the overall closeness of the euclidean distance of each output-related variable V and euclidean distance of recorded PV power output, an evaluation index EVA is generated in the form of:

$$\mathrm{EVA}_V = \frac{\sum_{d=2}^{S}\sum_{i=1}^{d-1}\left|\mathrm{ED}_V^N(i,d) - \mathrm{ED}_{PR}^N(i,d)\right|}{(S-1)\times S} \times 100\% \tag{3.4}$$

where, ED_{PR}^N is the normalized value of ED_{PR}, ED_V^N is the normalized value of ED_V, and S is the number of testing days. EVA_V measures the average daily difference between normalized euclidean distance of recorded PV power output and normalized euclidean distance of output-related variable V. EVA_V ranges from 0% to 100%. $\mathrm{EVA}_V = 0\%$ means the forecast euclidean distance of output-related variable V is exactly the same as the euclidean distance of recorded PV power output. $\mathrm{EVA}_V = 100\%$ means the forecast distance is totally different from the actual distance.

Table 3.6 lists EVA_V for each location. Based on the value of EVA_V, the euclidean distance of global horizontal irradiance has the least difference with the euclidean distance of recorded

Table 3.6: EVA_V for different forecast output-related variables at three testing location: San Diego, Catania and Braedstrup

EVA_V	GHI[1]	DNI[2]	DT[3]	WS[4]	TA[5]	SA[6]	LC[7]	MC[8]	HC[9]	TC[10]	Fog[11]
San Diego	14,26%	N/A	20.41%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Catania	16.35%	17.97%	25.33%	N/A	20.75%	25.93%	N/A	N/A	N/A	27.08%	N/A
Braedstrup	11.41%	N/A	21.56%	23.49%	18.34%	N/A	27.23%	32.44%	32.90%	25.65%	34.05%

[1] Global Horizontal Irradiance [2] Direct Normal Irradiance [3] Daytime Hour [4] Wind Speed [5] Ambient Temperature [6] Solar Altitute [7] Low Cloud Cover [8] Medium Cloud Cover [9] High Cloud Cover [10] Total Cloud Cover [11] Fog

PV power output. What's more, based on the the value of EVA_V, GHI is not greatly superior to the other output-related variables in contrast to its superiority in the analysis of the previous section. For example, the correlation coefficient of determination of forecast low cloud cover (LC) is 13 times lower than that of global horizontal irradiance, but the EVA_{LC} is only 3 times lower than the EVA_{GHI}. Through the above analysis, the best forecast output-related variable to build the euclidean distance between forecast day and historical days is the global horizontal irradiance. However, Table 3.6 shows the euclidean distance for other output-related variables is not that far away from the euclidean distance of actual PV power output. Thus, this study tries to build the euclidean distance of days using several different output-related variables and test this forecast euclidean distance against the euclidean distance of days using recorded power output. Global horizontal irradiance has the lowest EVA, hence, it was selected as the main variable. Other output-related variables can be combined with global horizontal irradiance to determine a bivariate euclidean distance ($\text{ED}_B^N(i,d)$) between day d and day i in the form of:

$$\text{ED}_B^N(i,d) = W_{GHI} \times \text{ED}_{GHI}^N(i,d) + W_V \times \text{ED}_V^N(i,d) \tag{3.5}$$

$$W_{GHI} + W_V = 100\% \tag{3.6}$$

$$\text{EVA}_B = \frac{\sum_{d=2}^{S} \sum_{i=1}^{d-1} \left| \text{ED}_B^N(i,d) - \text{ED}_{PR}^N(i,d) \right|}{(S-1) \times S} \times 100\% \tag{3.7}$$

where, W_{GHI} is the weight for global horizontal irradiance and W_V is the weight for an output related variable. By adjusting the value of W_{GHI} from 0% to 100%, this study

evaluates whether the evaluation index EVA_B for this new euclidean distance has a lower value compared to EVA_{GHI}. This test was conducted for all available output-related variables for the three sites.

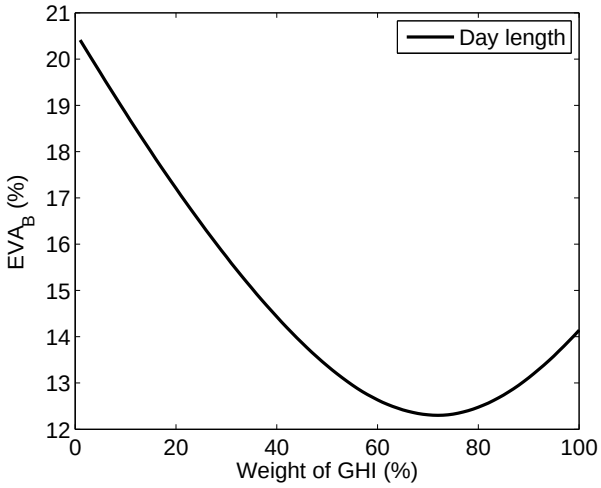

Figure 3.22: Influence of the weight for global horizontal irradiance on the value of EVA_B at San Diego

The only available candidate output-related variable is day length for the PV system at San Diego. Other variables (e.g. Forecast ambient temperature) are not available. Figure 3.22 shows the trend of EVA_B when adjusting the weight of GHI (W_{GHI}). When setting W_{GHI} equals to 71%, EVA_B drops to 12.3%. Compared to EVA_{GHI}, EVA_B is 13.40% lower than EVA_{GHI}. Hence, day length can reflect some similarity between days that cannot be reflected by only global horizontal irradiance.

Candidate secondary output-related variables are ambient temperature, low cloud cover, medium cloud cover, high cloud cover, total cloud cover, fog and wind speed at Braedstrup. Similar to the analysis done for San Diego and Catania, the analysis process is plotted in Figure 3.23. Figure 3.23 shows the trend of EVA_B when adjusting the weight of GHI (W_{GHI}). For example, the green line shows that lowest EVA_B value is achieved when W_{GHI} was 100%. This means that ED_B built by GHI and fog could not achieve a lower EVA_B value. The other lines show the trend of EVA_B when the ED_B is built by GHI and other candidate

secondary output-related variables. When setting corresponding W_{GHI} as 92%, 96%, 86% and 97% for low cloud cover, medium cloud cover, total cloud cover and wind speed, those variables could help GHI to gain a lower EVA_B. EVA_B for global horizontal irradiance and low cloud cover is 2.02% lower than EVA_{GHI}. EVA_B for global horizontal irradiance and medium cloud cover is 0.44% lower than EVA_{GHI}. EVA_B for global horizontal irradiance and total cloud cover is 5.10% lower than EVA_{GHI}. EVA_B for global horizontal irradiance and wind speed is 0.20% lower than EVA_{GHI}.

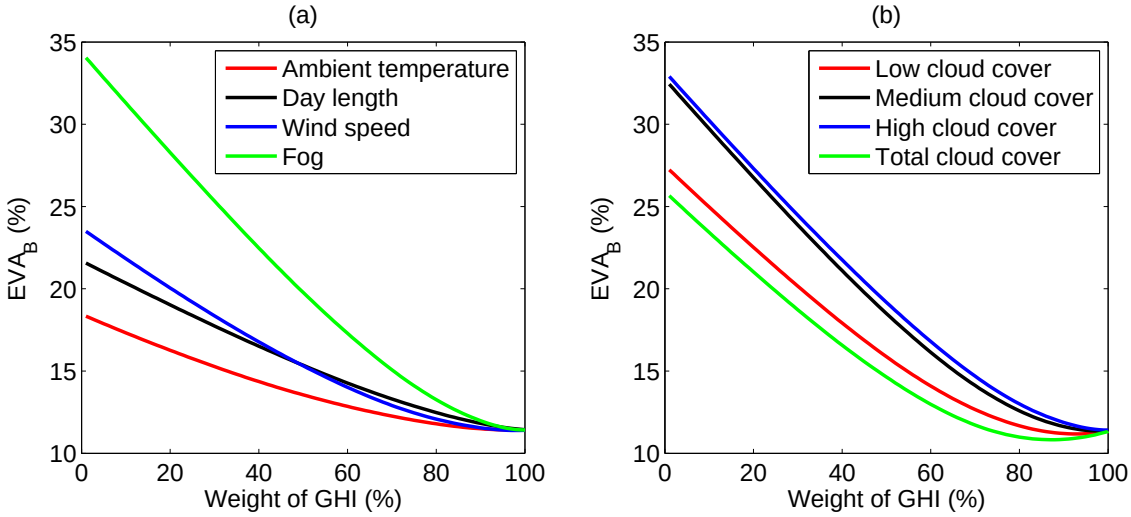

Figure 3.23: Influence of the weight for global horizontal irradiance on the value of EVA_B at Braedstrup

Candidate secondary variables are day length, ambient temperature, total cloud cover, direct normal irradiance and solar altitude for Catania. Figure 3.24 shows the trend of EVA_B when adjusting W_{GHI} value. Aside from solar altitude, other variables can help to decrease EVA_B. EVA_B for global horizontal irradiance and direct normal irradiance is 7.77% lower than EVA_{GHI}. EVA_B for global horizontal irradiance and ambient temperature is 2.81% lower than EVA_{GHI}. EVA_B for global horizontal irradiance and total cloud cover is 9.97% lower than EVA_{GHI}. The corresponding W_{GHI} is 65%, 80% and 76% for direct normal irradiance, ambient temperature and total cloud cover.

Based on Figure 3.22, 3.23, and 3.24, useful candidate secondary output-related variables

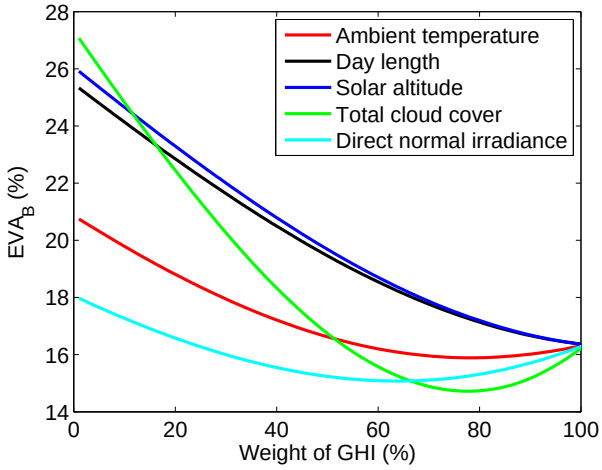

Figure 3.24: Influence of the weight for global horizontal irradiance on the value of EVA_B at Catania

that can help GHI to gain a lower EVA_B are found. Noting, a candidate variable that have high EVA_V could also help GHI to gain a lower EVA_B. Through stimulation, the weight of each secondary variable W_V (W_V=1-W_{GHI}) is found which will be used to build the euclidean distance of combination of output-Related variables.

3.5.2.3 Stage 3: Euclidean Distance of Combinations of Output-Related Variables

For San Diego, the bivariate euclidean distance for day length and global horizontal irradiance has a lower evaluation index. Thus, the hybrid euclidean distance for San Diego is built using global horizontal irradiance and day length using the following algorithm:

$$\frac{F_{GHI}}{F_{DT}} = \frac{W_{GHI}}{W_{DT}} = \frac{W_{GHI}}{1 - W_{GHI}} = \frac{71\%}{1 - 71\%}$$

$$F_{GHI} + F_{DT} = 1$$

$$\text{ED}^N_{HD} = F_{GHI} \times \text{ED}^N_{GHI} + F_{DT} \times \text{ED}^N_{DT} \tag{3.8}$$

where, F_{GHI} is the weight factor for GHI and F_{DT} is the weight factor for day length that is used to build the hybrid euclidean distance. W_{GHI} and W_{DT} is the weight calculated through the above EVA_B analysis for San Diego. W_{GHI} is 71% and W_{DT} is 29%. Noting GHI is the

only candidate variable for San Diego, W_V equals F_V. The final hybrid euclidean distance formula for San Diego is:

$$\mathrm{ED}_{HD}^{N} = 71\% \times \mathrm{ED}_{GHI}^{N} + 29\% \times \mathrm{ED}_{DT}^{N} \tag{3.9}$$

For Braedstrup, low cloud cover and total cloud cover can help to increase the evaluation index of the bivariate euclidean distance. Thus, the hybrid euclidean distance for Braedstrup uses global horizontal irradiance, low cloud cover and total cloud cover using the following algorithm:

$$\begin{aligned}
&\frac{F_{GHI}}{F_{LC}} = \frac{W_{GHI}}{W_{LC}} = \frac{W_{GHI}}{1 - W_{GHI}} = \frac{92\%}{1 - 92\%} \\
&\frac{F_{GHI}}{F_{TC}} = \frac{W_{GHI}}{W_{TC}} = \frac{W_{GHI}}{1 - W_{GHI}} = \frac{86\%}{1 - 86\%} \\
&F_{GHI} + F_{LC} + F_{TC} = 1 \\
&\mathrm{ED}_{HD}^{N} = F_{GHI} \times \mathrm{ED}_{GHI}^{N} + F_{LC} \times \mathrm{ED}_{LC}^{N} + F_{TC} \times \mathrm{ED}_{TC}^{N}
\end{aligned} \tag{3.10}$$

where, F_{LC} is the weight factor for low cloud cover and F_{TC} is the weight factor for total cloud cover that is used to build the hybrid euclidean distance. W_{GHI}, W_{LC} and W_{TC} are the weights calculated through the above EVA_B analysis for Braedstrup. W_{GHI} has two different values that relate to low cloud cover and total cloud cover. The final hybrid euclidean distance formula for Braedstrup is:

$$\mathrm{ED}_{HD}^{N} = 80\% \times \mathrm{ED}_{GHI}^{N} + 7\% \times \mathrm{ED}_{LC}^{N} + 13\% \times \mathrm{ED}_{TC}^{N} \tag{3.11}$$

For Catania, direct normal irradiance, ambient temperature and total cloud cover increase the evaluation index of the bivariate euclidean distance. Thus, the hybrid euclidean distance for Catania is built using global horizontal irradiance, direct normal irradiance, ambient temperature and total cloud cover using the following algorithm:

$$\frac{F_{GHI}}{F_{DNI}} = \frac{W_{GHI}}{W_{DNI}} = \frac{W_{GHI}}{1 - W_{GHI}} = \frac{65\%}{1 - 65\%}$$
$$\frac{F_{GHI}}{F_{TC}} = \frac{W_{GHI}}{W_{TC}} = \frac{W_{GHI}}{1 - W_{GHI}} = \frac{76\%}{1 - 76\%}$$
$$\frac{F_{GHI}}{F_{TA}} = \frac{W_{GHI}}{W_{TA}} = \frac{W_{GHI}}{1 - W_{GHI}} = \frac{80\%}{1 - 80\%}$$
$$F_{GHI} + F_{DNI} + F_{TC} + F_{TA} = 1$$

$$\mathrm{ED}_{HD}^{N} = F_{GHI} \times \mathrm{ED}_{GHI}^{N} + F_{DNI} \times \mathrm{ED}_{DNI}^{N} + F_{TC} \times \mathrm{ED}_{TC}^{N} + F_{TA} \times \mathrm{ED}_{TA}^{N} \quad (3.12)$$

where, F_{DNI} is the weight factor for direct normal irradiance, F_{TC} is the weight factor for total cloud cover, and F_{TA} is the weight factor for ambient temperature that is used to build the hybrid euclidean distance. W_{GHI}, W_{DNI}, W_{TC} and W_{TA} are the weight calculated through the above EVA_B analysis for Catania. W_{GHI} has three different values that relate to each different secondary variable. The final hybrid euclidean distance formula for Catania is:

$$\mathrm{ED}_{HD}^{N} = 48\% \times \mathrm{ED}_{GHI}^{N} + 25\% \times \mathrm{ED}_{DNI}^{N} + 15\% \times \mathrm{ED}_{TC}^{N} + 12\% \times \mathrm{ED}_{TA}^{N} \quad (3.13)$$

In summary, different hybrid euclidean distances are built for these three locations. Based on the need of the following forecasting engine, days with the smallest hybrid euclidean distance to the forecast days will be selected. Dates of those days will be fed into the forecasting engine.

3.5.2.4 The Algorithms for the Proposed Similar Day Method

In the previous section, the proposed similar day method is built specificity for three testing sites. In this section, the overall algorithm of the proposed method which could apply to any location is summarized.

- Step 1: Normalize the data including recorded power output and forecast output-related variables by their maximum values.

- Step 2: Calculate the euclidean distance of recorded power output using Equation (3.2) and calculate the euclidean distance of recorded power output using Equation (3.3) for day pairs within the training dataset.

