National College of Ireland

# Mid Term Forecasting of Solar Power Generation in India: A Statistical Approach

MSc Research Project
Data Analytics

Garima Gupta
Student ID: X18182160

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | | | |
|---|---|---|---|
| **Student Name:** | Garima Gupta | | |
| **Student ID:** | X18182160 | | |
| **Programme:** | Data Analytics | **Year:** | 2019-2020 |
| **Module:** | MSc Research Project | | |
| **Supervisor:** | Dr. Catherina Mulwa | | |
| **Submission Due Date:** | 17/08/2020 | | |
| **Project Title:** | Mid Term Forecasting of Solar Power Generation in India: A Statistical Approach | | |
| **Word Count:** | 10298 **Page Count:** 28 | | |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Garima Gupta |
| **Date:** | 17/08/2020 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Mid Term Forecasting of Solar Power Generation in India: A Statistical Approach

Garima Gupta
x18182160

**Abstract**

As a developing country, the electricity demand in India is increasing rapidly, but the generation using existing resources lacks in fulfilling the demand. The primary source of electricity generation is nonrenewable resources that are environmentally unfriendly and costly. Renewable resources such as solar power, wind power, hydropower, etc could solve this problem, due to the fact of its abundant availability, cost-effective and environment-friendly nature. Installation of solar photovoltaic plants can help in minimizing this energy crisis, but the intermittency in power generation can cause fluctuations at the supply end. This issue can be addressed using an accurate forecasting system that can handle seasonalities. Various machine learning techniques ARIMA, SES, DHR, Neural Network, TBATS, and Prophet have been implemented on historical data of Rajasthan and Andhra Pradesh and the main focus of this research project was forecasting solar power generation with the help of the best model among these by evaluating the performance using evaluation metrics RMSE, MAE, and MAPE. TBATS model has given the best performance over applied models with an RMSE of 3.395737 for Rajasthan data and ARIMA with an RMSE of 1.9261 was the best performing model for Andhra Pradesh time series data.

# 1 Introduction

This section comprises the background of this research project, its importance, why it was selected, and how it is beneficial for Indian government authorities. This project looks at finding a suitable location for installing photovoltaic plants to reduce the use of non-renewable resources for power generation, reducing electricity generation costs, analyzing different techniques for forecasting solar energy generation, and providing next one year forecast.

Due to the speedy growth of economics, India has secured a position in the fastest-growing country consuming energy and it is expected that by 2030 it will become the second-largest contributor in the world's energy consumption (Jiang, Dong and Xiao, 2017). Energy is directly related to the global challenges that India faces such as poverty alleviation, food security, and environmental change. In the current situation, India is facing a threatening challenge in fulfilling the energy needs at a competitive price (Purohit, Purohit and Shekhar, 2013). Most of the countries including India uses non-renewable resources for meeting their energy needs which are costly as well as hazardous to the environment. Global warming, climate change, and increasing energy demand have created an urgent need for finding an environment-friendly, sustainable, and cheap alternative solution for energy generation (Alsharif, Younes and Kim, 2019). Renewable energy resources can be a better choice for this problem, among various non-polluting resources solar energy has gained huge attention due to the tropical climate of India.

The following section provides a clear manifestation of the importance of solar power and illustrates the motivation behind the research of its generation forecasting. Furthermore,

the questions that this project aims to answer along with the research objectives to solve this research question and a clear roadmap of the technical report is documented in this section.

## 1.1 Background

Due to the increasing population, the demand for energy keeps increasing whereas its production is not increasing at the same pace. The primary source for electricity generation is thermal and coal which releases greenhouse gases that pollute the air and cause respiratory problems. Imminent climate changes and appeal for clean and cheap energy resources have raised an interest in researchers for forecasting energy generation by renewable means (Ramachandra, Jain and Krishnadas, 2011). At present, India produces only one-third of the total required energy (Varma and Sushil, 2019) using renewable resources. Figure 1 explains the share of different sources in electricity generation capacity.

