

# ARMA Time Series Modeling for Solar PV Generation Forecasting

Ninad Gaikwad and Karthikeya Devaprasad

**Abstract**—There has been a recent push to integrate renewable sources of energy into the power grid, with Solar PV being one of the most promising among them. Among other things, effective integration involves the ability to predict generation so that other sources can be curtailed or increased accordingly to meet the power demand. Time series modeling is the preferred technique due to the complexity of the factors. This project will explore the effect of various factors like training data size and resolution, and time series model order on the prediction accuracy. We will also be implementing custom programs to estimate the model parameters, make predictions and evaluate these models.

**Index Terms**—ARMA, Time Series, Solar PV, Forecasting

## I. INTRODUCTION

### A. Solar PV

Due to the irreversible effects of Global Warming, there is a worldwide initiative to integrate renewable sources of energy into the power grid. Solar Photo-Voltaic (PV) generation is the most prominent among these that is being explored and is promising due to the lower upfront costs involved, ease of installation and adoption, and the passive nature of generation with minimal human oversight. However, one of the major drawbacks associated with it is the restriction of power generation to hours of daylight. The demands of the grid that are not met by solar will have to be balanced out by other conventional and renewable sources. Therefore, the effective integration of solar will require the ability to predict the amount and duration of solar PV power generation, and effectively using other sources to match the excess demand.

### B. Time Series Forecasting

A time-series is a set of data points that is spaced at equal intervals of time. In time-series forecasting, we use time-series data to create system models which use past time instants as input and generate predictions of future time instants as outputs. This model is then used to make predictions about future instants.

Time series forecasting is a very important technique that is used in a wide array of fields like Economics, Statistics, Weather-Forecasting and Signal Processing. It is generally adopted when the system is too complex to be represented as a mathematical model of various other inputs resulting in an output. This may be because of the complex relationships between the various inputs affecting the output, or the lack of input data that is spaced at the same time intervals as the output training data. [1]

In these cases, a time series model is used, wherein the output prediction of a random variable at any time-instant is a linear

result of the random variable at past time instants. The output will not depend on other inputs to the system, unlike other conventional system models.

### C. Time Series Forecasting for Solar PV Generation Forecasting

The power generated by a Solar PV cell is dependent on various factors like Incident Solar Radiation Intensity, the ambient temperature as well as the meteorological conditions. All these factors are strongly correlated, which makes it very difficult to characterize the power generated as a function of these parameters. [2] Therefore, we choose to use a time-series model where the solar power generation is the only variable considered and the power generation at any time instant is a function of the power generated at past instants. [3]

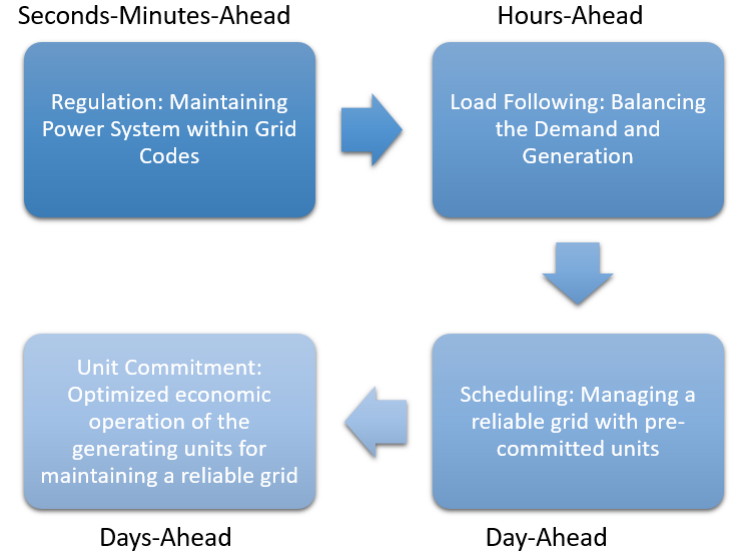


Fig. 1. Utilization of different forecasting temporal scales in Power Systems Operation

### D. Auto-Regressive Moving Average (ARMA) Model

“A stationary process is a stochastic process whose unconditional joint probability distribution does not change when shifted in time. Its parameters such as mean and variance do not change over time. Such a process can be described using an ARMA model, consisting of two polynomials. One of these is for the autoregression (AR) and the second is for the moving average (MA). The AR part involves regressing the variable on its own lagged values.