- Step 3: Calculate the EVA_V for each forecast output-related variable using Equation (3.4) and sort them in order. Select the output-related variable that has the lowest EVA value and choose it as the main variable (e.g. mostly should be forecast GHI).

- Step 4: Calculate bivariate euclidean distance ED_B of the main variable and other output-related variables using Equation (3.5). Through adjusting the weight of the main variable in ED_B, examine which secondary variable gains a lower EVA_B value. Choose those variables that have a lower EVA_B compared to EVA_{GHI} and record the weights W_{GHI} and W_V.

- Step 5: Use secondary output-related variables that could help the main variable to gain a lower EVA_B value to build a hybrid euclidean distance formula according to the weights that are calculated in the above step.

- Step 6: Calculate the forecast euclidean distance between the forecast day and all days in the training days using the above hybrid euclidean distance formula.

- Step 7: Select the right number of similar days that have the lowest forecast ED_{HD}^{N} based on the need of the forecasting engine.

3.5.3 Forecasting Engine

Four forecasting models; a persistence model, an ARIMA model, an LS-SVM model and an RBFNN model, are developed for these three sites. The ARIMA model and RBFNN

model is developed by Matlab ARIMA and neural network toolbox [68] and LS-SVM model is developed by LS-SVMlab v1.8 toolbox [69]. Following is the description for them.

The persistence model is a very simple and computationally effective forecasting model. The forecasting value $P_F(t, d)$ for the forecasting day d at hour t is calculated using:

$$P_F(t, d) = \frac{1}{j} \sum_{i=d-j}^{d-1} P_R(t, i) \tag{3.14}$$

where, $P_R(t, i)$ is the recorded power output at day i at hour t, $t \in [1, 24]$ and j is the number of training days in the persistence model. In this study, PM_j is defined as a persistence model using j training days.

The non-linear autoregressive RBFNN model is initially trained by selected similar days to generate one-step-ahead forecasting. The forecast power is then used to do 24-hour-ahead recursive forecasting. When the forecast day has passed, the model will be retrained to predict the next forecast day. The structure of the RBFNN model is $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-24})$.

The non-linear autoregressive LS-SVM model is modelled in $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-24})$ with radial basis kernel function and a simplex parameter optimization method. The forecasting power is then used to do 24-hour-ahead recursive forecasting. When the forecast day has passed, the model will be retrained to predict the next forecast day.

This model structure is achieved through trial and error and repeated The structure of the ARIMA model is $\mathrm{ARIMA}(2, 1, 1)(1, 0, 1)_{24}$ and the ARIMA model is also trained by similar days to generate one step ahead forecasting. Through trial and error and repeating the model building process, an $\mathrm{ARIMA}(2, 1, 1)(1, 0, 1)_{24}$ structure is identified. The forecasting power is then used to do recursive 24-hour-ahead forecasting. When the forecast day has passed, the model will be retrained to predict the next forecast day.

3.6 Results and Discussion

3.6.1 Error Measurement

Two error measurements are used in this study to measure daily forecasting accuracy: normalized mean absolute error (nMAE) and normalized root-mean-square error (nRMSE). They are defined as:

$$\text{nMAE} = \frac{100}{N} \sum_{i=1}^{N} \frac{\left|P_f^i - P_a^i\right|}{P_C} \tag{3.15}$$

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (P_f^i - P_a^i)^2}}{P_C} \tag{3.16}$$

where, N is the day length, P_f^i is the forecast power output at hour i, P_a^i is the actual power output at hour i and P_C is the capacity of the PV site. For a total N_T testing days, the average forecasting error and the standard deviation of the forecasting error are both calculated.

3.6.2 Forecasting Accuracy

Results were generated for the period from April 1 to June 30. The data-preprocessing engine utilized the proposed hybrid similar day method and the forecasting engine utilized four forecasting models including a persistence model, an ARIMA model, an LS-SVM model and an RBFNN model.

The average and standard deviation of forecasting error for San Diego in terms of nMAE and nRMSE are presented in Figure 3.25. When using the proposed data-processing engine, the persistence model and ARIMA model generated a more accurate forecast result compared to the LS-SVM model and the RBFNN Model. The average forecasting error and the standard deviation of the forecasting error of the persistence model and ARIMA are both lower than the LS-SVM and the RBFNN for both first order error measurement (nMAE)

and second order error measurement (nRMSE).

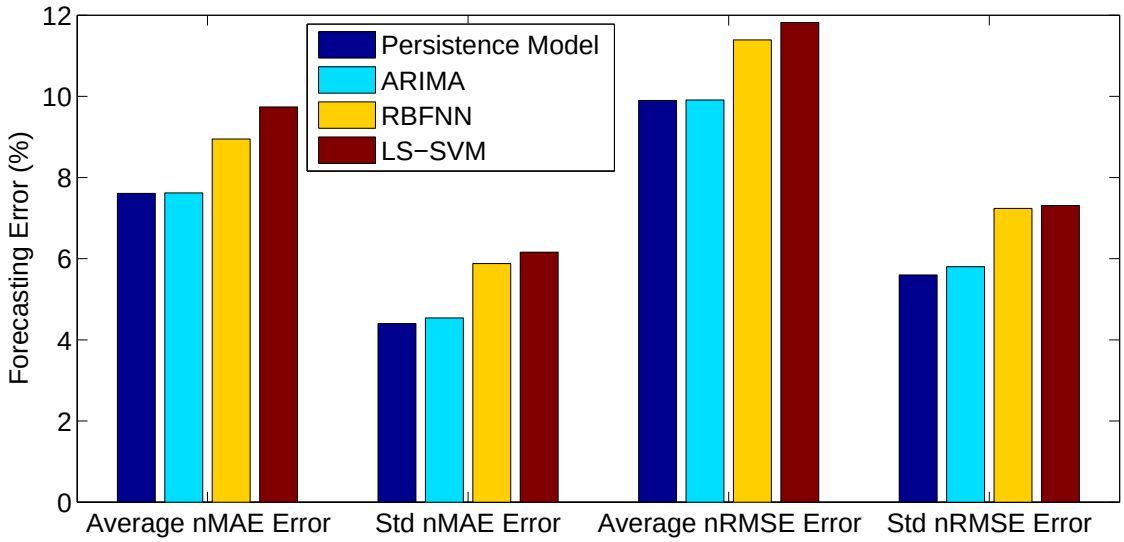

Figure 3.25: Average and standard deviation (Std) of forecasting error for San Diego in term of nMAE and nRMSE

Figure 3.26 shows the forecasting errors for Braedstrup. Similar to San Diego, the forecasting error at Braedstrup and Catania shows that when combining the similar day data-preprocessing engine, the persistence model and ARIMA model perform better than the LS-SVM and the RBFNN model. Figure 3.27 shows the forecasting error at Catania. It also indicates an accurate forecasting result from the persistence model and the ARIMA model. In terms of nMAE, the average error of the ARIMA and the persistence model are both less than 8%.

In summary, when array level forecasting uses a similar day method based data-preprocessing engine and ARIMA or a persistence model-based forecasting engine, better accuracy is achieved compared to RBFNN and LS-SVM.

3.6.3 Comparison with Different Similar Day Methods

In the literature, two different similar day methods were utilized [32, 33]. In the following, this study examines the forecasting accuracy improvement gained from a Data-Preprocessing Engine built based on these two similar day methods. The improvements are compared with

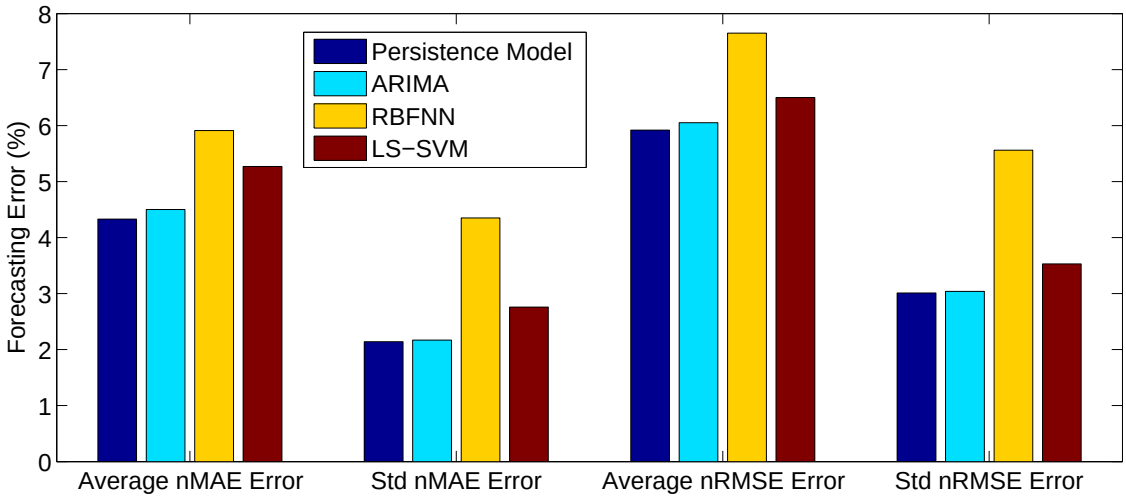

Figure 3.26: Average and standard deviation (Std) of forecasting error for Braedstrup in terms of nMAE and nRMSE

the one proposed in this thesis.

Similar Day Method 1 [32] searched similar days based on the euclidean distance of temperature. The forecast $\mathrm{ED}_1(i,j)$ is defined as:

$$\mathrm{ED}_1(i,j) = \left[\sum_{k=1}^{3}\left(T_i^k - T_j^k\right)^2\right]^{\frac{1}{2}} \tag{3.17}$$

where, day j is the forecast day and day $i \in [1,j]$ are training days, T_i^1, T_i^2, T_i^3 are the forecast high, low and average ambient temperatures for day i, T_j^1, T_j^2, T_j^3 is the forecast high, low and average ambient temperature for day j. Based on the value of ED_1, the day that has the shortest euclidean distance between forecast day and historical days will be selected.

Similar Day Method 2 [32] searched for similar days based on similarity of season, solar radiation, maximum temperature and minimum temperature. The similarity between day i and day j is defined as:

$$\begin{aligned}\mathrm{ED}_2(i,j) = &\left(1 - \left|\frac{D_i - D_j}{365}\right|\right) \times \left(1 - \left|\frac{1}{WA_i - WA_j}\right|\right) \times \\ &\left(1 - \left|\frac{1}{TA_{Max}^i - TA_{Max}^j}\right|\right) \times \left(1 - \left|\frac{1}{TA_{Min}^i - TA_{Min}^j}\right|\right)\end{aligned} \tag{3.18}$$

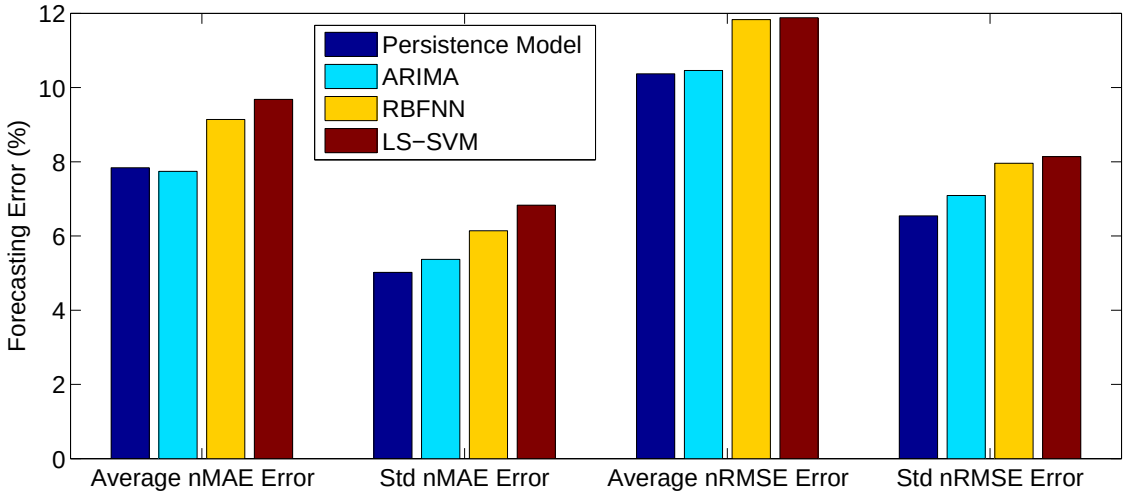

Figure 3.27: Average and standard deviation (Std) of forecasting error for Catania in term of nMAE and nRMSE

where, day j is the forecast day and day $i \in [1, j]$ is a training day, D_i is the day index for day i and D_j is the day index for day j (eg D_{32} is February 1. WA_i is the weather index for day i and WA_j is the weather index for day j, TA^i_{Max} is the highest temperature at day i, TA^j_{Max} is the highest temperature at day j, TA^i_{Min} is the lowest temperature at day i and TA^j_{Min} is the lowest temperature at day j. The similarity of season was measured by day of difference and similarity of solar radiation is measured by weather index difference. This method will find five historical days that have the smallest ED_2 between the forecast day as the training set.

In order to examine the improvement using different similar day methods, this study defines the average forecast error and standard deviation of forecasting error without similar day method as E^{NoSD}_{Mean} and E^{NoSD}_{Std}; the average forecast error and standard deviation of forecasting error with similar day method as E^{SD}_{Mean} and E^{SD}_{Std}. The accuracy improvement based on average forecasting error is defined as:

$$\mathrm{IMP}_{Mean} = \frac{E^{NoSD}_{Mean} - E^{SD}_{Mean}}{E^{NoSD}_{Mean}} \tag{3.19}$$

The improvement based on standard deviation of forecasting error is defined as:

$$\mathrm{IMP}_{Std} = \frac{E_{Std}^{NoSD} - E_{Std}^{SD}}{E_{Std}^{NoSD}} \tag{3.20}$$

In the following, three data-processing engines were built using these three similar day methods, combined with four different forecasting engines. Those different forecasting tools were built and compared with each other to examine which data-processing engine is more effective.

The first test was conducted using a persistence model. Without data-processing, the forecasting engine chose the PM_1 model. Since Similar day method 1 chooses only one most similar day from the training days, the forecasting engine chose the PM_1 model as well. PM_5 was chosen as the forecast Engine when the Data-Preprocessing Engine is built by Similar day method 2 which selects five most similar days. Since the proposed similar day method can choose any number of similar days, PM_{12} is chosen for San Diego; PM_{18} is chosen for Catania and PM_4 is chosen for Braedstrup.

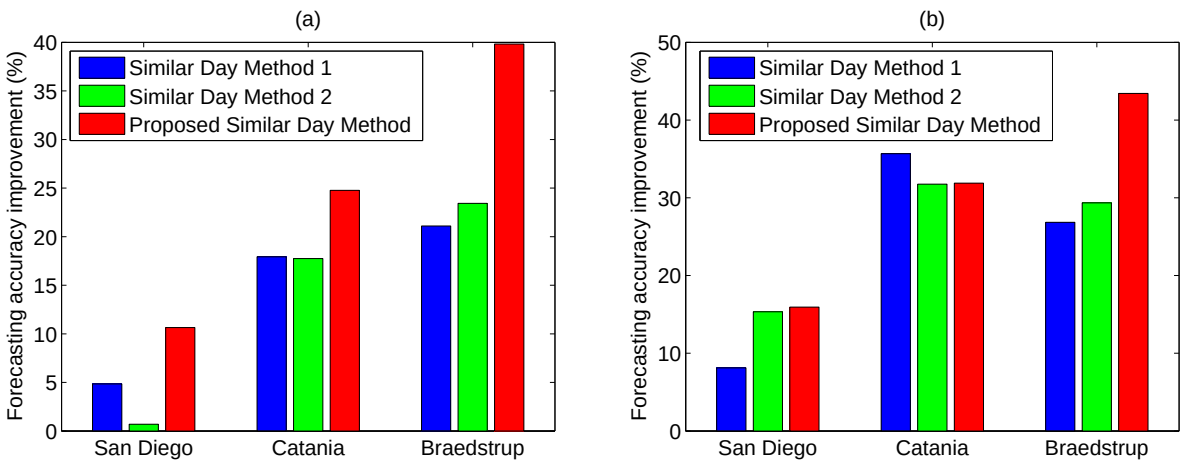

Figure 3.28: Forecasting accuracy improvement from three similar day methods for persistence model: (a) Improvement based on IMP_{Mean}, (b) Improvement based on IMP_{Std}

Figure 3.28 (a) and (b) shows the plot of IMP_{Mean} and IMP_{Std} for San Diego, Catania and Braedstrup using a persistence model method. In terms of both IMP_{Mean} and IMP_{Std}, these

three methods can all help to improve the forecasting accuracy for all the testing locations. What's more the proposed similar day method gives a higher improvement compared to Similar Day Methods 1 and 2. For all the three testing locations, the proposed similar day method can help to gain a higher IMP_{Mean} improvement. In terms of IMP_{Std}, only the improvement for Catania is slightly less than the other two locations. One reason for this is that the proposed similar day method has a better similar day selection formula, whereby the forecast similarity is close to actual similarity. The other reason is that the proposed similar day method is designed to select any number of training days for different sites. But Similar Day Method 1 is designed to select one most similar day and Similar Day Method 2 is designed to select 5 most similar days.