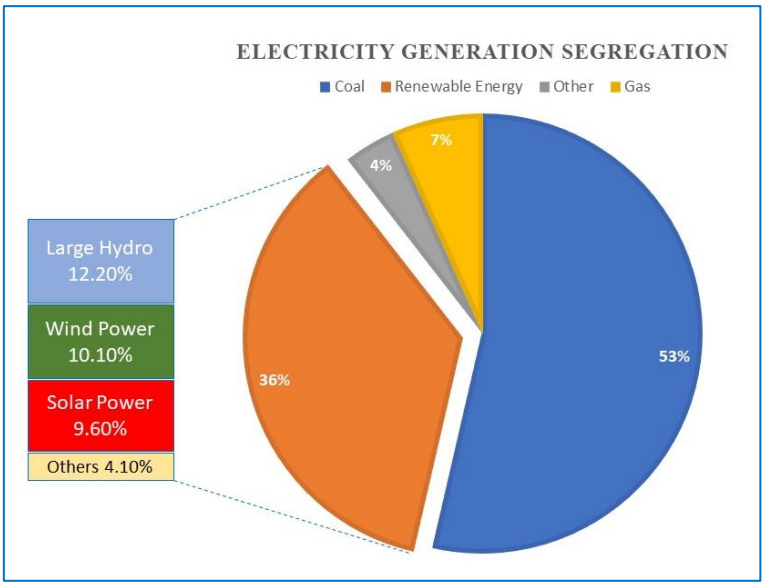


Figure 1: Sources of electricity generation in India by installed capacity

India is one of the best receivers of solar insolation due to its geographical location in the solar belt (40◦S to 40◦N), which can be used in generating electricity using this means. Despite the abundance of solar irradiance, it can be seen from figure 1 that, it contributes to only 9.6 % of electricity generation. So a lot of researches has been done in this field, but the power generation from solar energy is very erratic due to its dependency on weather, location, and seasonal changes (Prema and Rao, 2015). Due to this intermittent nature of solar irradiance, there is a need for forecasting models to schedule the power generation. In current times, the methods used for solar power forecasting are mostly Artificial Neural Networks (ANNs), Support Vector Machine (SVM) (Sharma *et al.*, 2011), and time-series models such as ARMA, ARIMA, ARMAX, SARIMA (Alsharif, Younes and Kim, 2019).

Autoregressive Integrated Moving Average (ARIMA) model has given very reliable results in forecasting solar power, but in previous studies, other models such as Dynamic Harmonic Regression(DHR), Neural Network(NN), and Simple Exponential Smoothing(SES) has also provided significant results. So in this research project neural networks, time-series models, TBATS model, along with the newly developed model prophet (novel) model have been applied to the historical data from June 2017 till June 2020 of two states of India: Rajasthan and Andhra Pradesh. The focus of this research project was on finding a machine learning model which provides improved accuracy in forecasting result and using that model to find a state which can reduce the energy crisis in India.

## 1.2 Motivation

One major motivation for forecasting solar power generation is the high cost paid by the Indian government bodies for purchasing electricity from other countries. The high demand for electricity forces the government for importing it. India has various renewable resources that can be used for generating electricity at a lesser cost. Another motivating attribute for this research project is global warming which is caused due to greenhouse gas emissions. India produces its two-third of electricity with the help of coal or natural gas which emits a huge amount of hazardous gases which cause pollution and health problems for people. Rajasthan and Andhra Pradesh have been chosen as the area of interest for this research due to the highest power requirement and potential to generate the power as these states have the highest barren or unculturable land (Ramachandra, Jain and Krishnadas, 2011) which can be utilized for installing power plants. The main question this research target to answer and objectives that need to be fulfilled to complete the research question effectively are discussed in the next section.

## 1.3 Research Question

In order to reduce the gap between requirement and production of energy in India new, more efficient, and reliable means of production should be implemented. This can also be done with the help of more precise energy generation forecasts of easily available, environment-friendly energy resources. The questions of this research had help in finding a better forecast technique for addressing this issue.

**RQ:** *"To what extent can time series forecasting on historical data of Indian states (Rajasthan and Andhra Pradesh) using forecasting models((Prophet, ARIMA, SES, DHR, NN, TBATS) help government bodies for photovoltaic plant installation to reduce the energy deficit of the country?*

**Sub RQ:** *"Can the state with the maximum anticipated capacity help in increasing the solar power generation of India?"*

This research has compared the performance of business time series forecasting model Prophet with other time series models ARIMA, SES, DHR, NN, TBATS on historical data (2017 -2020) of Rajasthan and Andhra Pradesh and calculated the forecasted solar energy of state with maximum capacity using the best performing model.