The MA part involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The model is usually referred to as the ARMA (p,q) model where p is the order of the AR part and q is the order of the MA part.”

“ARMA models can be estimated using the Box-Jenkins method.” [4] [5] [6]

## II. PROBLEM STATEMENT

Using ARMA time series modeling to create prediction models for forecasting power generated by Solar Photovoltaic (PV) arrays using the Diamond Solar data. Developing and evaluating ARMA prediction models for various time scales of the Diamond Solar Data. Finally, evaluating the effect of different estimation techniques used to estimate the parameters of the ARMA models on the performance of prediction models.

## III. AIMS AND OBJECTIVES

The aims and objectives are as follows;

- Transforming the given Diamond Solar Data (5min Resolution) to different time scales – 15 min, 60 min, 120 min, 180 min, Daily and Monthly resolution data.
- Using MATLAB’s Econometrics Toolbox to develop ARMA models for the above-mentioned time scales and predict solar power generation.
- Developing ARMA estimation models based on Least Squares (LS) and Maximum Likelihood Estimation (MLE) to develop solar power generation models at different time scales using the appropriate model orders obtained previously from the Econometrics Toolbox.
- For intra-day time scales training data of one week and two weeks will be used respectively for training the above-mentioned models and a forecast of one week into the future will be generated.
- Comparing the accuracy of the generated forecasts from the MATLAB’s Econometrics Toolbox and the developed ARMA Estimation models, different developed ARMA estimation models, different amount of training data and different time scales will be done.
- Evaluating the effect of different ARMA parameter estimation models on the forecast accuracy.
- Evaluating the effect of amount of training data on the forecast accuracy.
- Evaluating the effectiveness of ARMA prediction on different time scales

## IV. METHODS

The method comprises of three main parts: Data Pre-Processing, Development of ARMA models using MATLAB’s Econometrics Toolbox and Development of ARMA Parameter Estimation techniques based on Least Squares and Maximum Loglikelihood Estimation. The schematic of the method is illustrated in Fig. 2.

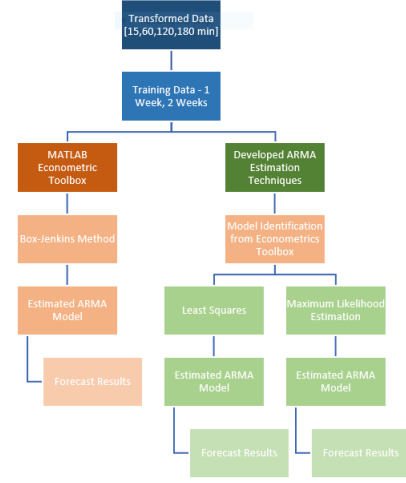


Fig. 2. ARMA Model Development Schematic

### A. Data Pre-Processing

The data file Diamond\_Solar\_data.csv consists of three different time series at 5 minutes resolution corresponding to Diamond300, Diamond304 and Diamond306 contiguously arranged in a single column. This file is processed to give individual files corresponding to Diamond300, Diamond304 and Diamond306 at different time resolutions of 5 min, 15 min, 60 min, 2 hours (120 min), 3 hours (180 min), daily and monthly. The schematic for data pre-processing is illustrated in Fig. 3. The transformed data for 15 minute and one hour intervals are plotted against the original five minute data in Fig. 4. We can observe that averaging the data over an entire hour removes the large fluctuations seen with the five-minute and fifteen-minute data. While this will provide better prediction models, it will also result in a loss of resolution.

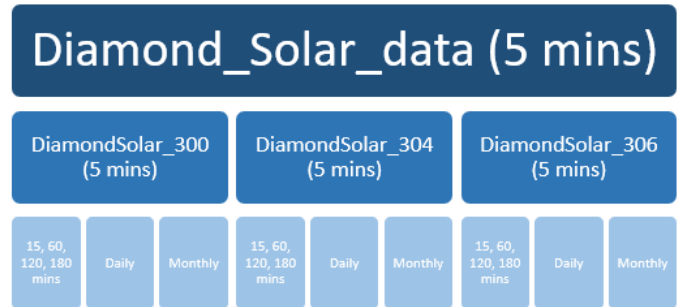


Fig. 3. Data Pre-Processing Schematic

### B. ARMA Model Development using MATLAB’s Econometrics Toolbox

The ARMA model development is done using Box-Jenkins Method [4], which has the following steps as illustrated in the Fig. 5.