The second test is conducted using the ARIMA model. Since it is impossible to build ARIMA using only one training day, method 1 is not included in this study. ARIMA trained with 25 previous days was chosen as the benchmark model. ARIMA trained with 5 similar days selected by similar day method 2 and ARIMA trained with 25 similar days selected by HD SD method is used for comparison.

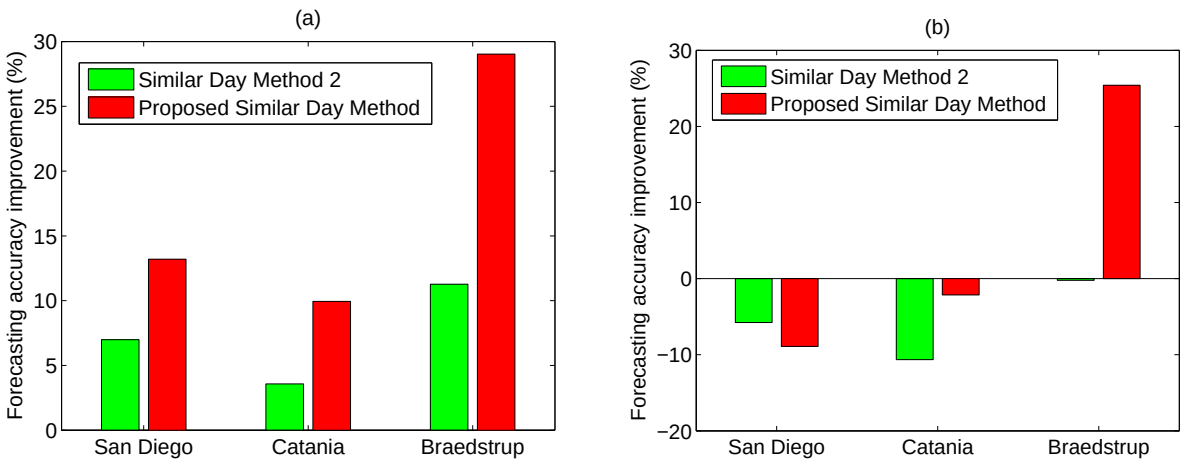

Figure 3.29: Forecasting accuracy improvement from two similar day methods for ARIMA model: (a) Improvement based on IMP_{Mean}, (b) Improvement based on IMP_{Std}

Figure 3.29 (a) and (b) is the plot of IMP_{Mean} and IMP_{Std} for San Diego, Catania

and Braedstrup using the ARIMA model. Similar to the result when using the persistence model, the proposed similar day method also can help to gain a higher improvement based on IMP_{Mean}. For Braedstrup, the improvement reached 30%. In terms of IMP_{Std}, improvement at Braestrup is still significant, but IMP_{Std} at San Diego and Catania decreases. IMP_{Mean} for Catania is 9.94% and IMP_{Std} is only -2.14%, the proposed similar day method is still effective when considering these two improvements together.

The third test is conducted using LS-SVM. The benchmark LS-SVM model is trained using 10 previous days. Again five training days were selected for Similar Day Method 2. The proposed similar day method selected 10 similar days. For simulation for Similar Day Method 1, the training set is 10 days without similar day method, but the input data is chosen by Similar Day Method 1.

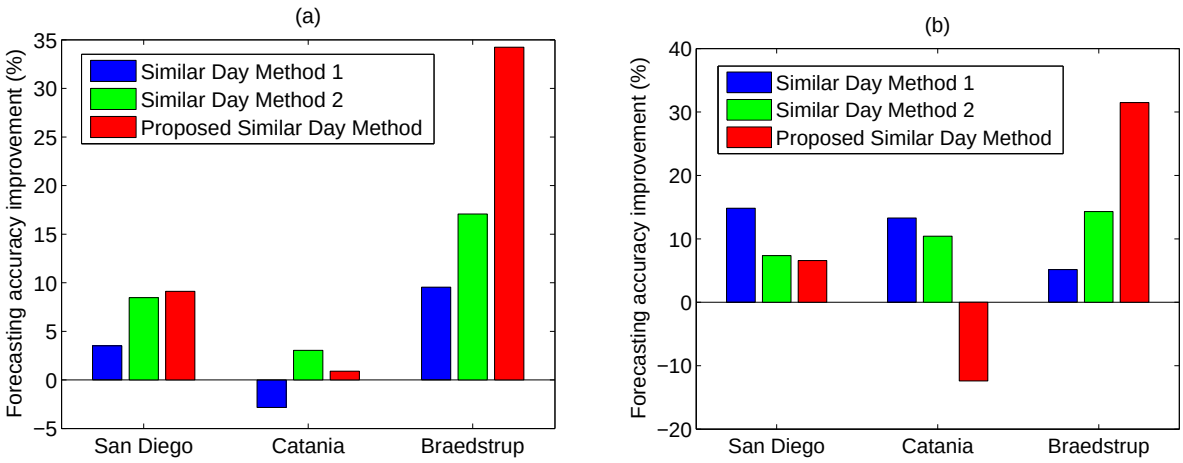

Figure 3.30: Forecasting accuracy improvement from three similar day methods for LS-SVM model: (a) Improvement based on IMP_{Mean}, (b) Improvement based on IMP_{Std}

Figure 3.30 (a) and (b) is the plot of IMP_{Mean} and IMP_{Std} for San Diego, Catania and Braedstrup using the LS-SVM model. For San Diego and Braedstrup, method 3 performs better than the other two methods, but method 3 does not work well as well at Catania. Adding it will actually harm the forecasting accuracy.

The last test is conducted using the RBFNN model. The benchmark RBFNN model is trained by 50 days. Five training days were selected for Similar Day Method 2. The proposed

similar day method selects 50 similar days. For Similar Day Method 1, the training set is 50 days without similar day method, but the input data is chosen by Similar Day Method 1. Figure 3.30 (a) and (b) is the plot of IMP_{Mean} and IMP_{Std} for San Diego, Catania and Braedstrup using the RBFNN model. At San Diego, only the proposed similar day method can help to gain accuracy improvement. At Braedstrup, the proposed similar day method can help to gain substantial IMP_{Mean} improvement, but loses some IMP_{Std} improvement. The IMP_{Mean} from the other two methods is not significant. At Catania, no similar day method works; adding a similar day method will decrease the forecasting accuracy.

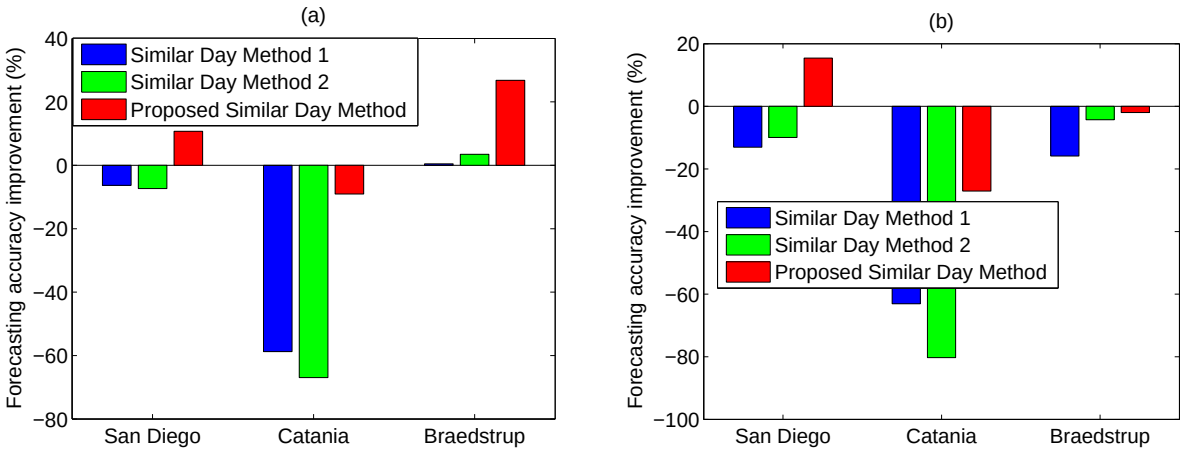

Figure 3.31: Forecasting accuracy improvement from three similar day methods for RBFNN model: (a) Improvement based on IMP_{Mean}, (b) Improvement based on IMP_{Std}

In summary, the proposed similar day method performed better than the other two methods. With the help of the proposed similar day methods, the persistence model can gain a very significant improvement. This is consistent with a general rule in forecasting that model sophistication does not always mean a better result [70]. For all three testing sites, the proposed similar day method all have a better performance than other methods when combined with different forecasting model. In fact, all methods fail to show improvement for Catania using the RBFNN model. The RBFNN model needs more training days than the other models. Because of the forecasting error of output-related variables, it cannot select too many similar days and give them a right order. In general, similar day method is especially

useful when combined with a computational effective models which need less training set. Moreover, accurate output-related variable forecasting is crucial for a similar day method. For example, the EVA of GHI is 16.35% for Catania, but 11.41% for Braedstrup. Thus, the similar day method works better at Braedstrup.

3.6.4 Comparison with ARX

From the above simulation, the best array level forecasting tool is built by the proposed similar day method based data-preprocessing engine and a persistence model-based forecasting engine. This subsection compares the accuracy of this forecasting tool with an Auto Regressive with Exogenous Input (ARX) model [23] using the same data, i.e., Data from Site 2: Braedstrup, and the same error measurement. The error measurement used in this comparison is nRMSE which is normalized by mean peak power of the 21 PV systems over 2006 (2769 kW). The ARX model runs at 12:00 and can predict up to 36 hours, while the proposed array level forecasting tool runs at 24:00 and predicts up to 24 hours. In the following, a six hour period from 13:00 to 18:00 is selected as the testing period. For this period, ARX did a six-hour-ahead forecast, while the proposed array level forecasting tool did a 18-hour-ahead forecast. The comparisons are shown in Table 3.7. The general decrease in forecasting error with time is due entirely to the decrease in power output as sunset is approached and is not in itself an indication of increasing accuracy. Only comparative performance of the two models can be assessed from this table. At 13:00 and 14:00, the ARX has better accuracy than the proposed model, but at 15:00 to 18:00, the proposed method has better accuracy. Overall, the average accuracy of the proposed method is 7% lower than the ARX model in this period. Considering that the forecasting horizon (up to 18 hours) of the proposed model is longer than the ARX model (up to six hours), the proposed forecasting tool is effective.

Table 3.7: Forecasting error comparison between ARX and proposed forecasting model for the same time period with different forecasting horizon

Time	13:00	14:00	15:00	16:00	17:00	18:00
Forecasting Horizon of ARX	1 h	2 h	3 h	4 h	5 h	6 h
Average Forecasting Error (nRMSE)	9.0%	9.5%	9.0%	6.75%	5.4%	4.5%
Forecasting Horizon of Proposed Model	13 h	14 h	15 h	16 h	17 h	18 h
Average Forecasting Error(nRMSE)	11.1%	10.1%	8.6%	4.6%	4.3%	2.0%

3.7 Summary

In summary, this chapter proposed an array level forecasting tool. This tool combined a data-preprocessing engine and a forecasting engine. The foundation of the data-preprocessing engine is a similar day method which searches similar days based on the hybrid combination of the output-related variables. When this data-preprocessing engine was combined with four widely used forecasting models, a simple persistence model and the ARIMA model generated more accurate forecasts compared to the LS-SVM and the RBFNN models. Through the analysis of the power output, this study found the proposed similar day method works better than similar day methods in the literature and the overall forecasting tool generated a more accurate forecast compared to the ARX model in the literature.

Chapter 4

System Level Short-Term PV Output Forecasting

4.1 Introduction

In this chapter, the short-term forecasting tool designed for aggregated system level PV output is introduced. This 24-hour-ahead PV output forecasting can help system operators to maintain the reliability of the system, especially when this system involves or plans to involve large-scale PV power into the grid. This study is conducted using the aggregate PV power output in California, which plans to integrate a significant amount of renewable energy into the grid. In Section 4.2, a brief introduction of solar power in California, and an analysis of the hourly PV output is conducted. In Section 4.3, the system level forecasting tool and four forecasting models used in this tool are described. In Section 4.4, the forecasting results are discussed. In Section 4.5, the conclusions for this system level forecasting tool are summarized.

4.2 Data Description and Analysis

The studied system is the California power grid. Because of the geographical diversity of solar arrays across a wide area, the aggregated solar power output is less prone to local weather changes. Hence, the characteristics of the system level output are expected to be different from those of a single array level PV system.

4.2.1 Solar power in California

The California power grid covers three quarters of California and part of Nevada, and this grid is operated by California ISO, an independent system operator in North America. One

goal of California ISO is integrating more renewable energy into the grid. By 2010, 17% of the load was served by renewable resources and 33% of the load is expected to be supplied by renewable energy by 2020 [71].

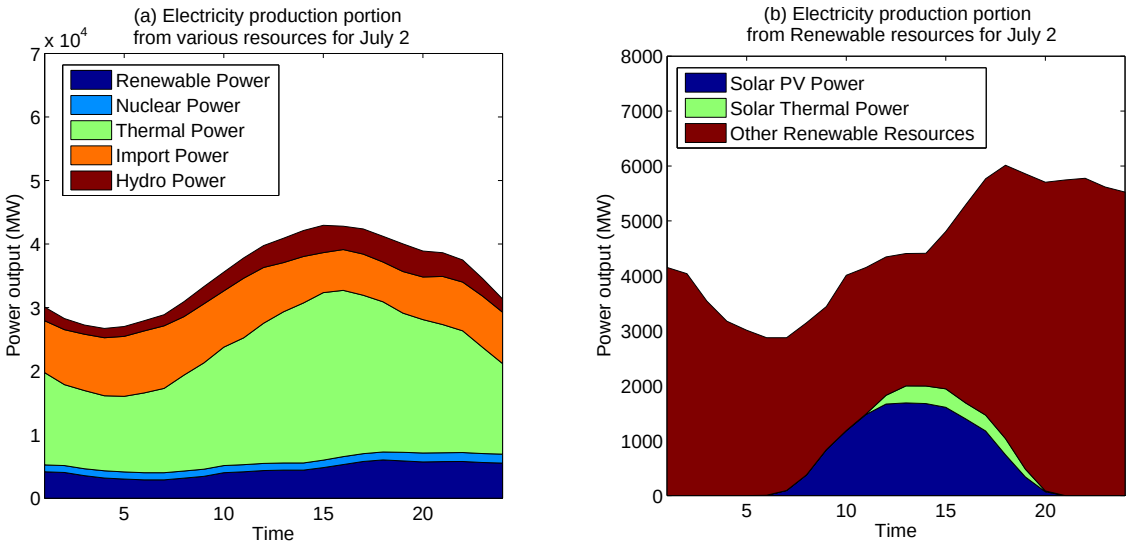

Figure 4.1: California power supply shares from various resources on July 2 [68]

The total net capacity of the California grid is 60,703 MW and 18.3% of the load is currently supplied by renewable energy. Solar energy counts 15.6% of the total renewable energy capacity [71]. To illustrate the daily power supply within California, the power output distribution from various resources on July 2 (randomly selected) is analysed. Figure 4.1 (a) shows the power supply share from various resources on July 2, 2013. Most of the power is generated by thermal power, but the renewable power also has its significant share as shown in the dark blue area. On July 2, the 24-hour system demand was 851,623 MWh on this day, and 106,858 MWh, i.e., 12.55% of the load is powered by renewable power. Figure 4.1 (b) shows the share of different renewable resources within the grid on July 2. Solar power has a clear portion in the renewable resources. The installed PV capacity has surpassed the solar thermal by 2008 [72]. Currently, solar PV has a significantly larger portion compared to solar thermal power. On July 2, 2013, solar PV produced 13,312 MWh and solar thermal has a total production of 2,166 MWh. The newly added PV capacity in 2012 was over 670

MW and California plans to install 12,000 MW capacity by 2020 [73].