## 1.4 Research Objectives

This project is implemented using machine learning in R programming language on historical data of the last three years of two Indian states Rajasthan and Andhra Pradesh. Different time series models were applied to this data and the model with the best performance was used to forecast the power generation of these states. The power generation forecasts have been visualized using Tableau software (Wade and Nicholson, 2010) and can be presented to government authorities for helping them in deciding the photovoltaic plant installation in the most productive city.

This project is subdivided into several research objectives discussed in Table 1 that were implemented and analyzed thoroughly to answer the research questions of this project.

Table 1: Summary of Research objectives

| Objective No | Objective | Description | Evaluation method |
|---|---|---|---|
| 1. | Literature review | Identify and critically review the literature of factors affecting Solar irradiance and forecasting Solar power generation. | Critical review of literature of solar energy field |
| 2. | Data precocessing and EDA | Collect the daily data of Rajasthan and Andhra Pradesh from the CPCB website and prepare it for modelling by performing Exploratory data analysis like finding the correlation between factors using scatterplots, and removing Null. | Correlation coefficient and Exploratory data analysis |
| 3. | Implementation of Models | Implement a range of time series models for forecasting solar irradiance. | RMSE, MAE, MAPE |
| 3.1 | ARIMA | Modelling the time series data using ARIMA in Rstudio. | |
| 3.2 | SES | Applying simple exponential smoothing on preprocessed data. | |
| 3.3 | DHR | Dynamic Harmonic Regression Model on historical data of both states. | |
| 3.4 | TBATS | Modelling of TBATS machine learning technique on prepared data. | |
| 3.5 | Neural Network | Neural network model application over preprocessed data. | |
| 3.6 | Prophet | Implementing business time series of Facebook for forecasting solar irradiance. | |
| 4. | Evaluation | Evaluating and comparing the performance of implemented models. | |
| 5. | Result comparison | Comparing the performance of implemented models with existing models | RMSE |
| 6. | Energy forecast | Selecting the preferred model to find the forecasted energy generation. | |

## 1.5 Contribution

The implementation of the Prophet model, a newly designed forecasting model developed by Facebook, in forecasting solar power generation was the major contribution of this research project in the area of knowledge. This will ensure the addition of a reliable, automated, and efficient forecasting model in the field of solar power forecasting. The minor contribution of this project is the forecasting of solar irradiance in Rajasthan and Andhra Pradesh using meteorological data to help government authorities in taking decisions about energy management.

The rest of this technical report is organized in a manner where the next section provides similar sorts of researches that have been done in the area of solar power forecasting and different machine learning techniques associated with forecasting. The third section illustrates the methodology, design of this research, and the data gathering process. The fourth segment explains the implementation process of research which is followed by section five that cites the performance and results of implemented models. The sixth section discusses the overall

outcome of the research project and the last section concludes the state which has more potential to generate solar power along with the best forecasting model for the used dataset.

# 2 Literature Review of Solar Power Forecasting (2009-2019)

## 2.1 Introduction

In this section, the literature related to time series forecasting of solar irradiance has been reviewed. The various aspects considered for the literature review in this research were parameter consideration for forecasting and time series forecasting models. The area of this research was cognitive systems and prediction (due to the prediction of solar irradiance as outcome). The research papers of the last 10 years(2009-2019) were taken into consideration and were critically reviewed. This section aimed to fulfil the research objective 1 of Table 1.

## 2.2 A critique of Parameter Selection for Solar Irradiance

In research by (Heidari Kapourchali, Sepehry and Aravinthan, 2019) temperature, pressure, humidity, and wind speed were taken into consideration for short term multivariate forecasting of solar irradiance.