- 1) **Model Identification:** The univariate time series is checked for stationarity and if not, stationary it has to be converted into a stationary series by applying differencing. The Auto-Correlation Function (ACF) and

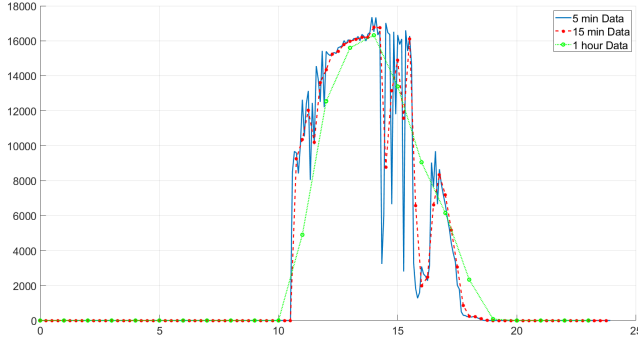


Fig. 4. Plot of Transformed Data for 15 min and 1 hour

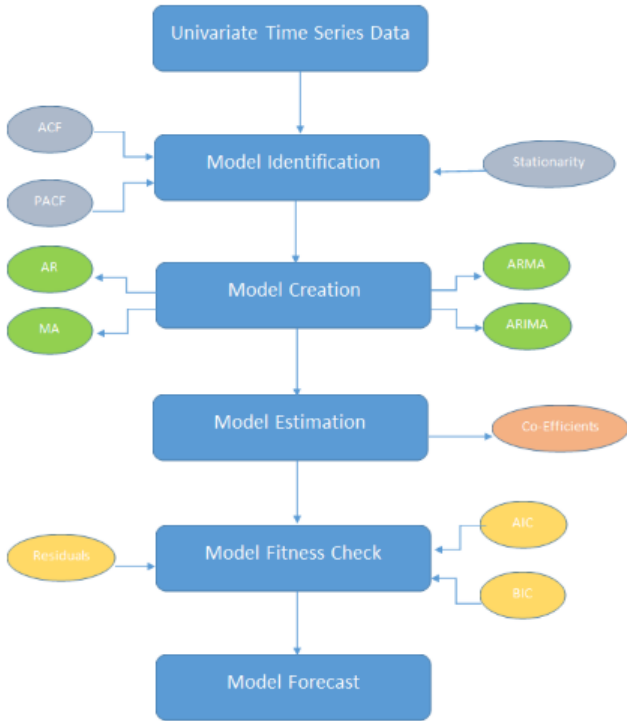


Fig. 5. ARMA Model Development Schematic

the Partial Auto-Correlation Function (PACF) of the stationary univariate time series are checked for determining the order of the AR and MA process generating the series.

- 2) **Model Creation:** On the basis of the Model Identification a number of models around the determined AR and MA orders are created.
- 3) **Model Estimation:** These models are then estimated to give the appropriate coefficients.
- 4) **Model Fitness Check:** The estimated models are then evaluated using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Lower the AIC and BIC better the model.
- 5) **Model Forecast:** The best model determined from the previous step is used to forecast the series.

### C. ARMA Parameter Estimation Techniques

1) *Least Squares Estimation Method:* The parameter estimation of an ARMA process using Least Squares is as follows;

$$\hat{y}_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + \dots + \epsilon_t$$

Now;

$$\hat{y}_t = ay + b\epsilon$$

Where;

$$a = [a_1 \ a_2 \ a_3 \dots]$$

$$b = [b_1 \ b_2 \ b_3 \dots]$$

$$y = [y_{t-1} \ y_{t-2} \ y_{t-3} \dots]^T$$

$$\epsilon = [\epsilon_{t-1} \ \epsilon_{t-2} \ \epsilon_{t-3} \dots]^T$$

The Least Squares Estimate for the Parameters is given by;

$$\hat{\theta}(a, b) = \arg \min_{\theta} \frac{1}{2} [(y_t - \hat{y}_t)^2]$$

2) *Maximum Likelihood Estimation Method:* The maximum likelihood function for an ARMA process when the error is assumed to be distributed normally is give as follows;

$$\hat{\theta}(a, b) = \arg \max_{\theta} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}}$$

## V. INITIAL EXPERIMENTS AND RESULTS

We have performed some initial experiments on the fifteen minute data and made a few forecasts assuming an AR(4) model with seasonality of 24 hours. The results are as shown in Fig. 6 and Fig. 7.