In summary, PV will produce more electricity within the grid in the future and so will impose more influence on the grid. Thus, its output characteristics need to be analyzed.

4.2.2 Analysis of Aggregated PV Power Output

The impact of PV power on the grid will increase with the development of PV capacity. Because of the policy support and good insolation resource, solar power in California leads the USA California in this area, having the largest number of installed rooftop PV systems in the USA. In this section, the characteristics of aggregated PV power output are analysed. The system level data is collected from California ISO, who publish PV power production within the ISO grid with one-hour intervals [71]. Starting from December 1, 2012, the PV power and thermal solar power were published separately.

4.2.2.1 Analysis of Daily Energy Production

The daily energy production feature of the system level output is analysed here. Unlike a array level PV power system, which has constant capacity, the capacity of system level PV power is not a constant value. Thus, the daily energy production of system level and array level outputs may have a different feature.

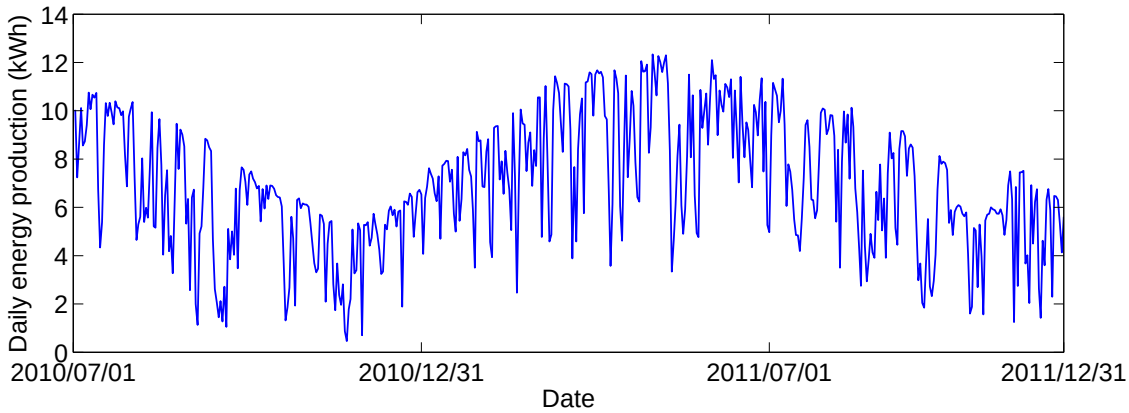

Figure 4.2: The output of a array level PV system in California (kW)

Figure 4.2 shows the daily energy production from one array level PV system in Cali-

fornia (San Diego in Chapter 3) from July 1, 2010 to December 31, 2011. Although, there are fluctuations among this production series, the trend of this array level daily energy production shows a clear annual pattern. For example, on the same data over different years (e.g July 1, 2010 and July 1, 2011), the daily production will not exceed 12 kWh due to irradiance strength and installation capacity limitation.

Figure 4.3 shows the daily energy production from all grid-connected solar energy (Solar PV and solar thermal) from April 20, 2010 to July 8, 2013 within California. Because of the continuous newly added solar energy, the output is increasing, which is especially obvious for 2013, due to the fast development of solar systems. There were 438 MW new PV capacity added in California during the first-quarter of 2013 and 409 MW new PV capacity during the second-quarter of 2013 [74]. On April 20, 2010, the daily production was less than 200 MWh, however, it was more than 500 MWh on April 20, 2013. As stated in Section 4.2.1, PV capacity surpassed solar thermal by 2008. Thus, although Figure 4.3 shows production of solar energy, PV energy production is expected to dominate in the future.

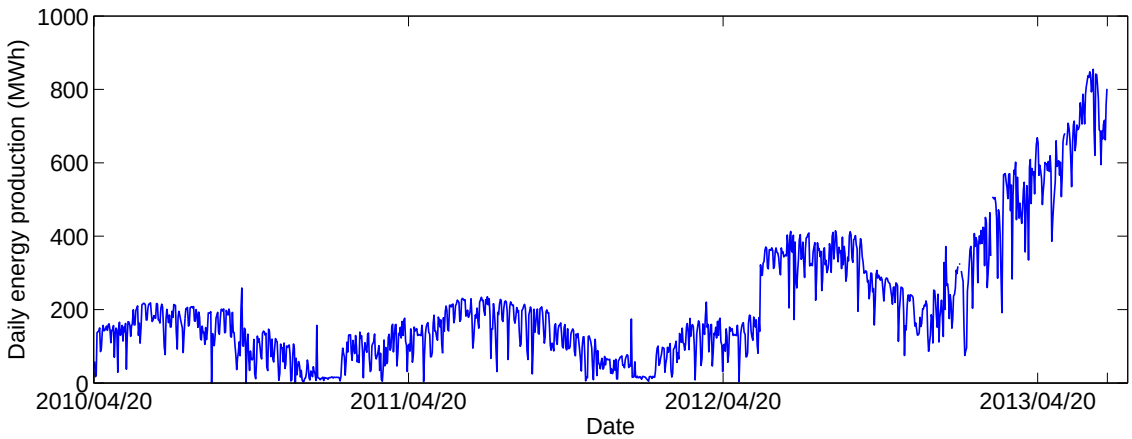

Figure 4.3: On-grid solar energy production within California (Both solar PV and solar thermal energy) (MW)

4.2.2.2 Analysis of Daily Energy Production Period

Solar energy production period for a single site has a close relationship with the day length. The irradiance can even be approximately calculated by the day length [17]. In the following,

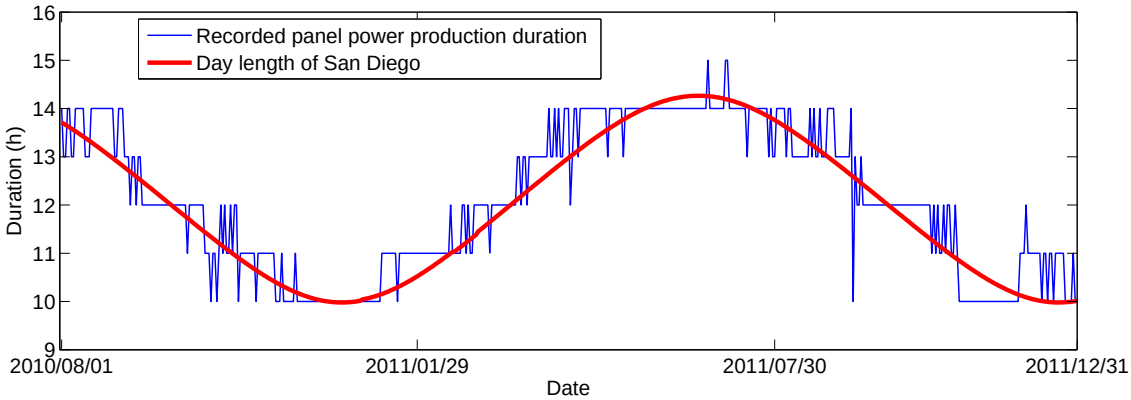

Figure 4.4: Comparison of power production duration from a PV system in San Diego and the day length of San Diego

this relationship is examined for system level power data and is compared to array level power data. Figure 4.4 compares the power production duration from one PV system in San Diego and the day length of San Diego from August 1, 2010 to December 31, 2011. The array level power production duration is correlated strongly with the day length. The difference between production duration and day length is less than two hours. Since the production duration is recorded as an integer, this difference is acceptable.

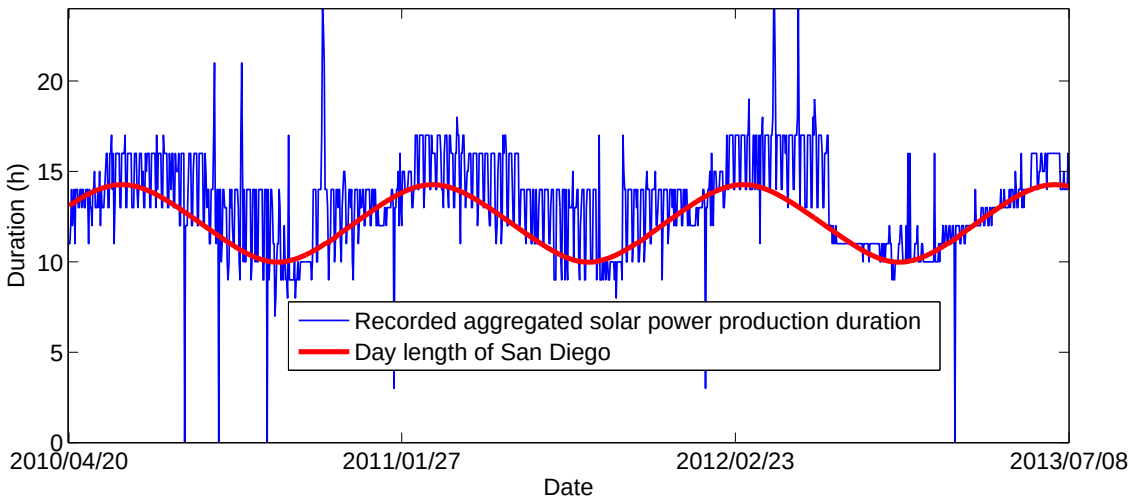

Figure 4.5: Comparison of aggregated solar power production duration (Solar PV and solar thermal) in California and the day length of San Diego

The power production period of aggregated system level solar does not have the close

relationship with day length as the array level power output. Figure 4.5 plots the relationship between the aggregated solar power production duration in California and the day length of San Diego. San Diego was chosen as the representation site in California to make the day length comparison because it is in the south of the states where the solar power is likely to be located. Latitude and longitude effect on day length are ignored. This power production duration has a larger fluctuation compared to the duration of an array PV system. Some days even have 24-hour energy production. Day-to-day difference in daily production changes are frequently over one hour. Solar thermal power is the main reason behind this significant fluctuation. Since solar thermal plants can save the heat of solar energy and use it when they want to produce electricity, there may be no power production for some days and may produce power after sundown. Thus, the power production from solar thermal plants may not follows the weather or irradiance strength. For example, the power output on December 31, 2012 and January 2, 2013 have a production period of 16 hours which is much longer than the day length of California (10 hours for December).

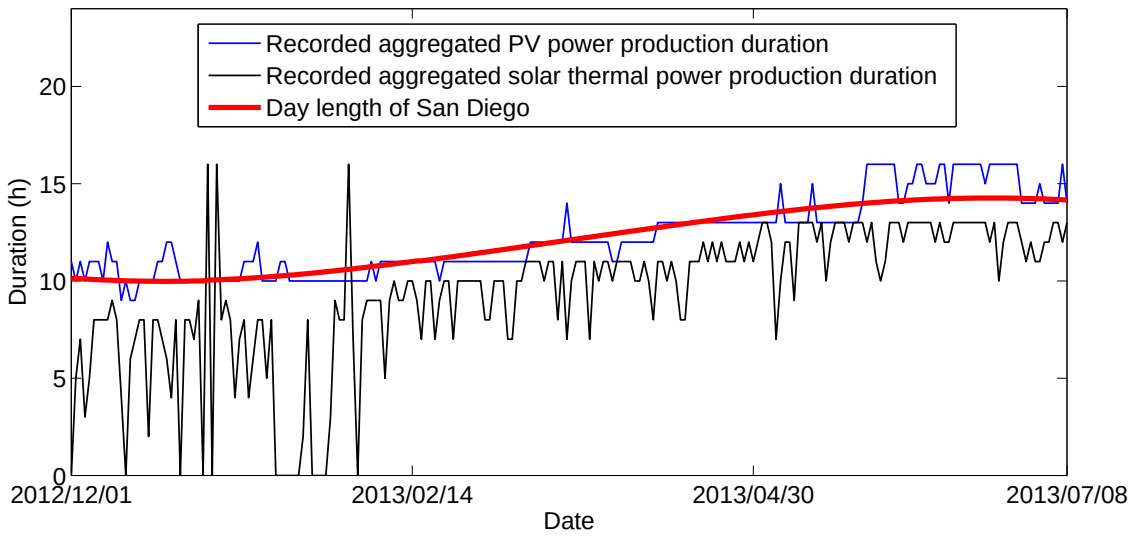

Figure 4.6: Comparison of aggregated PV power production duration and aggregated solar thermal power production duration in California

Figure 4.6 shows the comparison of aggregated PV power production duration and aggregated solar thermal power production duration from December 1, 2012 to July 8, 2013.

Clearly, although the day length is recorded for San Diego, the aggregated PV power output production period has a very similar trend to the day length of San Diego. In contrast, the aggregated solar thermal power output production period has quite a large deviation from the day length. This figure shows that the fluctuation of the overall solar power data is mainly from the solar thermal power output. Because of the large power production period variation of solar thermal power, it should be excluded for the analysis.

4.2.2.3 Analysis of Intra-day Fluctuation

When removing the solar thermal power from the solar data, the remaining PV power data has less power production duration deviation. Moreover, because of the geophysical smoothness, the output of aggregated system has less fluctuation than the individual sites examined in the previous chapter. In the following, the intra-day fluctuation of this system level PV output series was analysed. Similar to the analysis of the array level fluctuation, the differenced series P_t^D for the aggregated system level PV output is plotted in Figure 4.7 (a). Visually, the fluctuation is much lower than the fluctuation level of array level output, which is plotted in Figure 4.7 (b) (Figure 4.7 (b) is the same as Figure 3.8 (a) in Chapter 3). In terms of maximum fluctuation, aggregated power is 0.4581 which is much less than an array level system in California (0.8217). In term of the standard deviation of the intra-day fluctuation, array level power is 0.1237 and system level power is 0.0526.

The historical power data before December 1, 2012 is the sum of solar PV and solar thermal. This part of the data is not suitable to be used as a training set, even if the majority is solar PV power. Figure 4.8 compares the maximum and standard deviation of intra-day fluctuation levels for solar PV series, solar thermal series and their summation for the period from December 1, 2012 to July 8, 2013. The solar thermal series has a significantly higher intra-day fluctuation level than the solar PV series. Although solar PV is the major part of the solar power, mixing solar thermal and solar PV together will increase the intra-day fluctuation in terms of standard deviation. Thus, from the viewpoint of intra-day fluctuation,

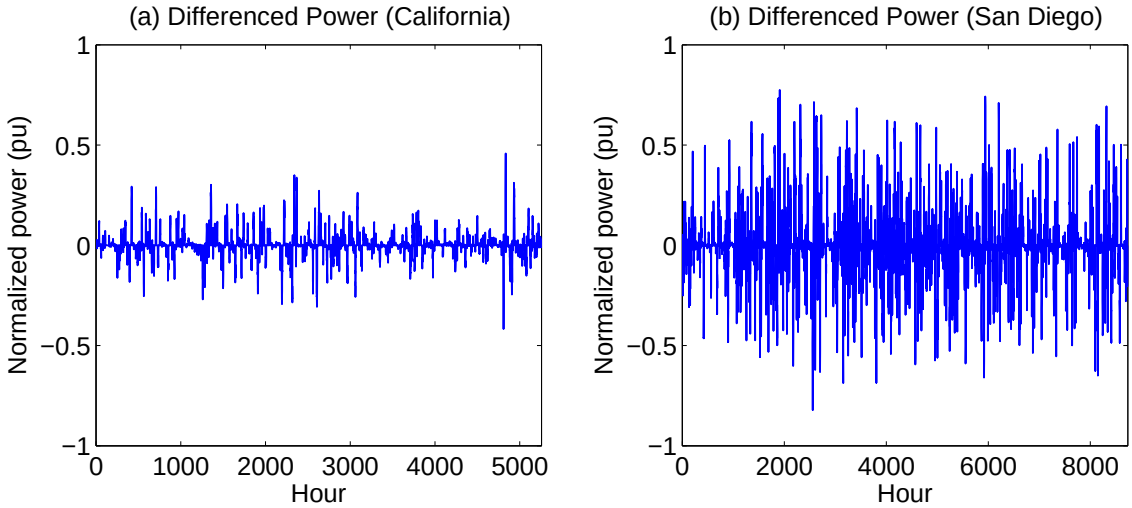

Figure 4.7: Hourly intra-day fluctuation series over 220 days

the training set should be the PV power data alone.