(Amarasinghe, 2019) has also considered weather parameters such as cloud cover, solar irradiance, and wind speed as significant variables for solar power generation forecasting. Three indices of sky cover named Total Sky Imager (TSI) Infrared Radiation (IR), Global Horizontal Irradiance (GHI) were used by researchers (Marquez, Gueorguiev and Coimbra, 2011) for forecasting global horizontal irradiance.

(Tiwari, Sabzchgar and Rasouli, 2018) has implemented a numerical weather prediction model using temperature, relative humidity, pressure, perceptible water, wind speed, dew point, cloud type, and Global Horizontal Irradiance (GHI). The result obtained from this research was good enough to consider.

(Gensler *et al.*, 2017) has used temperature, the direct solar radiation, and diffuse solar radiation to forecast the solar irradiance of Germany and achieved very accurate forecast results with a root mean square error of 0.0713. (Premalatha and Valan Arasu, 2016) have examined the monthly solar irradiance using latitude, longitude, altitude, year, month, mean ambient air temperature, mean station, level pressure, mean wind speed, and mean relative humidity using artificial neural networks with various algorithms.

By critically analyzing the considered research papers, it can be concluded that bar pressure, Relative humidity, wind speed, solar irradiance are the most significant variables in the forecasting of solar power. Hence these variables were considered while executing this research project.

## 2.3 Critical Review of Statistical Models of Solar Power

(Li and Niu, 2009) has used Markov chain model for power generation forecasting on the Beijing data of one sunny day using Markov chain theory. The model was implemented without the use of weather parameters, only past values are enough for forecasting. Markov chain has a noticeable feature of "no memory", as power generation is growing and fluctuating based on previous years, so this property can be used in the analysis. Even without the use of meteorological data, this model was able to provide acceptable forecast value. The accuracy of forecasts depends upon the truthfulness of raw data.

There can be missing values in solar time series data due to communication failure, and invisible solar sites during recording. (Heidari Kapourchali, Sepehry and Aravinthan, 2019)

have discussed a Design of Experiments (DOE) approach for handling this situation and did 6 hour ahead forecast using hourly data of kanas from the NREL meteorological database. The dataset was of one year and split into 4 parts for training the model. The performance of the model was calculated using the normalized root mean square error. The model was able to perform accurately with an NRMSE ranging from 0.05 to 0.1.

The researchers (Alsharif, Younes and Kim, 2019) have studied the hourly data of 37 years of Seoul, South Korea, and forecasted daily and monthly solar radiation using the SARIMA model. ARIMA(1,1,2) and SARIMA (4,1,1) were used as a forecasting model for daily and monthly solar radiation respectively. The performance of the model was evaluated using RMSE and $R^{2.}$ For both monthly and daily models.

(Amarasinghe, 2019) has conducted a study on the Buruthakanda solar farm of Sri Lanka using Artificial neural networks along with back propagation algorithm and compared its performance with the smart persistence method. The datasets used include the minutely, hourly, and daily data of three years 2011-2013. Normalization was performed on the dataset, divided into training-testing with a 70-30 ratio and ANN was implemented using Relu, Tanh, and sigmoid function.

Forecasting of solar irradiance with the help of three sky cover indices has been analyzed by (Marquez, Gueorguiev and Coimbra, 2011). In this study data of 3 months was considered and ANN model was applied using the different combinations of sky cover indices and it was observed that the performance of model including all indices have given the best results as compared to other implemented model on this dataset.

(Dewangan, Singh and Chakrabarti, 2018) has employed the Wavelet Neural Network with Levenberg-Marquardt (LM) algorithm for forecasting solar energy using solar irradiance data for three years of Chicago and Sandiego. The wavelet basis was used as an activation function in this study. The model was implemented using Matlab 2014. The performance was evaluated using MAPE, RMSE, correlation coefficient (r), and standard deviation (σ). Its performance was compared with conventional sigmoid Neural Network (SNN) and it was observed that WNN outperformed sigmoid neural network.

(Tiwari, Sabzchgar and Rasouli, 2018) has applied numerical weather prediction technique using Gradient Boost regression using one-year hourly data of San Diego, U.S. This approach was executed in three stages: data processing, algo 1, and algo 2. In the data preprocessing step, missing values were replaced by linear interpolation, in algo 1 sampling and bagged trees algorithm was used to reduce variance, and in final step boosting gradient was applied on the output of algo 1 step. MSE, RMSE, MAP, MAPE was used as evaluation metrics in this study.