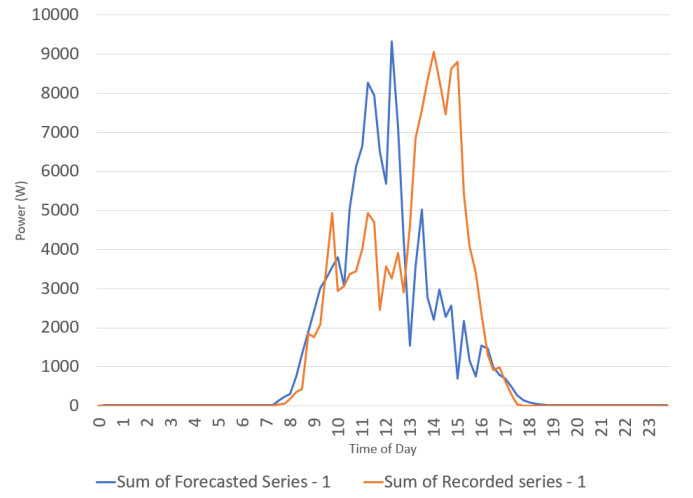


Fig. 6. Forecast results for 1st day with a week of training data

We can see that the forecast with just one week of training data appears to be closer to the actual observations than the forecast made with two weeks of training data. Therefore, we will need to use an appropriate amount of training data based on the forecast duration we are looking to make, with

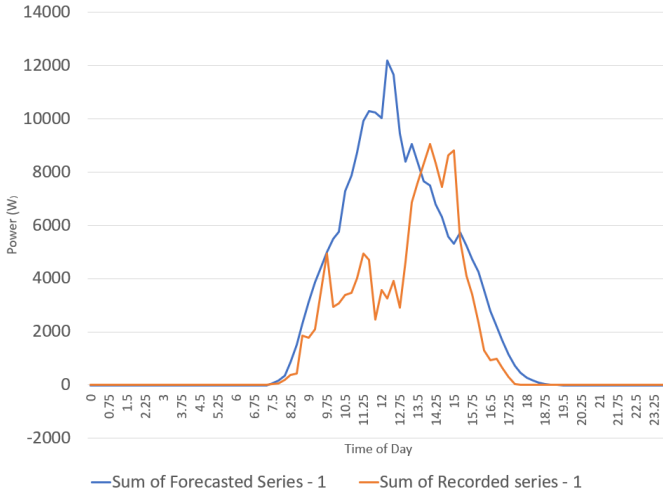


Fig. 7. Forecast results for 1st day with two weeks of training data

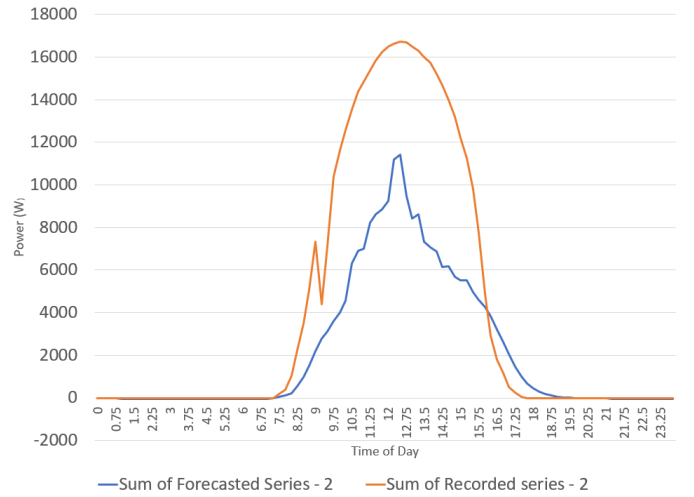


Fig. 9. Forecast results for 2nd day with two weeks of training data

a week's data sufficient to make a forecast for a single day. Apart from that, we can see that the day for which the forecast has been made has anomalous recordings due to possible cloud conditions. Therefore, we can see that even a well-tuned ARMA model cannot account for it and will not be able to predict the actual output under such conditions.