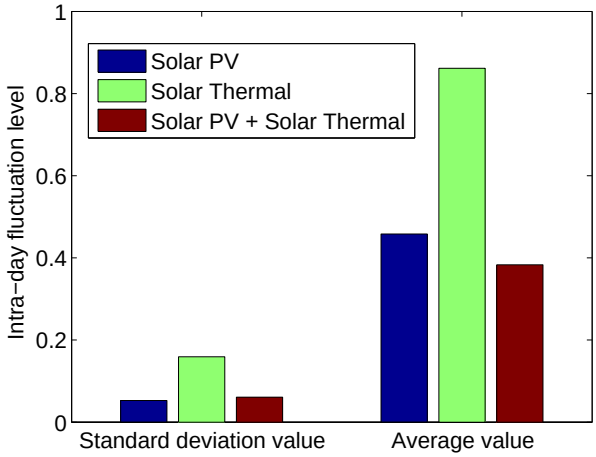

Figure 4.8: Intra-day fluctuation of solar PV, solar thermal and total solar power

4.2.2.4 Summary

In summary, the aggregated PV output has an increasing trend and lower intra-day function level. The increasing trend in solar production may influence the importance of historical data that far from the forecast day. Moreover, the solar thermal and solar PV output are published together as solar energy before September 1, 2012. Therefore, data recorded before this date may not be suitable to be used as a training set because of the high intra-

day fluctuation of solar thermal data. Thus, it is suitable to use PV output data alone as a training set. In the following, the modeling process is done by only aggregated power output from September 1, 2012 to July 8, 2013.

4.3 Forecasting Tool for System Level PV Output Forecasting

In this section, the system level forecasting tool is introduced. The proposed tool does not have a similar day-based data-preprocessing engine, which will be a suggested future research direction. The forecasting engine is built by four models including persistence, RBFNN, ARIMA and LS-SVM, which are frequently used for PV output forecasting [23,30, 31]. Available training data for all models is the recorded hourly PV output from September 1, 2012 to June 17, 2013 and the testing period is from June 18, 2013 to July 8, 2013. In the following, these four models are developed to gain their best performances.

4.3.1 ARIMA based Forecasting

The available historical PV output series data for a specific ARIMA model can be denoted as:

$$P_t; t = 1, \cdots T. \tag{4.1}$$

This series includes historical power data up to hour 24 of day $d-1$ (day d is the forecast day). The value of T denotes the size of the training set. For example, $T = 1344$ means the past two months data. When an ARIMA model is trained by this training series, it will generate 24 future values $(T+1, \cdots T+24)$ for day d. When the forecast day has passed, the model will be retrained to predict the next forecast day. In the following, the size of the training set, as well as the model structure, are investigated to gain the best performance.

The initial structure of the ARIMA model is found through analyzing the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plot. Figure 4.9 is the ACF and PACF plot for overall power series P_t. A clear 24-hour period of ACF plot

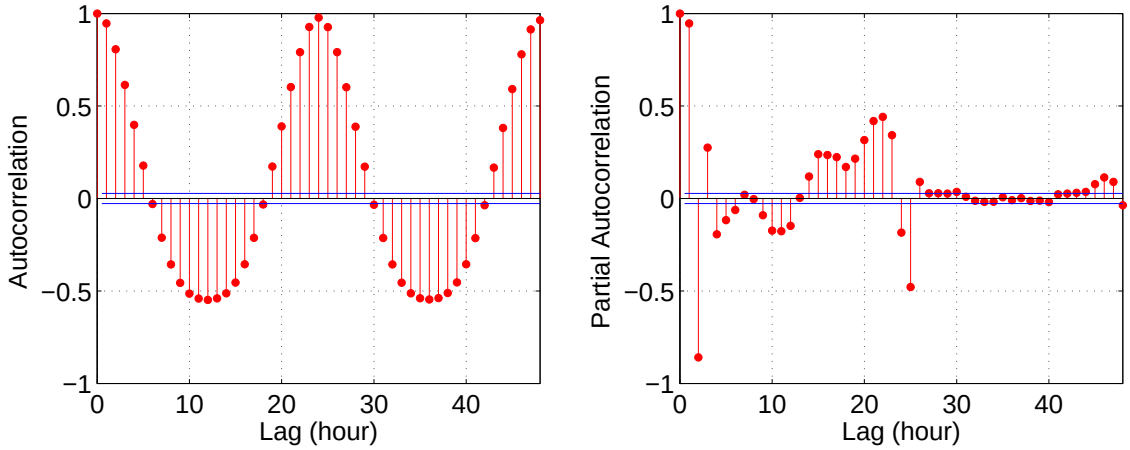

Figure 4.9: ACF and PACF plot of the original aggregated system level PV power output

indicates P_t the diurnal variation. As well, the PACF plot does not decay. These two plots indicate that P_t is non-stationary. To meet the stationary requirement of the ARIMA model, the original power data needs differencing to remove this non-stationary. Thus, a 24 order differencing is needed to remove the seasonality. Recalling the increasing trend of this series, another first order differencing is needed. Figure 4.10 shows the ACF and PACF plot for the differenced series. The auto correlation decreases sharply after one hour and partial correlation decreases sharply after two hours. Moreover both auto correlation and partial correlation is large and negative at 24 hours. Hence, ARIMA$(1,1,1)(1,0,1)_{24}$ could be used as the tentative model.

The initial model identified by ACF and PACF plot is the tentative model. The final model of ARIMA must be determined by trial and errors. In practice, all the orders of (p,d,q,P,D,Q) of the ARIMA model are smaller than 2 [75]. Among this range, these tentative models are used to generate forecasts for the period from June 18, 2013 to July 8, 2013. In this test, the size of the training set is 2400 (100 days) for all tentative models. Table 4.1 lists the forecasting error from different combinations. Through testing, the best structure is ARIMA$(0,1,1)(1,0,1)_{24}$, the average error is 5.42%, and the standard deviation of the forecasting error is 3.85%.

The size of the training set also has influence on the forecasting accuracy of the ARIMA

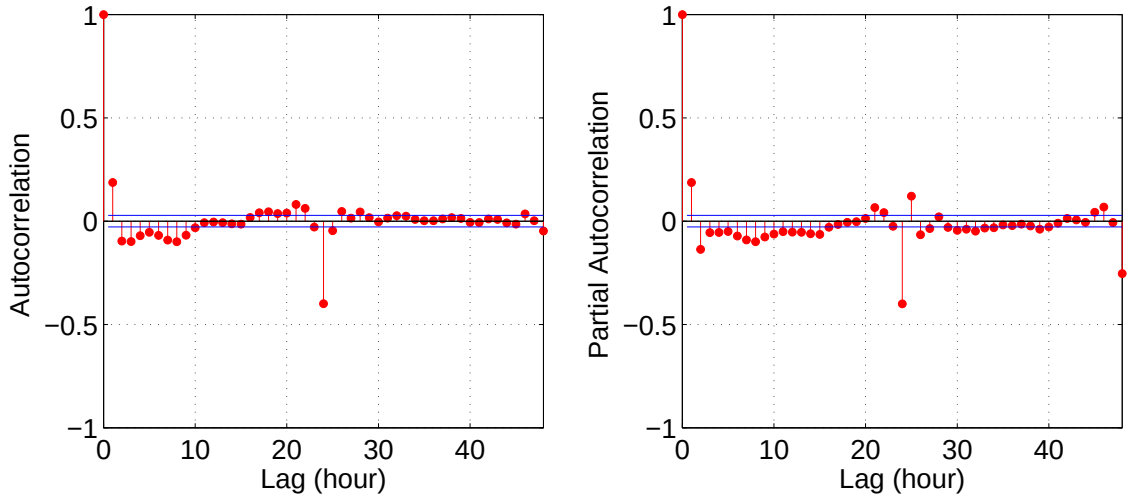

Figure 4.10: ACF and PACF plot of the differenced aggregated system level PV power output

model. This analysis is conducted using the initial ARIMA model (ARIMA$(1,1,1)(1,0,1)_{24}$). Figure 4.11 (a) and (b) shows the average forecasting error and standard deviation of forecasting error of the ARIMA$(1,1,1)(1,0,1)_{24}$ model when changing the size of the training set from 10 days to 200 days. Both the average and standard deviation of forecasting error became stable when ARIMA is trained by more than 30 days and the best accuracy is achieved when the size of the training set is 100 days.

Using the final configuration, the final model structure is ARIMA$(0,1,1)(1,0,1)_{24}$ and the training data is 100 previous days. To illustrate the forecasting process of the ARIMA model, one day in the testing period (July 2) is used as an example. The ARIMA$(0,1,1)(1,0,1)_{24}$ can be specified as:

$$(1-\Phi_{24}B^{24})(1-B)(1-B^{24})P_t = c + (1+\theta_1 * B)(1+\Theta_{24}B^{24})\epsilon_t \tag{4.2}$$

Through training, the parameters in this equation are calculated and the final model for July 2 is:

$$(1-0.0560611B^{24})(1-B)(1-B^{24})P_t = 0 + (1+0.162325B)(1-0.828111B^{24})\epsilon_t \tag{4.3}$$

Figure 4.12 is the residual ACF plots for this ARIMA model. The horizontal blue line in this figure is the significance limits of ACFs. Clearly, there is no significant correlation for

Table 4.1: Influence of model structure on the forecasting accuracy of ARIMA measured by nRMSE

Structure	Size	Average	Std	Max	Min
$\text{ARIMA}(0,1,1)(1,0,1)_{24}$	**100**	**5.42%**	**3.85%**	**15.56%**	**1.57%**
$\text{ARIMA}(0,1,2)(1,0,1)_{24}$	100	5.43%	3.85%	15.57%	1.60%
$\text{ARIMA}(1,1,0)(1,0,1)_{24}$	100	5.43%	3.85%	15.56%	1.57%
$\text{ARIMA}(1,1,1)(1,0,1)_{24}$	100	5.43%	3.84%	15.56%	1.58%
$\text{ARIMA}(1,1,2)(1,0,1)_{24}$	100	5.58%	4.12%	16.58%	1.53%
$\text{ARIMA}(2,1,0)(1,0,1)_{24}$	100	5.43%	3.85%	15.57%	1.60%
$\text{ARIMA}(2,1,1)(1,0,1)_{24}$	100	5.55%	4.12%	16.62%	2.04%
$\text{ARIMA}(2,1,2)(1,0,1)_{24}$	100	6.23%	4.90%	18.89%	1.14%
$\text{ARIMA}(2,1,2)(1,0,0)_{24}$	100	6.36%	4.18%	14.97%	0.95%
$\text{ARIMA}(2,1,2)(0,0,1)_{24}$	100	5.61%	4.26%	17.25%	1.55%
$\text{ARIMA}(2,1,2)(0,0,0)_{24}$	100	6.23%	4.90%	18.89%	1.14%
$\text{ARIMA}(1,1,1)(0,0,0)_{24}$	100	6.19%	4.88%	18.90%	1.27%

the residual series. Thus, this is a trained ARIMA model. The forecasting result from this ARIMA model is plotted in Figure 4.13. The hourly forecast power output is very close to the recorded power output and the nRMSE forecasting error is only 2.04%.

4.3.2 RBFNN based Forecasting

The non-linear autoregressive RBFNN model as shown in 4.14 can be expressed as $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-Lag})$. To train this model, the historical PV output series data P_t needs to be modified as a training matrix. For example, when Lag equals to 24, P_t will be modified into the following matrix format:

$$\begin{bmatrix} P_1 & P_2 & \cdots & P_{24} \\ P_2 & P_3 & \cdots & P_{25} \\ \vdots & \vdots & & \vdots \\ P_{T-24} & P_{T-23} & \cdots & P_{T-1} \end{bmatrix} \Rightarrow \begin{bmatrix} P_{25} \\ P_{26} \\ \vdots \\ P_T \end{bmatrix} \tag{4.4}$$

where, the left matrix is the input matrix and the right matrix is the target matrix. This matrix indicates the structure of RBFNN is $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-24})$ and the numbers of training samples are $T - 25$ for day d. When the RBFNN model is trained by $T - Lag - 1$

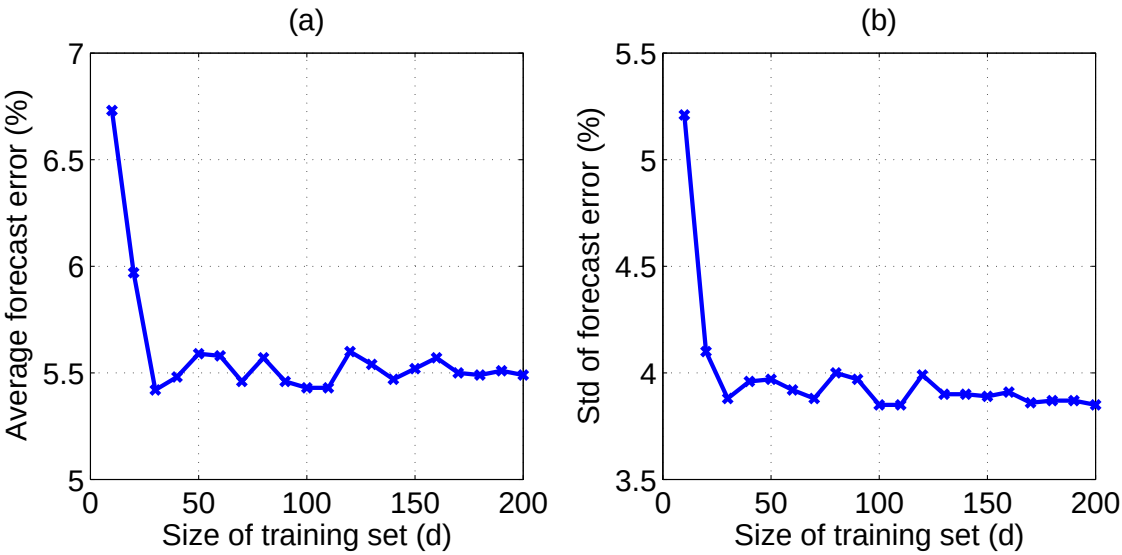

Figure 4.11: Influence of the size of the training set on the forecasting accuracy of ARIMA measured by nRMSE

training sample, RBFNN will use the past Lag data point from 24:00 of day $d-1$ as input to generate the forecast output at 1:00 of day d ($\hat{P_T}$). The $\hat{P_T}$ is then input along with $(P_{t-1}, P_{t-2} \cdots P_{t-Lag+1})$ into the original model to generate an additional forecast data point $\hat{P_{T+1}}$. This process repeats until all 24 hours $(\hat{P_T} \cdots \hat{P_{T+23}})$ have been forecasted. When the forecast day has passed, the model will be retrained to predict the next forecast day.

The forecasting accuracy of the RBFNN model is influenced by the number of training samples, the structure of the RBFNN model and the stopping criteria of the training process of RBFNN. These three aspects are simulated in the following so as to contribute to the

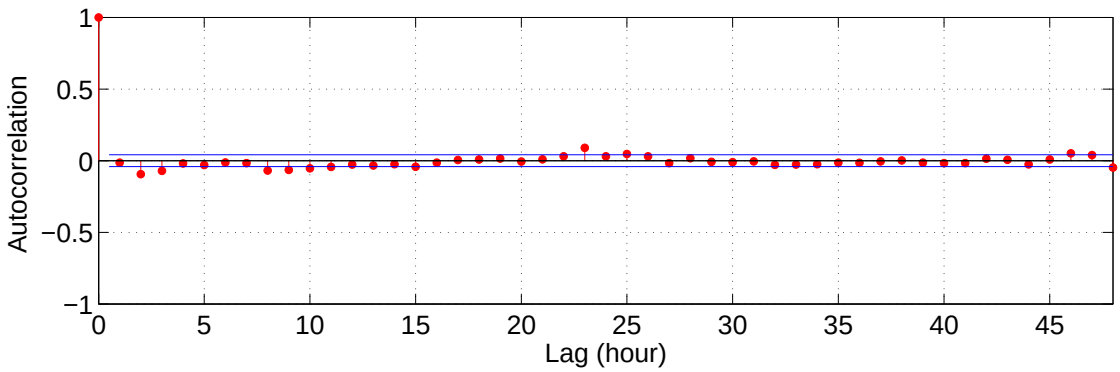

Figure 4.12: Residuals ACF plots for ARIMA model

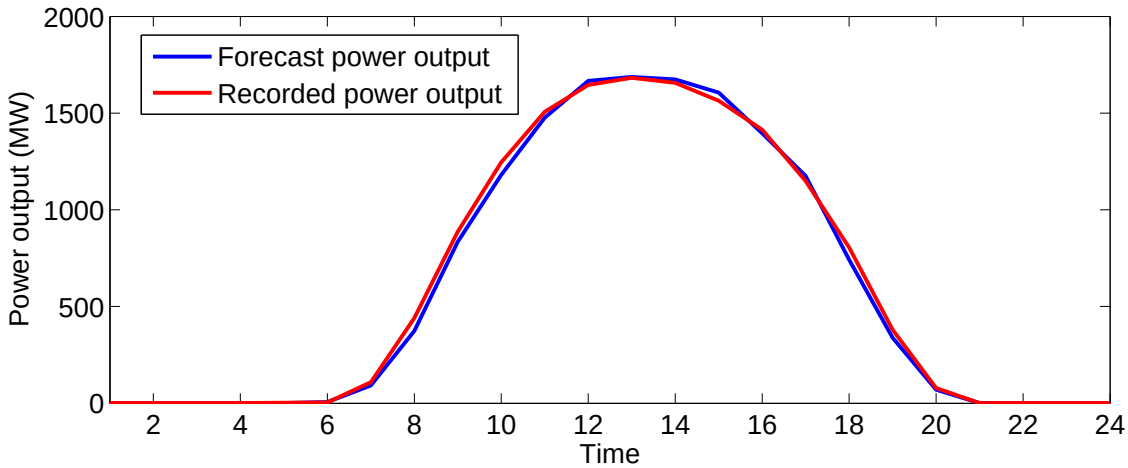

Figure 4.13: Forecasting result for July 2 using ARIMA

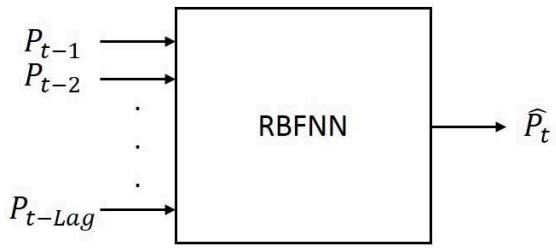

Figure 4.14: Black box structure of the RBFNN model

forecasting accuracy of RBFNN.