(Gensler *et al.*, 2017) have forecasted solar irradiance data using 990 days of data of Germany using deep learning techniques autoencoder and long short term memory neural networks (LSTM). Before modelling the data was converted in the scale of 0 and 1 by normalization function, data was split into a 75-25 ratio. Relu activation function was used for modelling and the performance of the model was evaluated using RMSE, MAE, and absolute deviation.

(Premalatha and Valan Arasu, 2016) have implemented artificial neural networks using different algorithms on ten-year data of four Indian cities Chennai, Kolkatta, New Delhi, Bangalore, and Mumbai to forecast the monthly average solar radiation. The data was normalized within the range of -1 and +1, split in training testing set of 80-20 ratio, and then a 3 layer feed-forward network with tangent sigmoid activation in the hidden layer and linear activation function in the output layer was implemented using Matlab software. The number of neurons in the hidden layer was chosen based on low MSE and high correlation coefficient.

After careful consideration of all the techniques applied for solar irradiance forecasting, it can be inferred that the neural network was the most commonly used technique, along with time series forecasting models, and the most common training testing split ratio was either 75-25 or 80-20. So in this research project, two training testing sets 75-25 or 80-20 were considered, and the performance of neural network and time series forecasting models were evaluated using RMSE, MAE, and MAPE evaluation metrics.

## 2.4 Performance of Existing Models

It can be observed from the studies discussed in section 2.1, that time series models and different types of neural networks have given accurate forecasting results. This section will discuss the performance of existing models using root mean square error metric. Table 2 shows the detailed explanation of models' performance.

Table 2: Performance of Existing Models

| Model | RMSE | Data Duration | Reference |
|---|---|---|---|
| ARIMA | 33.18 | 37 years | (Alsharif, Younes and Kim, 2019) |
| Artificial Neural Network | 0.035-0.036 | 3 years | (Amarasinghe, 2019) |
| Artificial Neural Network | 27.9 | 3 months | (Marquez, Gueorguiev and Coimbra, 2011) |
| Sigmoid Neural Network | 9.11 | 3 years | (Dewangan, Singh and Chakrabarti, 2018) |
| Wavelet Neural Network | 6.68 | 3 years | (Dewangan, Singh and Chakrabarti, 2018) |
| Numerical Weather Prediction | 38.7 | 1 year | (Tiwari, Sabzchgar and Rasouli, 2018) |
| Auto-LSTM | 0.0713 | 990 days | (Gensler *et al.*, 2017) |

## 2.5 Conclusion

In this chapter, various research papers associated with the implementation techniques, methodologies, and parameters selection of solar irradiance forecasting were critically analyzed. The performance of statistical models ARIMA, ARMA, SES, and SARIMA as well as machine learning models SVM, ANN, and Random forest on time series data was also examined. This critical review of the literature of this solar energy field has fulfilled objective 1 of Table 1. In the next section, solar power forecasting methodology and design specification has been discussed.

# 3 Solar Power Methodology and Design Specification

## 3.1 Introduction

Successful implementation of any research project requires a standard and conventional research methodology that can be easily interpreted and accepted by industries and stakeholders. In this project, Cross-Industry Process for Data Mining (CRISP-DM) research

methodology suggested by (Shearer et al., 2000) has been followed. Similar to the architecture of the CRISP-DM, this project has also been executed in the form of six different phases: Business understanding, Data understanding, Data preparation, Data Modelling, Evaluation, and Deployment.

A detailed description of each phase of the used methodology has been provided in the following report. The main focus of this research is to bring insights of renewable energy generation forecasting into the Indian industry so that more efficient decisions can be made by stakeholders and government authorities. A three-tier design approach showing the project flow has also been discussed in this section.

## 3.2 Solar Power Methodology Approach

This research project aimed to forecast the solar power of Rajasthan and Andhra Pradesh that can be generated from the historical data of solar irradiance of these states. These two states have been selected for analysis because of the highest solar radiation received throughout the year as well as the capacity of generating huge energy due to available unculturable land (Ramachandra, Jain and Krishnadas, 2011). This project will help the Indian government in meeting energy requirements by installing photovoltaic plants in the more capable state.