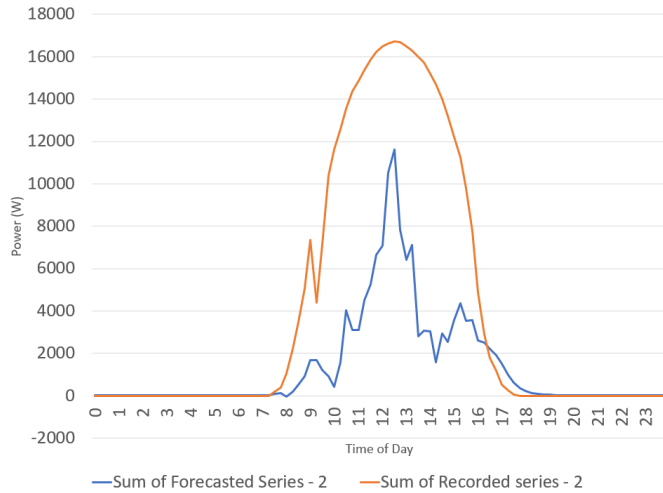


Fig. 8. Forecast results for 2nd day with a week of training data

Fig. 8 and Fig. 9 show the forecasts for the second day after the training period. We can observe that the accuracy reduces as we make forecasts further away from the training period.

## VI. RESULTS

### A. Dealing with Non-stationary Data

TABLE I  
THE KPSS AND ADF TEST FOR STATIONARITY

Result \ Test	KPSS Test	ADF Test
<b>Null Hypothesis</b>	Trend Stationary	Unit Root is present
<b>Alternative Hypothesis</b>	Unit Root is present	No Unit Root, can be stationary or trend stationary

TABLE II  
FINDING ARMA MODEL ORDER FROM ACF AND PACF PLOTS

Model	ACF	PACF
<b>AR</b>	Decaying	Significant till p Lags
<b>MA</b>	Significant till q Lags	Decaying
<b>ARMA</b>	Decaying	Decaying

### B. ARMA Model Selection and Evaluation

TABLE III  
ARMA MODELS SPECIFICATIONS FOR EVALUATION

Model	Single Lag	AR	MA	SAR	SMA
<b>AR-1</b>	0	1,2,3	0	0	0
<b>AR-2</b>	0	1,288		0	0
<b>AR-3</b>	0	1,288,576, 864,1152		0	0
<b>MA-1</b>	1	0	1,2,3	0	0
<b>MA-2</b>	1	0	1	0	288
<b>MA-3</b>	1	0	1,576, 864,1152	0	288
<b>ARMA-1</b>	1	1,2,3	1,2,3	0	0
<b>ARMA-2</b>	1	1,288	1,288	288	288
<b>ARMA-3</b>	1	1,288,576	1,288,576	288	288

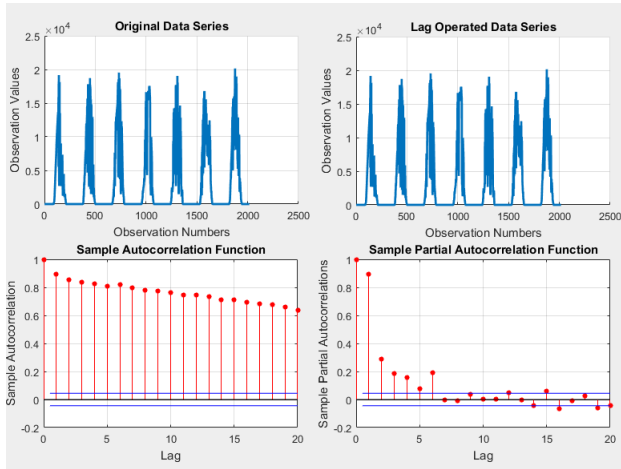


Fig. 10. Diamond Solar 300 undifferentiated data with ACF and PACF Plots

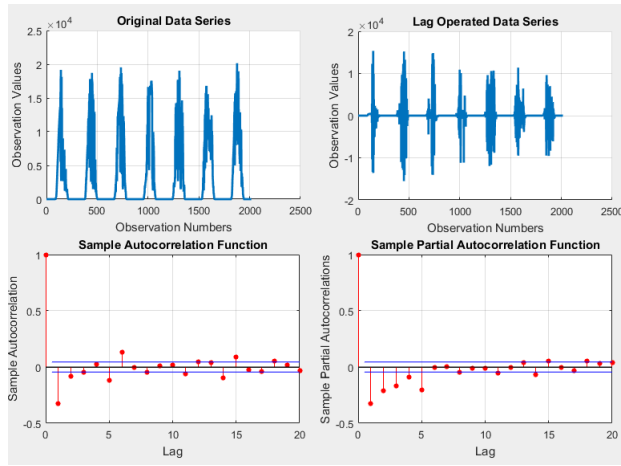


Fig. 11. Diamond Solar 300 original and single differenced data with ACF and PACF Plots