The model structure (or the number of inputs) is determined by the value of Lag. In order to find the value of Lag that can generate the most accurate forecast, a test was conducted with 100 training days. The stopping criteria for training is set as follows: the maximum number of hidden layer neurons is 70 and the performance goal was set as 0.0001 (MSE). The training process stopped when it reached either of them. Table 4.2 shows the forecasting accuracy changes with different Lag values. When $Lag = 24$, the forecasting error is the lowest. The average nRMSE error is 7.42%, which is much lower than others. Hence, $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-24})$ is selected as the model structure for RBFNN.

The test for the influence of the training set is conducted using $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-24})$ model. The stopping criteria of the RBFNN training were the same as the above test. This test is used to show the forecasting error changes when the training days was increased from

Table 4.2: Influence of model structure on the forecasting accuracy of RBFNN measured by nRMSE

Lag	Average	Std	Max	Min
16	17.54%	17.37%	72.71%	2.47%
20	12.44%	15.54%	72.45%	1.51%
24	**7.42%**	**3.91%**	**14.88%**	**1.54%**
36	10.25%	9.21%	39.79%	1.07%

10 days to 190 days. Figure 4.15 (a) and (b) shows the forecasting accuracy changes when adjusting the size of the training set in terms of nRMSE. The training set should be at least 30 to gain an average forecasting accuracy less than 10%. Within the range of 90 days to 110 days, the forecast error is low and stable. In this study, 100 days was chosen as the number of training days.

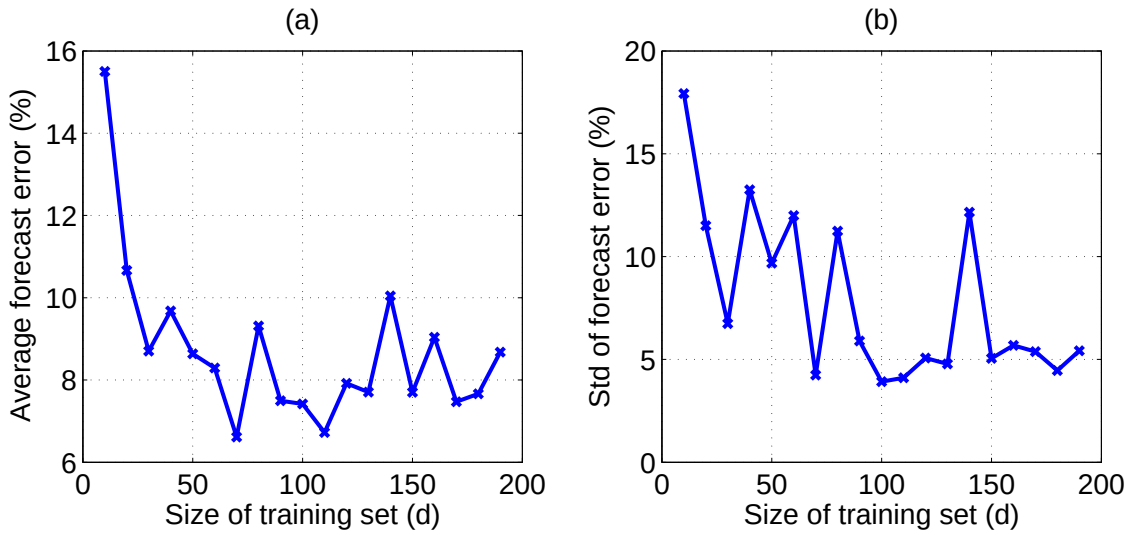

Figure 4.15: Influence of the size of the training set on the forecasting accuracy of RBFNN measured by nRMSE

The influence of the stopping criteria for the training process was tested through trial and errors for the testing period. In the following, different stopping criteria combinations of the maximum number of hidden layers neurons (MN) and the performance goal was simulated with $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-24})$ model and 100 training days. The training process will stop

Table 4.3: Influence of the stopping criteria for the training process on the forecasting accuracy of RBFNN measured by nRMSE

goal	0.0001	0.0001	**0.0001**	0.0001	0.0001	0.0001	0.001	0.001	0.001
MN	90	80	**70**	60	50	40	90	80	70
Average	6.82%	6.98%	**6.72%**	7.03%	7.78%	9.17%	7.46%	7.46%	7.46%
Std	4.41%	4.43%	**4.11%**	4.38%	4.69%	6.37%	4.56%	4.56%	4.56%
Max	17.50%	16.79%	**17.13%**	17.13%	19.96%	26.71%	17.74%	17.74%	17.74%
Min	1.54%	1.13%	**1.81%**	1.78%	1.92%	1.42%	2.02%	2.02%	2.02%
goal	0.001	0.001	0.001	0.01	0.01	0.01	0.01	0.01	0.01
MN	60	50	40	90	80	70	60	50	40
Average	7.46%	7.46%	7.46%	12.41%	12.41%	12.41%	12.41%	12.41%	12.41%
Std	4.56%	4.56%	4.56%	5.35%	5.35%	5.35%	5.35%	5.35%	5.35%
Max	17.74%	17.74%	17.74%	23.18%	23.18%	23.18%	23.18%	23.18%	23.18%
Min	2.02%	2.02%	2.02%	5.44%	5.44%	5.44%	5.44%	5.44%	5.44%

when either of these two criteria are reached. Figure 4.17 shows the training process for July 2, when MN is set as 70 and goal as 0.0001. As the increase of newly added hidden neurons in each epochs, the performance drops. The training process stoups when it reaches the limit of maximum MN (70) first, while the performance does not reach the goal. The simulation results are summarised in Table 4.3. The best forecasting accuracy is achieved when setting the stopping goal as 0.0001 and MN as 70.

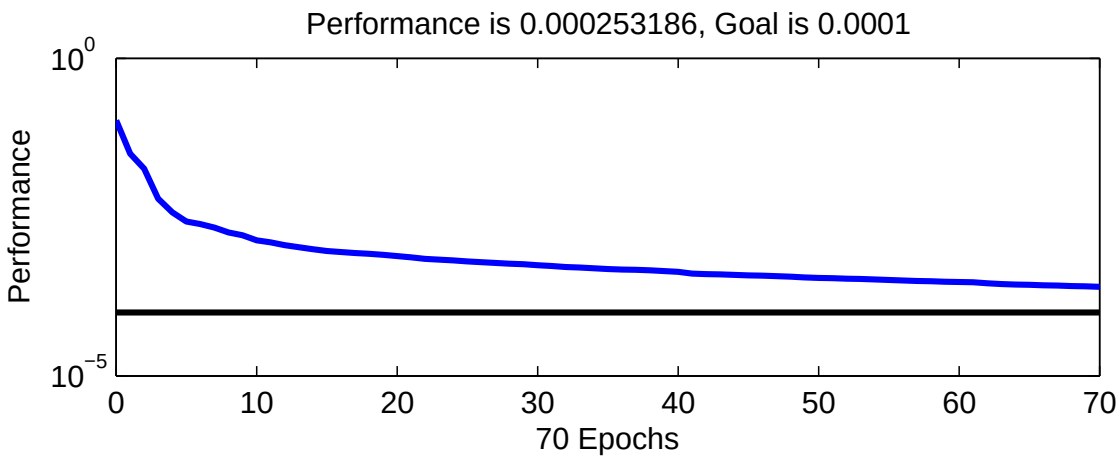

Figure 4.16: Training process of the RBFNN model for July 2 prediction

Using the above configuration, RBFNN is used to predict output for July 2. Figure 4.16 shows the training process of RBFNN model for July 2 prediction. The training process

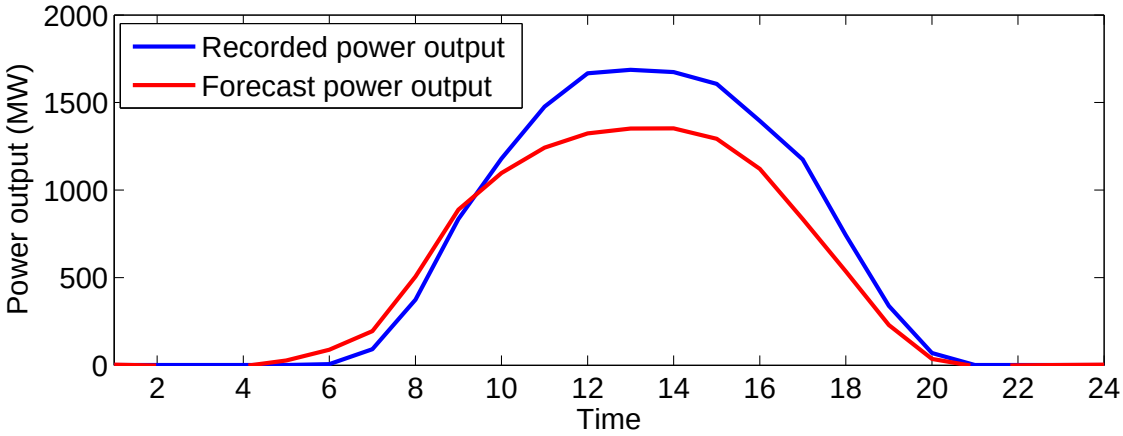

Figure 4.17: Forecasting result for July 2 using RBFNN

stopped when it reached the limit of maximum number of hidden layer neurons. Thus, network trained for this day has 70 hidden layer neurons. Figure 4.17 shows the forecasting result for July 2 using RBFNN with the above configuration. The forecasting error is 12.16% (nRMSE).

4.3.3 LS-SVM based Forecasting

The non-linear autoregressive LS-SVM model as shown in 4.18 can also be expressed as $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-Lag})$. To train this model, the historical PV data P_t needs to be modified as a training matrix and fed into the LS-SVM toolbox to generate the forecast [69]. Similar to the RBFNN model, LS-SVM will recursively generate the next day's output.

The forecasting accuracy of the LS-SVM model is influenced by the number of training samples, the structure of the LS-SVM model and the model parameters' optimization method. These three aspects are simulated in the following so as to contribute to the forecasting accuracy of LS-SVM.

The model structure (or the number of inputs) for LS-SVM is also determined by the value of Lag. In order to find the value of Lag that can generate the most accurate forecast, a test was conducted with 5 training days. Radial basis function was selected as the kernel function and the parameter of this function is optimized by simplex method. Table 4.2 shows

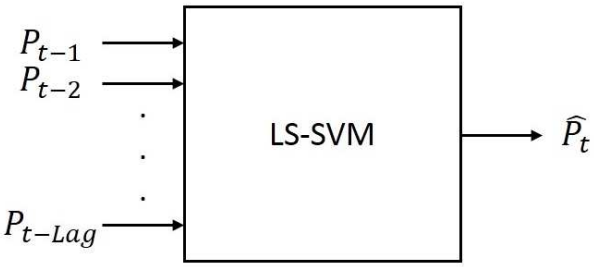

Figure 4.18: Black box structure of the LS-SVM model

the forecasting accuracy changes with different Lag values. When $Lag = 12$, the forecasting error is the lowest. The average nRMSE error is 5.96%. Hence, $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-12})$ is selected as the model structure for LS-SVM.

Table 4.4: Influence of model structure on the forecasting accuracy of LSSVM measured by nRMSE

Lag	Average	Std	Max	Min
12	**5.96%**	**4.11%**	**16.54%**	**2.10%**
24	6.51%	4.82%	16.42%	1.44%
36	6.23%	4.89%	18.44%	0.90%
48	5.93%	4.53%	18.81%	1.28%

The test for the influence of training set is conducted using $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-12})$ model, radial basis kernel function and simplex optimization method, by changing the number of training days between 5 and 25. Figure 4.15 (a) and (b) shows the forecasting accuracy changes in terms of nRMSE. The best forecasting accuracy is achieved by 5 training days when considering both the average forecasting error and standard deviation of the forecasting error. In this study, 5 training days were chosen.

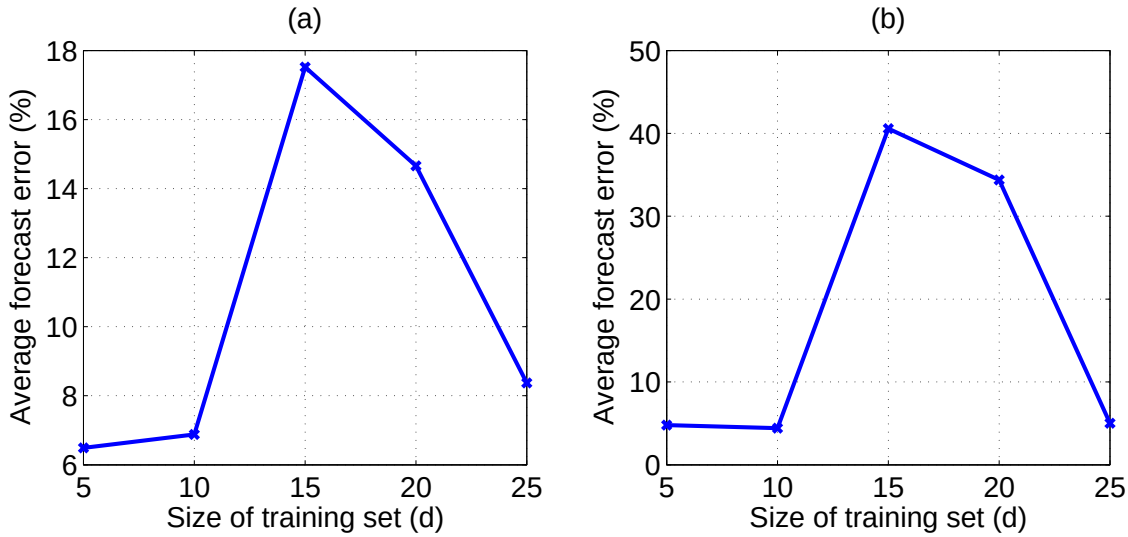

Figure 4.19: Influence of the size of the training set on the forecasting accuracy of LS-SVM measured by nRMSE

Table 4.5: Influence of kernel function and optimization method for tuning kernel parameters of LSSVM measured by nRMSE

Kernel Function	Optimization Method of kernel parameters	Average	Std
RBF	**Simplex**	**5.96%**	**4.11%**
RBF	Grid search	6.05%	4.34%
Linear	Simplex	14.60%	3.93%
Linear	Grid search	14.61%	3.94%
Polynomial	Simplex	6.80%	3.61%

The influence of kernel function and the corresponding parameters optimization method is tested through trial and error for the testing period. One challenge of LS-SVM forecasting is that there is no optimal method to select the kernel function and free parameters [76]. Therefore scenarios were developed to examine different combinations when using 5 days as the training set and $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-12})$ as the model structure. Table 4.5 shows radial basis function with simplex method has a better performance.