The segments of solar power generation forecasting methodology have been explained in detail in Figure 2.

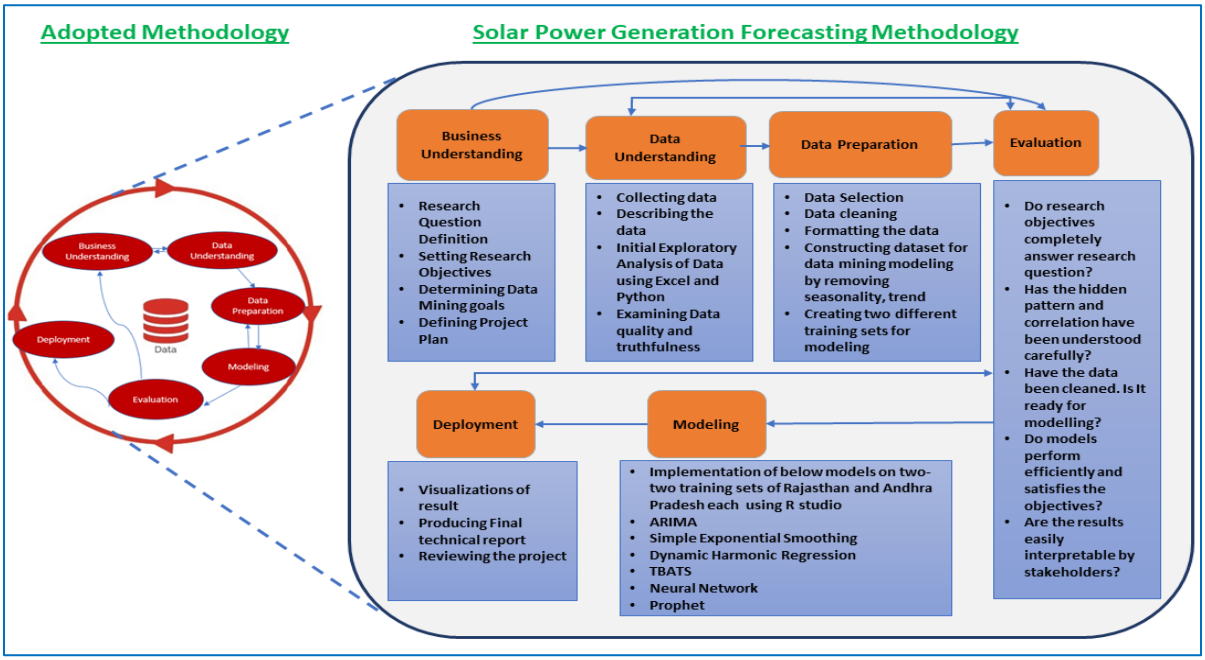


Figure 2: Solar Power Generation Forecasting Methodology

## 3.3 Data Gathering

It has been seen in various pieces of literature of the solar power forecasting field that solar irradiance is the most significant variable for its generation, so its historical data can be used as a good source to forecast the capability of a state to generate solar power. With the help of this knowledge, historical data of solar radiation along with other environmental factors

measured hourly at Jaipur and Vishakhapatnam station of Rajasthan and Andhra Pradesh respectively has been collected from Central Pollution Control Board (CPCB)[1].

To utilize the forecasting results more efficiently in business and industries, a midterm forecasting will be more suitable, short term forecasting results will not be able to help government bodies in deciding where to install solar plants for reducing energy crisis in India. Midterm forecasting implies a daily forecast of solar power, so daily data will suffice in this analysis. Hence, hourly data were converted into daily data by aggregating the hourly data of 24 hours.

For maintaining the consistency of datasets, data from June 2017 till June 2020 has been chosen for both states and two different datasets were downloaded in .xlsx format. There were various columns named relative humidity (RH), bar pressure (BP), wind speed (WS) along with information on solar irradiance (SR). For a few records, there is no value in the dataset, some have negative values and some has out of range value. These errors are due to instrumental errors caused by automatic monitoring machines used for recording meteorological parameters. For better forecasting results datasets were cleaned and preprocessed using R and Python. The detailed process of data cleaning and preprocessing is explained in further sections.