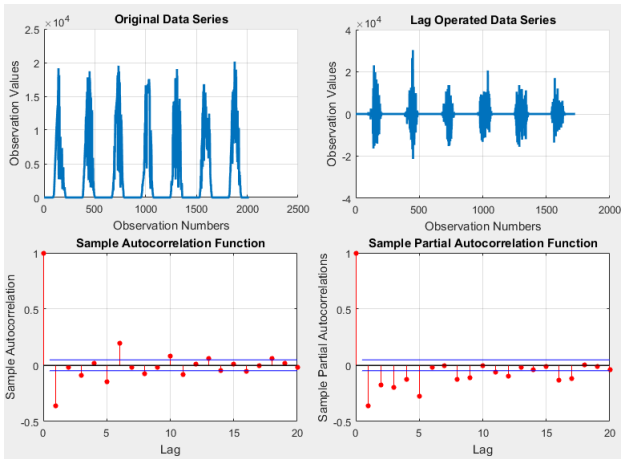


Fig. 12. Diamond Solar 300 original and single-seasonal differenced data with ACF and PACF Plots

TABLE IV  
FORECAST ERROR PERCENTAGE PER FORECAST PERIOD FOR DIFFERENT ARMA MODELS

FP\Model	AR-3	MA-2	ARMA-3
<b>5 minute</b>	<b>20.1</b>	30.4	52.8
<b>15 minute</b>	<b>7.2</b>	18.3	31.5
<b>30 minute</b>	<b>6.6</b>	28.9	19.9
<b>1 hour</b>	17.6	20.2	<b>15.0</b>
<b>3 hour</b>	35.4	<b>17.7</b>	26.9
<b>6 hour</b>	27.9	<b>18.8</b>	26.6
<b>12 hour</b>	70.6	<b>22.2</b>	30.1
<b>24 hour</b>	83.4	<b>26.2</b>	35.6
<b>2 day</b>	87.6	<b>38.3</b>	43.2
<b>3 day</b>	145.3	98.7	<b>94.1</b>
<b>4 day</b>	127.7	79.5	<b>76.9</b>
<b>5 day</b>	123.2	69.1	<b>68.3</b>
<b>6 day</b>	132.7	68.3	<b>66.2</b>
<b>7 day</b>	167.9	114.2	106.2

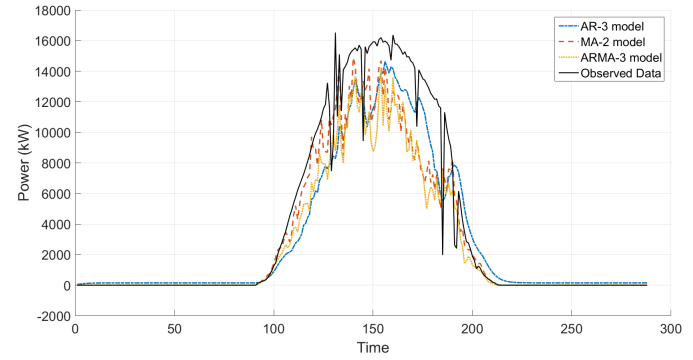


Fig. 13. Day-Ahead Power Prediction with AR-3, MA-2 and ARMA-3 Models

TABLE V  
FORECAST ERROR PERCENTAGE PER FORECAST PERIOD FOR DIFFERENT AMOUNTS OF TRAINING DATA WITH MA-2

FP\Training Data	1 Week	2 Weeks	3 Weeks	4 Weeks
<b>5 minute</b>	<b>22.7</b>	30.4	35.8	36.0
<b>15 minute</b>	<b>14.7</b>	18.3	20.2	20.3
<b>30 minute</b>	<b>26.7</b>	28.9	28.8	28.6
<b>1 hour</b>	20.0	20.2	19.7	<b>19.5</b>
<b>3 hour</b>	<b>15.7</b>	17.7	18.1	18.0
<b>6 hour</b>	<b>18.4</b>	18.8	19.4	19.3
<b>12 hour</b>	23.1	<b>22.2</b>	23.3	23.3
<b>24 hour</b>	27.3	<b>26.2</b>	27.6	27.5
<b>2 day</b>	<b>38.3</b>	38.3	39.1	39.1
<b>3 day</b>	100.1	98.7	96.1	<b>96.0</b>
<b>4 day</b>	80.5	79.5	79.0	<b>78.9</b>
<b>5 day</b>	70.0	<b>69.1</b>	70.4	70.3
<b>6 day</b>	70.4	68.3	67.1	<b>67.1</b>
<b>7 day</b>	117.7	114.2	111.2	<b>111.2</b>