Similarly, July 2 is used to illustrate the prediction process of theLS-SVM model. Trained by five previous days, parameter γ and σ^2 of the radial basis kernel function is optimized

as 212.025 and 9.8278. Figure 4.20 shows the function estimated by power data from June 28 to July 1 and the above kernel function. The blue dots are the power data points and the red line is the estimated function. When using the last 12 data points as input for this trained LS-SVM model, the result of the forecasting is plotted in Figure 4.21. The predicted output shown by the blue line this figure is very close to the recorded power output, shown as the red line. The forecasting error in terms of nRMSE for July 2 is 2.25%.

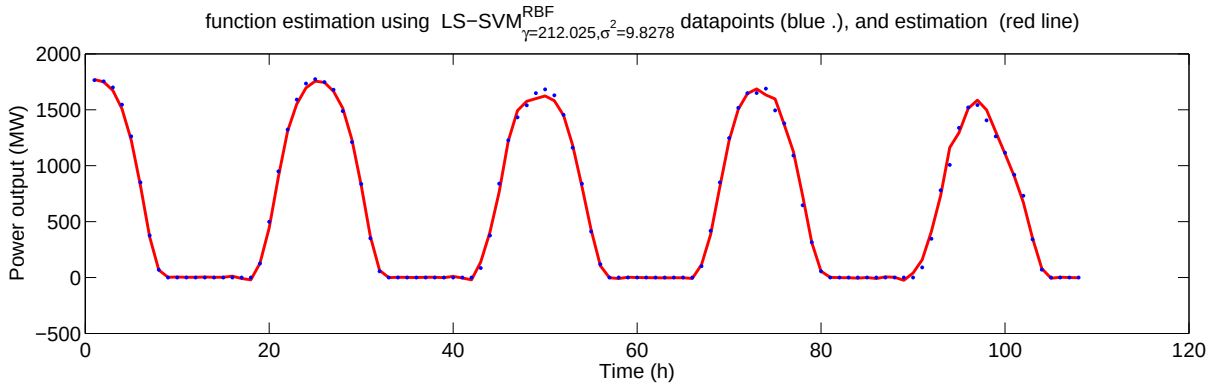

Figure 4.20: LS-SVM estimation result in the training environment

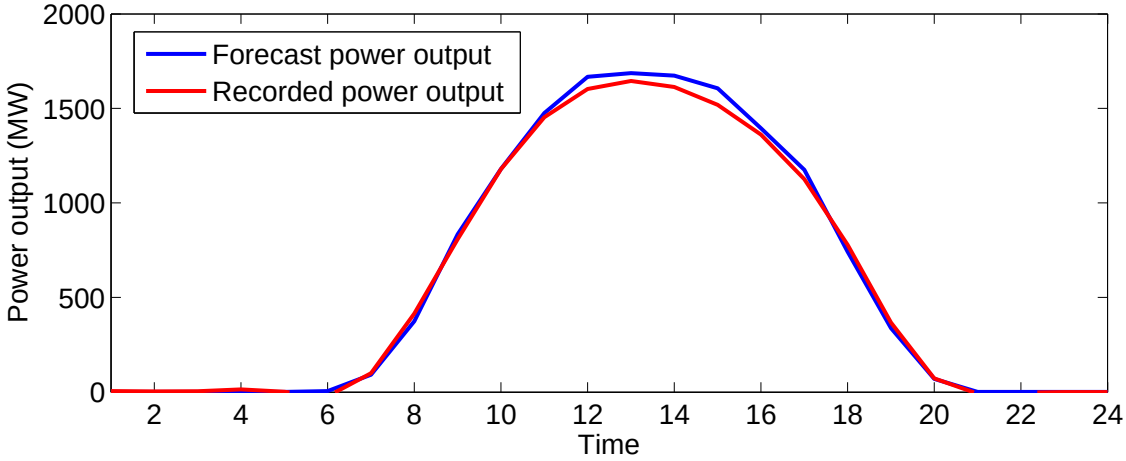

Figure 4.21: Forecasting result for July 2 using LS-SVM

4.3.4 Summary

Through the above modelling process, the final model structures were found for each of them. The best structure of the ARIMA model is ARIMA$(0,1,1)(1,0,1)_{24}$ and the optimal training

set is 100 days. The best structure of the RBFNN model is $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-24})$ which is trained by 100 days. In the training process the stopping goal is 0.0001 and the maximum number of hidden layer neurons is 70. The best structure of the LS-SVM model is $\hat{P}_t = f(P_{t-1}, P_{t-2} \cdots P_{t-12})$ which is trained by 5 days. Radial basis function was selected as the kernel function and the free parameters are optimized by the simplex method.

4.4 Results and Discussion

In this section, the forecasting accuracy of the system level forecasting tool built by the three different models is evaluated and compared to a bench mark persistence model. Through comparison, this study examines which model is suitable for aggregate system level forecasting. In the end, the forecasting accuracy of the array level forecast tool and the system level forecast tool is compared and analyzed.

4.4.1 Comparison of Different Models

In this section, the forecasting accuracy of the above three models are evaluated and then compared to a benchmark persistence model. Through evaluation, this study examines which forecast models are more suitable to build a system level forecasting tool.

For the three-weeks testing period, ARIMA, LS-SVM and RBFNN models were compared to a persistence model. Figure 4.22 (a) shows the daily error (nRMSE) of the persistence model and Figure 4.22 (b) shows the forecasting accuracy improvement over the persistence model using the other three models. The improvement is based on the error difference. For example, the persistence model's forecasting error for June 25 was 18.89%, and the forecasting error of ARIMA, LS-SVM and RBFNN models for June 25 is 6.10%, 6.40% and 7.95% respectively. Correspondingly, the forecast improvement over PM for ARIMA, LS-SVM and RBFNN is 12.79% (18.89%-6.10%), 12.49% (18.89%-6.40%) and 10.94% (18.89%-7.95%). Seen from figure 4.22 (a), persistence model can generate very accurate forecasts

for some days (e.g June 18, June 22 and June 26). However, forecasting accuracy is not constant and the persistence model produces very inaccurate forecasting for days like June 21 and June 25.

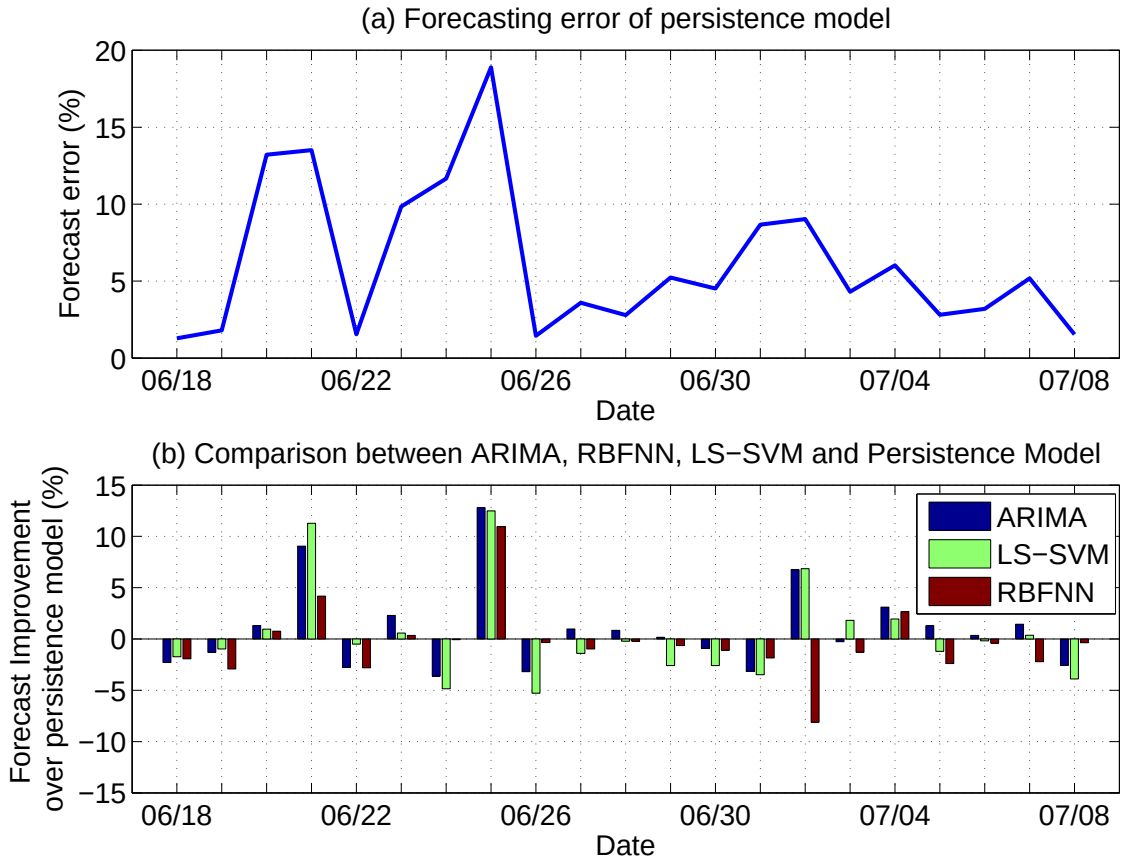

Figure 4.22: Comparison about ARIMA,LS-SVM and RBFNN over persistence model

Compared to the persistence model, the chosen models may not perform well for days like July 18 and 19, when the persistence model has very good accuracy. Although the low intra-day fluctuation level of system level data allows the persistence model to have very good accuracy for some days, it cannot give an accurate prediction when the output for the day before the forecast day is significantly different from the forecast day, like June 25. For those days, the ARIMA, LS-SVM and RBFNN models have better performance. Since we do not know the intra-day fluctuation between forecast day and the day before it in real time, it is suitable to select models that can handle high intra-day fluctuation well, such as the

ARIMA model. Although they have a slightly higher error for days with very low intra-day fluctuation, they can handle forecast days that have very high intra-day fluctuation. The RBFNN performs the worst of these models. It may produce big errors for days that RBFNN is not trained well on, like July 2.

Figure 4.23 plots the measured power output and the predicted power output for four testing days to show the difference in the forecasting models. Obviously, the RBFNN model is untrained for June 21, as the predicted output is higher than 2000 MW at noon, which is impossible. Similarly, RBFNN is not trained very well for July 2 and predicted a bad result. For June 19 and 20, these models have a similar performance.

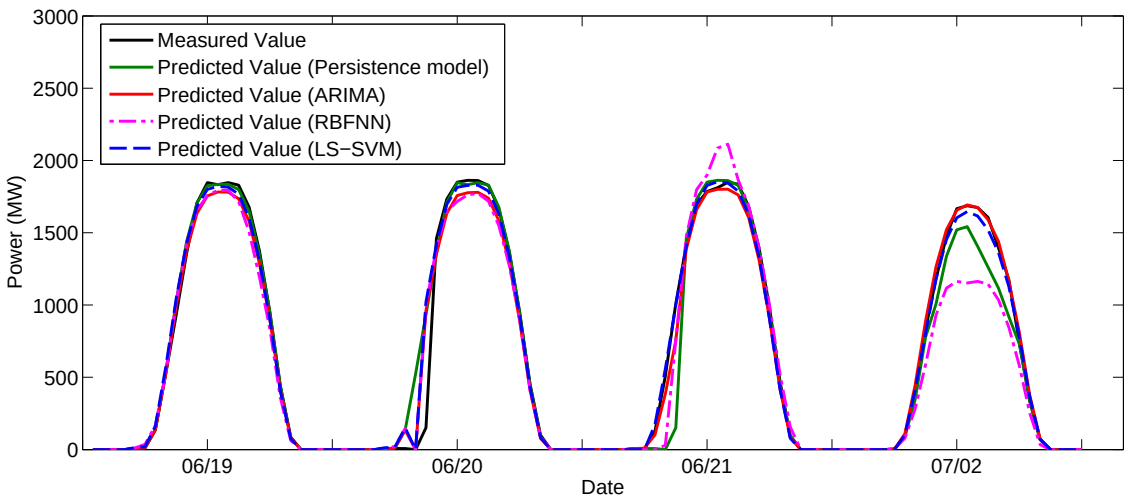

Figure 4.23: Measured power and predicted power from Persistence Model, ARIMA, LS-SVM and RBFNN

Table 4.6 shows the forecasting accuracy summary for the testing period from the above three models. Compared to the LS-SVM and the RBFNN, ARIMA has the best forecasting accuracy. In term of average nMAE error, ARIMA is 14.11% less than LS-SVM and 28.21% less than RBFNN. In terms of the standard deviation of nMAE error, ARIMA is 6.16% less than LS-SVM and 8.70% less than RBFNN. Moreover, ARIMA performs better than RBFNN and LS-SVM using RMSE and nRMSE as well. Hence, ARIMA has better performance than the other two models.

Table 4.6: Forecasting accuracy of ARIMA, LS-SVM and RBFNN

	ARIMA		LS-SVM		RBFNN	
	nMAE	nRMSE	nMAE	nRMSE	nMAE	nRMSE
Mean	3.97%	5.21%	4.53%	5.85%	5.09%	6.61%
Std	2.76%	3.71%	2.93%	3.91%	3.00%	3.94%

In summary, when the output pattern of two consecutive days are very similar, the persistence model usually performed better than the other three models. But when the daily power difference is significant, ARIMA and LS-SVM usually performed better than persistence model. The RBFNN model is vulnerable to high intra-day fluctuation, so it produced a big errors for some days. In summary, ARIMA is the best choice for aggregated system level power forecasting.

4.4.2 Comparison Between System Level and Array Level Forecasting

In comparison, the maximum intra-day fluctuation of system level power output is 180% lower than that of array level power data and the standard deviation of intra-day fluctuation of system level power output is 235% lower than that of array level power data. A lower intra-day fluctuation level makes system level data is easer to predict.

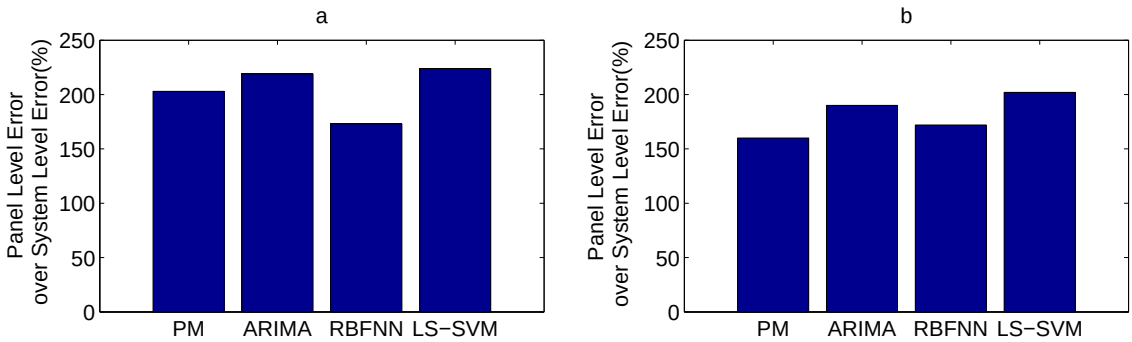

Figure 4.24: Forecasting accuracy comparison between system level forecasting tool and array level forecasting tool based on average nRMSE error: (a)The array level tool is built without data-preprocessing engine,(b) The array level tool is built with data-preprocessing engine

To illustrate the influence of intra-day fluctuation, the forecasting accuracy of system level

forecasting is compared with array level forecasting in Figure 4.24. Figure 4.24 (a) shows the difference of system level forecasting error and the array level forecasting error using four models without data-preprocessing engine. Each bar shows the array level forecasting error over the system level forecasting error. For example, the average forecasting error for array level forecasting tool using the persistence model without data-preprocessing engine is 12.55%, while the average forecasting error for the system level forecasting tool using the persistence model is 6.19%. Thus, the difference is 203%, as shown in the bar for persistence model in Figure 4.24 (a). Without using data-preprocessing, the forecasting error difference of the array level forecasting tool and the system level forecasting tool ranges from 173% to 224%. This difference is similar to the intra-day fluctuation. Figure 4.24 (b) shows the difference of system level forecasting error and array level forecasting error using four models with the data-preprocessing engine. The difference drops to the range of 160% to 202%. On the one hand, the data-preprocessing engine can help the forecasting engine to deal with the fluctuation and reduce the array level forecasting error. On the other hand, the influence of intra-day fluctuation is significant, a lower fluctuation of system level power data makes it easer to forecast, even without the data-preprocessing engine.