## 3.4 Exploratory Analysis and Data Preparation

### 3.4.1 Exploratory Data Analysis

**Feature Engineering:** The unprocessed data collected from the source contains a column named "To Date" which contains 24 rows for each date, this column was converted to a new column named Date by applying feature engineering. This new column contains only one record per date and the solar irradiance column contains the average value of solar irradiance of all 24 records. After this process, the datasets containing 26304 records were left with only 1096 records. The cleaning process of datasets such as the removal of missing values, incorrect data, outliers, and the trend is discussed in the following section.

The exploratory analysis was conducted on raw data of both states collected from the central pollution control board using Jupyter notebook. Datasets were imported to Jupyter using python library pandas, this library offers excellent data manipulation and analysis functions for the data frame. The environmental factors pressure (BP), humidity (RH), wind speed (WS) affects solar irradiance significantly, so these factors were considered for analysis.

In statistics, scatterplots are the method of finding the relationship of one factor with another by plotting the data points in the vertical and horizontal axis. Scatterplot of all variables included in the research with respect to solar irradiance was drawn using matplotlib.pyplot library and it was found that there is a very weak relation among these variables. Figure 3 explains the scatterplot of Rajasthan data whereas Figure 4 discusses the same for Andhra Pradesh.

[1] https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data/%7B%22state%22:%22Rajasthan%22,%22city%22:%22Jaipur%22,%22station%22:%22site_134%22%7D