TABLE VI  
FORECAST ERROR PERCENTAGE PER DAY FOR DIFFERENT TIME RESOLUTION TRAINING DATA WITH MA-2

Time Resolution	Forecast Error Percentage [1st Day]
<b>5 min</b>	26.2
<b>15 min</b>	21.6
<b>1 hour</b>	21.8
<b>2 hour</b>	21.1

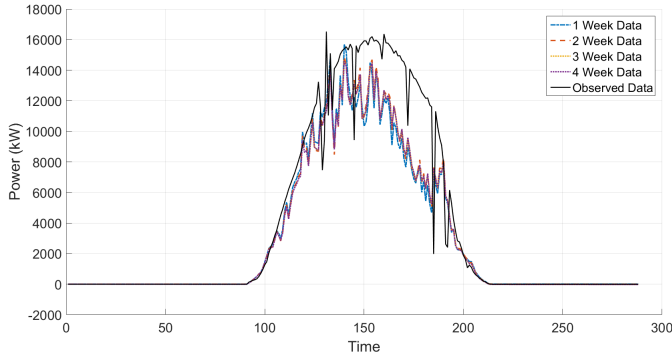


Fig. 14. Day-Ahead Power Prediction with MA-2, 2-Week Data at 5 min resolution with different amounts of training data

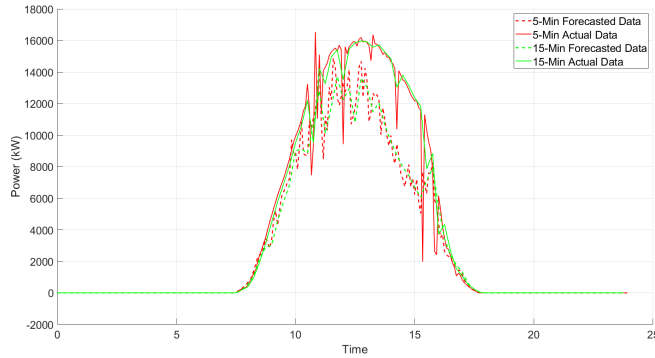


Fig. 15. Day-Ahead Power Prediction with MA-2, 2-Week Data at 5 and 15 min resolution

### C. Effect of Amount of Training Data

### D. Effect of Time Resolution of Data Series

### E. Comparison of MATLAB and Self-Developed AR and MA Models

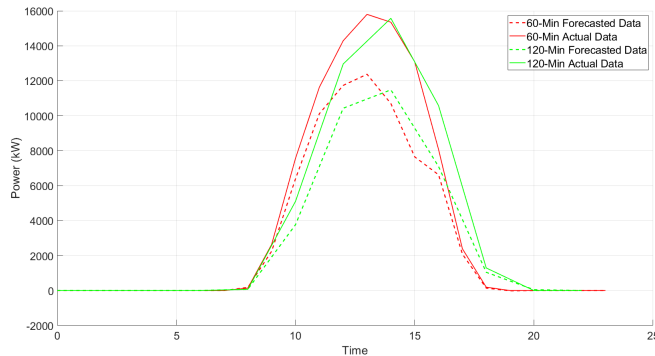


Fig. 16. Day-Ahead Power Prediction with MA-2, 2-Week Data at 60 and 120 min resolution

TABLE VII  
FORECAST ERROR PERCENTAGE PER FORECAST PERIOD FOR AR-3  
ESTIMATED WITH MATLAB, SELF-DEVELOPED LS AND MLE

FP\Models	AR-3 MATLAB	AR-3 LSE	AR-3 MLE
<b>5 minute</b>	20.1	46.9	46.9
<b>15 minute</b>	7.2	51.2	51.2
<b>30 minute</b>	6.6	47.5	47.5
<b>1 hour</b>	17.6	51.3	51.3
<b>3 hour</b>	35.4	50.8	50.8
<b>6 hour</b>	27.9	38.2	38.2
<b>12 hour</b>	70.6	143.4	143.4
<b>24 hour</b>	83.4	169.2	169.2
<b>2 day</b>	87.6	164.0	164.0
<b>3 day</b>	145.3	205.5	205.5
<b>4 day</b>	127.7	198.6	198.6
<b>5 day</b>	123.2	203.0	203.0
<b>6 day</b>	132.7	217.2	217.2
<b>7 day</b>	167.9	237.4	237.4