4.5 Summary

In this chapter, a short-term system level forecasting tool was established. Because of the geographical diversity of solar arrays across a wide area, it shows a different feature of a single array level PV system. System level data has an increasing trend which is lead by the new PV capacity. In terms of intra-day fluctuation characteristics, system level data is 50% less than array level power output. After that, a forecasting tool was generated for system level forecasting. Through simulation, ARIMA was found to have better forecasting accuracy compared to LS-SVM and RBFNN. In terms of nRMSE, the average forecasting error of ARIMA is 5.42% for a three-week testing period. The forecasting accuracy of the

system level forecasting tool was then compared to the array level forecasting tool. Through comparison, a lower intra-day fluctuation leads to a similar range of forecast error reduction. In the future, through analyzing the changes of intra-day fluctuation of system level data, the potential forecasting error of the proposed tool could be appropriately estimated.

Chapter 5

Conclusions

5.1 Summary and Conclusions

In this thesis, a short-term forecasting tool was proposed to predict array level and aggregated system level PV power output. Because of the high intra-day fluctuation of array level power output, the forecasting tool includes both a similar day based data-preprocessing engine and a forecasting engine. The data-preprocessing engine is designed through analyzing the relationship between the power output and output-related variables. Through analysis, a group of forecast output-related variables is used to build the similar day selection formula to choose the similar day training set and then fed into the forecasting engine. This tool is tested for three locations around the world with substantially different climates. The forecasting result shows that this new hybrid similar day-based data-preprocessing engine is more effective than previous methods described in the literature. Moreover, the overall forecasting accuracy of this array level forecasting tool is compared to an ARX model in the literature. For the location of Breadstrup in Denmark, better forecasting accuracy is achieved through the proposed tool.

The intra-day fluctuation of aggregated system level data is relatively lower and the forecasting tool includes only the forecasting engine. Candidate models for this forecasting engine included the persistence model, ARIMA, LS-SVM and RBFNN. The persistence model is used to generate benchmark results. The ARIMA, LS-SVM and RBFNN models are simulated to find the optimal structure, training set size and training process. Through result comparison, the ARIMA model has the best forecasting accuracy compared to the other four models.

5.2 Contributions

The focus of this study is short-term PV power forecasting. The contributions of array level forecasting in Chapter 3 are:

- A hybrid similar day method is proposed to build the data-processing engine. The similar day selection algorithm is proposed from the available output-related variables. Compared to the similar day methods in the literature, the proposed similar day method improves forecasting accuracy.
- The overall accuracy of the array level forecasting tool is better than an ARX model in the literature.

Following is the summary of the contributions of aggregated system level forecasting in Chapter 4:

- Currently, most of the research concentrates on array level forecasting. This system forecasting tool is proposed to fill the gap for the system level PV output field.
- The intra-day fluctuation level of system level data is analyzed. Compared to an array level PV system, the fluctuation is 50% lower. Thus, the forecasting accuracy of system level data has a better accuracy.
- Through the forecasting results comparison between array level and system level forecasting, this study finds the persistence model does not work very well for aggregated power output forecasting. This computationally effective model works for array level forecasting because of the assistance of a data pre-processing engine.

5.3 Directions for Future Work

Based on the above research works, further research work may look at the following subjects:

1. Investigate the application of a multivariate forecasting engine in array level forecasting. Forecast output-related variables will be allocated to the data-preprocessing engine and the forecasting engine, so as to test whether better forecasting accuracy can be achieved.

2. Currently, the aggregated system level forecasting only has the forecasting engine. The data pre-processing engine functions well in the array level forecasting tool. In the future, the regional weather of California as well as the general distribution of PV sites within this grid should be investigated to build a system level similar day-based data pre-processing engine.

Bibliography

[1] J. Byrne and L. Kurdgelashvili, "The role of policy in PV industry growth: past, present and future," *Handbook of photovoltaic science and engineering. 2nd ed. UK: John Wiley & Sons, Ltd*, 2011.

[2] "Global market output for photovoltaic until 2016," EPIA, Tech. Rep., 2012. [Online]. Available: http://files.epia.org/files/Global-Market-Outlook-2016.pdf

[3] A. Brown, S. muller, and Z. Dobrotkova, "Renewable energy markets and prospects by technology," Tech. Rep., 2011 November.

[4] "Technology roadmap. solar photovoltaic energy," International Energy Agency, Tech. Rep., 2010.

[5] NREL. National renewable energy laboratory. [Online]. Available: http://www.nrel.gov/ncpv/

[6] J. Tian, Y. qiang Zhu, and J. neng Tang, "Photovoltaic array power forecasting model based on energy storage," in *Critical Infrastructure (CRIS), 2010 5th International Conference on*, Sep 2010, pp. 1 –4.

[7] E. Lorenz, J. Hurka, D. Heinemann, and H. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 2, no. 1, pp. 2 –10, March 2009.

[8] V. Lara-Fanego, J. Ruiz-Arias, D. Pozo-Vazquez, F. Santos-Alamillos, and J. Tovar-Pescador, "Evaluation of the WRF model solar irradiance forecasts in Andalusia (southern Spain)," *Solar Energy*, vol. 86, no. 8, pp. 2200 – 2217, 2012, progress in Solar Energy 3.

[9] S. Pelland, G. Galanis, and G. Kallos, "Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model," *Progress in Photovoltaics: Research and Applications*, vol. 21, no. 3, pp. 284–296, 2013. [Online]. Available: http://dx.doi.org/10.1002/pip.1180

[10] R. Perez, K. Moore, S. Wilcox, D. Renn, and A. Zelenka, "Forecasting solar radiation-preliminary evaluation of an approach based upon the national forecast database," *Solar Energy*, vol. 81, no. 6, pp. 809 – 812, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X06002404

[11] P. Mathiesen and J. Kleissl, "Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States," *Solar Energy*, vol. 85, no. 5, pp. 967 – 977, 2011.

[12] J. Remund, R. Perez, and E. Lorenz, "Comparison of solar radiation forecasts for the USA," in *Proc. of the 23rd European PV Conference*, 2008, pp. 1–9.

[13] P. Mathiesen, C. Collier, and J. Kleissl, "A high-resolution, cloud-assimilating numerical weather prediction model for solar irradiance forecasting," *Solar Energy*, vol. 92, no. 0, pp. 47 – 61, 2013.

[14] A. Hammer, D. Heinemann, E. Lorenz, and B. Luckehe, "Short-term forecasting of solar radiation: a statistical approach using satellite data," *Solar Energy*, vol. 67, pp. 139 – 150, 1999.

[15] R. Perez, S. Kivalov, J. Schlemmer, K. H. Jr., D. Renn, and T. E. Hoff, "Validation of short and medium term operational solar radiation forecasts in the US," *Solar Energy*, vol. 84, no. 12, pp. 2161 – 2172, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10002823

[16] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, and

B. Washom, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," *Solar Energy*, vol. 85, no. 11, pp. 2881 – 2893, 2011.

[17] F. Besharat, A. A. Dehghan, and A. R. Faghih, "Empirical models for estimating global solar radiation: A review and case study," *Renewable and Sustainable Energy Reviews*, vol. 21, no. 0, pp. 798 – 821, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032112007484

[18] A. Mellit and S. A. Kalogirou, "Artificial intelligence techniques for photovoltaic applications: A review," *Progress in Energy and Combustion Science*, vol. 34, no. 5, pp. 574 – 632, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360128508000026

[19] A. Sfetsos and A. Coonick, "Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques," *Solar Energy*, vol. 68, no. 2, pp. 169 – 178, 2000.

[20] C. Voyant, M. Muselli, C. Paoli, and M.-L. Nivet, "Hybrid methodology for hourly global radiation forecasting in mediterranean area," *Renewable Energy*, vol. 53, no. 0, pp. 1 – 11, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148112007008

[21] J. Zeng and W. Qiao, "Short-term solar power prediction using a support vector machine," *Renewable Energy*, vol. 52, no. 0, pp. 118 – 127, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148112006465

[22] C. Voyant, M. Muselli, C. Paoli, and M.-L. Nivet, "Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation," *Energy*, vol. 39, no. 1, pp. 341 – 355, 2012, sustainable Energy and Environmental Protection 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360544212000114

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

[24] B. Zhao, X. Ge, M. Xue, X. Zhang, and W. Xu, "Research on model for photovoltaic system power forecasting," in *Electricity Distribution (CICED), 2010 China International Conference on*, sept. 2010, pp. 1 –5.

[25] 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*, oct. 2010, pp. 1 –6.

[26] D. Caputo, F. Grimaccia, M. Mussetta, and R. Zich, "Photovoltaic plants predictive model by means of ANN trained by a hybrid evolutionary algorithm," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, july 2010, pp. 1 –6.

[27] T. Cai, S. Duan, and C. Chen, "Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement," in *Power Electronics for Distributed Generation Systems (PEDG), 2010 2nd IEEE International Symposium on*, Jun. 2010, pp. 773 –777.

[28] C. Chupong and B. Plangklang, "Forecasting power output of pv grid connected system in Thailand without using solar radiation measurement," *Energy Procedia*, vol. 9, no. 0, pp. 230 – 237, 2011, 9th Eco-Energy and Materials Science and Engineering Symposium.

[29] F. Grimaccia, M. Mussetta, and R. Zich, "Neuro-fuzzy predictive model for pv energy production based on weather forecast," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, june 2011, pp. 2454 –2457.

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

[31] 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," *Industry Applications, IEEE Transactions on*, vol. 48, no. 3, pp. 1064 –1069, May-Jun 2012.

[32] M. Ding, L. Wang, and R. Bi, "An ANN-based approach for forecasting the power output of photovoltaic system," *Procedia Environmental Sciences*, vol. 11, Part C, no. 0, pp. 1308 – 1315, 2011.

[33] 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.

[34] P. Mandal, S. T. S. Madhira, A. U. haque, J. Meng, and R. L. Pineda, "Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques," *Procedia Computer Science*, vol. 12, no. 0, pp. 332 – 337, 2012, complex Adaptive Systems 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050912006710

[35] S. K. Chow, E. W. Lee, and D. H. Li, "Short-term prediction of photovoltaic energy generation by intelligent approach," *Energy and Buildings*, vol. 55, no. 0, pp. 660 – 667, 2012, cool Roofs, Cool Pavements, Cool Cities, and Cool World. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037877881200415X

[36] H. T. Pedro and C. F. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," *Solar Energy*, vol. 86, no. 7, pp. 2017 – 2028, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X12001429

[37] L. A. Fernandez-Jimenez, A. Muoz-Jimenez, A. Falces, M. Mendoza-Villena, E. Garcia-Garrido, P. M. Lara-Santillan, E. Zorzano-Alba, and P. J. Zorzano-Santamaria, "Short-term power forecasting system for photovoltaic plants,"

Renewable Energy, vol. 44, no. 0, pp. 311 – 317, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148112001516

[38] B. Kraas, M. Schroedter-Homscheidt, and R. Madlener, "Economic merits of a state-of-the-art concentrating solar power forecasting system for participation in the spanish electricity market," *Solar Energy*, vol. 93, no. 0, pp. 244 – 255, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13001527

[39] M. Wittmann, H. Breitkreuz, M. Schroedter-Homscheidt, and M. Eck, "Case studies on the use of solar irradiance forecast for optimized operation strategies of solar thermal power plants," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 1, no. 1, pp. 18–27, 2008.

[40] M. Roger and J. Messenger, *Photovoltaic systems engineering.* CRC Press, 2000.

[41] T. Markvart, *Solar electricity.* John Wiley & Sons, 2000.

[42] A. B. Meinel and M. P. Meinel, "Applied solar energy. an introduction," 1976.

[43] H. C. Hottel, "A simple model for estimating the transmittance of direct solar radiation through clear atmospheres," *Solar Energy*, vol. 18, no. 2, pp. 129 – 134, 1976. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0038092X76900451

[44] A. Goetzberger and V. U. Hoffmann, *Photovoltaic solar energy generation.* Springer, 2005, vol. 112.

[45] H.-L. Tsai, C.-S. Tu, and Y.-J. Su, "Development of generalized photovoltaic model using matlab/simulink," in *Proceedings of the World Congress on Engineering and Computer Science*, 2008, pp. 846–851.

[46] T. Markvart and L. Castaner, *Solar cells: materials, manufacture and operation.* Access Online via Elsevier, 2004.

[47] T. Bruton, "General trends about photovoltaics based on crystalline silicon," *Solar Energy Materials and Solar Cells*, vol. 72, pp. 3 – 10, 2002, eMRS 2001 Symposium E: Crystalline Silicon for Solar. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0927024801001453

[48] A. Yona, T. Senjyu, and T. Funabashi, "Application of recurrent neural network to short-term-ahead generating power forecasting for photovoltaic system," in *Power Engineering Society General Meeting, 2007. IEEE*, Jun 2007, pp. 1 –6.

[49] M. Santarelli and S. Macagno, "A thermoeconomic analysis of a PV-hydrogen system feeding the energy requests of a residential building in an isolated valley of the alps," *Energy Conversion and Management*, vol. 45, no. 3, pp. 427 – 451, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196890403001560

[50] G. Stapleton and S. Neill, *Grid-connected Solar Electric Systems.* Routledge, 2012.

[51] Enphase Energy. [Online]. Available: https://englighten.enphaseenergy.com/pv/public_systems

[52] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control.* Wiley, 2011, vol. 734.

[53] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342 – 349, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X08002107

[54] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[55] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[56] D. Wang, M. Wang, and X. Qiao, "Support vector machines regression and modeling of greenhouse environment," *Computers and Electronics in Agriculture*, vol. 66, no. 1, pp. 46–52, Apr. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169908002305

[57] D. Lowe and D. Broomhead, "Multivariable functional interpolation and adaptive networks," *Complex systems*, vol. 2, pp. 321–355, 1988.

[58] M. Benghanem and A. Mellit, "Radial basis function network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia," *Energy*, vol. 35, no. 9, pp. 3751 – 3762, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S036054421000294X

[59] W. Blog, "Weather underground". wunderground. com," Retrieved 2011-03-17, Tech. Rep.

[60] M. Lave, J. Kleissl, and E. Arias-Castro, "High-frequency irradiance fluctuations and geographic smoothing," *Solar Energy*, vol. 86, no. 8, pp. 2190 – 2199, 2012.

[61] Jan. [Online]. Available: http://maeresearch.ucsd.edu/kleissl/

[62] Peder, "Short-term solar power forecasting," Master's thesis, Technical University of Denmark, 2008.

[63] H. Zareipour, K. Bhattacharya, and C. A. Caizares, "Electricity market price volatility: The case of Ontario," *Energy Policy*, vol. 35, no. 9, pp. 4739 – 4748, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030142150700153X

[64] P. Mandal, T. Senjyu, A. Yona, J.-W. Park, and A. Srivastava, "Sensitivity analysis of similar days parameters for predicting short-term electricity price," in *Power Symposium, 2007. NAPS '07. 39th North American*, 2007, pp. 568–574.

[65] T. Hiyama and K. Kitabayashi, "Neural network based estimation of maximum power generation from pv module using environmental information," *Energy Conversion, IEEE Transactions on*, vol. 12, no. 3, pp. 241 –247, Sep 1997.

[66] I. Gultepe, *Fog and boundary layer clouds: fog visibility and forecasting.* Springer, 2007.

[67] P. Mandal, A. Srivastava, and J.-W. Park, "An effort to optimize similar days parameters for ANN-based electricity price forecasting," *Industry Applications, IEEE Transactions on*, vol. 45, no. 5, pp. 1888–1896, 2009.

[68] MATLAB, *version 7.10.0 (R2010a).* Natick, Massachusetts: The MathWorks Inc., 2010.

[69] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J. Suykens, "Ls-svmlab toolbox users guide," *KU Leuven Leuven, Belgium*, 2011.

[70] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting methods and applications.* Wiley. com, 2008.

[71] CISO. California independent system operator. [Online]. Available: http://www.caiso.com/

[72] L. Sherwood, "US solar market trends 2010," Interstate renewable energy council, Tech. Rep., 2011.

[73] "Building a sustainable energy future 2013-2016 strategic plan," California ISO, Tech. Rep., 2013. [Online]. Available: http://www.caiso.com/Documents/2014-2016StrategicPlan-ReaderFriendly.pdf

[74] "U.S. solar market insight report — Q2 2013 — excutive summary," Solar Energy Industries Association, Tech. Rep., 2013.

[75] A. Pankratz, *Forecasting with dynamic regression models.* Wiley. com, 2012.

[76] N. Sapankevych and R. Sankar, "Time series prediction using support vector machines: A survey," *Computational Intelligence Magazine, IEEE*, vol. 4, no. 2, pp. 24 –38, May 2009.