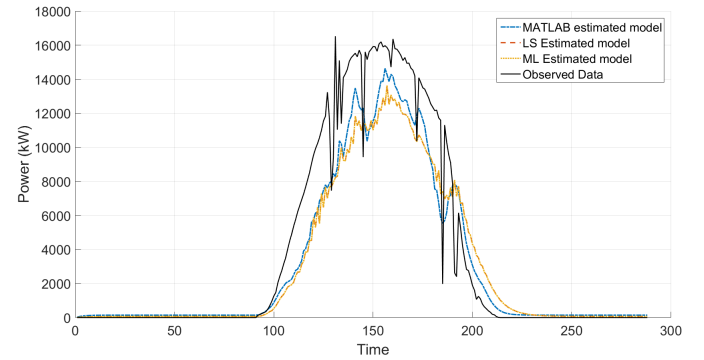


Fig. 17. Day-Ahead Power Prediction with AR-3 estimated model, 2-Week Data at 5 min resolution with MATLAB, self-developed LS and MLE

TABLE VIII  
FORECAST ERROR PERCENTAGE PER FORECAST PERIOD FOR MA-2  
ESTIMATED WITH MATLAB, SELF-DEVELOPED LS AND MLE

FP\Models	MA-2 MATLAB	MA-2 LSE	MA-2 MLE
<b>5 minute</b>	30.4	3.0	3.0
<b>15 minute</b>	18.3	7.3	7.9
<b>30 minute</b>	28.9	11.3	17.6
<b>1 hour</b>	20.2	16.4	23.6
<b>3 hour</b>	17.7	19.9	23.4
<b>6 hour</b>	18.8	20.6	25.6
<b>12 hour</b>	22.2	27.9	30.5
<b>24 hour</b>	26.2	32.9	36.0
<b>2 day</b>	38.3	48.3	52.6
<b>3 day</b>	98.7	123.5	128.0
<b>4 day</b>	79.5	99.9	105.0
<b>5 day</b>	69.1	85.7	91.5
<b>6 day</b>	68.3	85.1	90.6
<b>7 day</b>	114.2	141.9	146.9



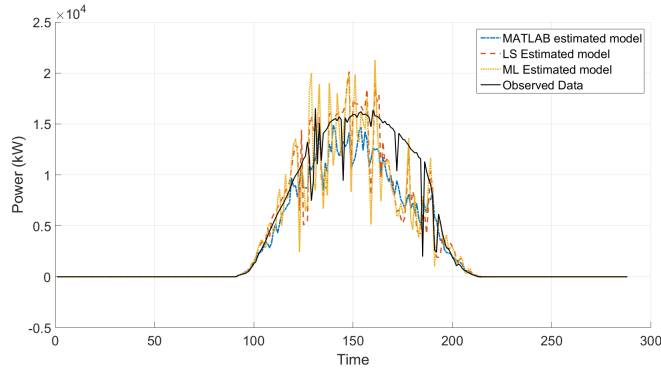


Fig. 18. Day-Ahead Power Prediction with MA-2 estimated model, 2-Week Data at 5 min resolution with MATLAB, self-developed LS and MLE

## VII. CONCLUSION

We have successfully created all the code to preprocess the data and convert it to a series spaced at an appropriate time interval. We have also used this data to create an ARMA model using the in-built MATLAB functions and find estimates for the parameters in the model. The model is able to predict the power generation to a reasonable accuracy level, though further improvements can be made. The final submission will include programs to create a model and estimate its parameters using Least Squares and Max Likelihood estimates.

## REFERENCES

- [1] P. A. Gagniuc, *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons, 2017.
- [2] W. Zhou, H. Yang, and Z. Fang, "A novel model for photovoltaic array performance prediction," *Applied energy*, vol. 84, no. 12, pp. 1187–1198, 2007.
- [3] R. Huang, T. Huang, R. Gadh, and N. Li, "Solar generation prediction using the arma model in a laboratory-level micro-grid," in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*. IEEE, 2012, pp. 528–533.
- [4] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- [5] R. Baillie, "Maximum likelihood estimation of time series models," 2017.
- [6] N. Sandgren, P. Stoica, and P. Babu, "On moving average parameter estimation," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 2348–2351.