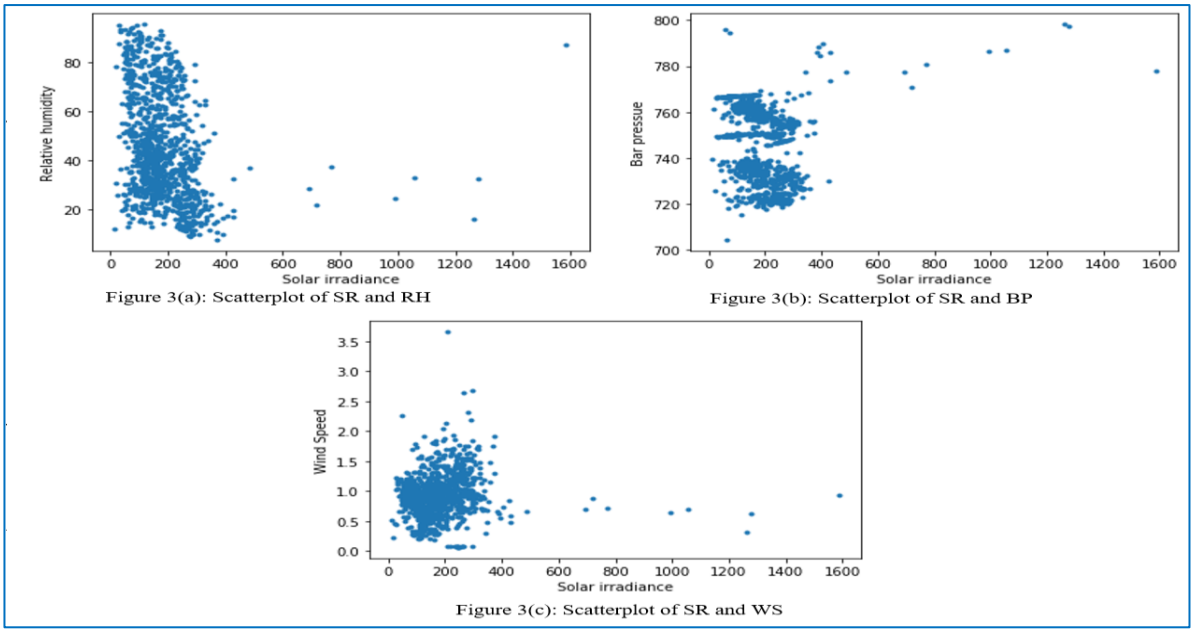


Figure 3: Scatterplots of variables of Rajasthan dataset

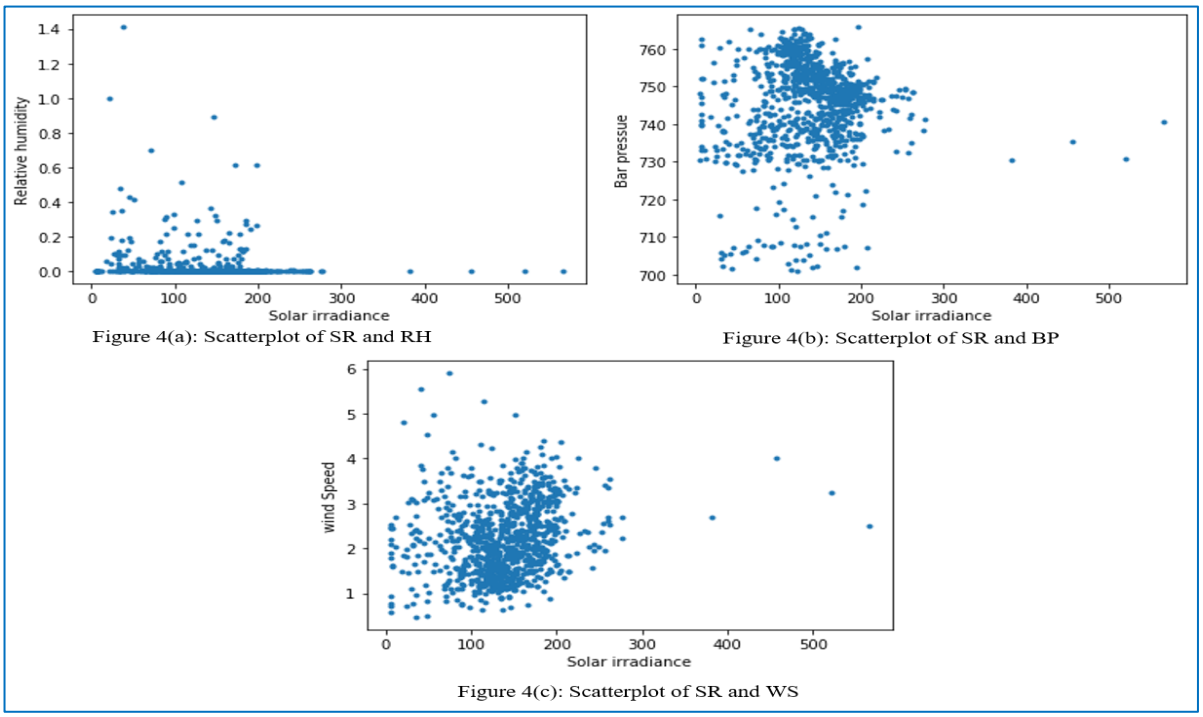


Figure 4: Scatterplots of variables of Andhra Pradesh dataset

Heatmap showing the correlation between all variables has also been analyzed using corr() function and plot has been created using the seaborn library. Figure 5 and 6 shows the heatmap of Rajasthan and Andhra Pradesh datasets. It can also be observed from both heatmaps that there is a very weak correlation among variables.

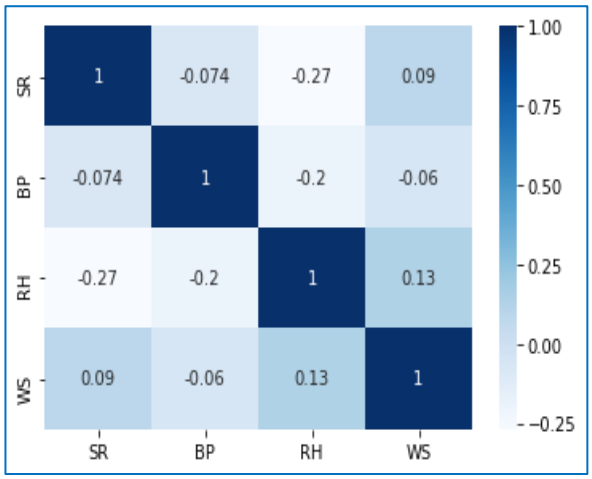


Figure 6: Heatmap of Rajasthan Data

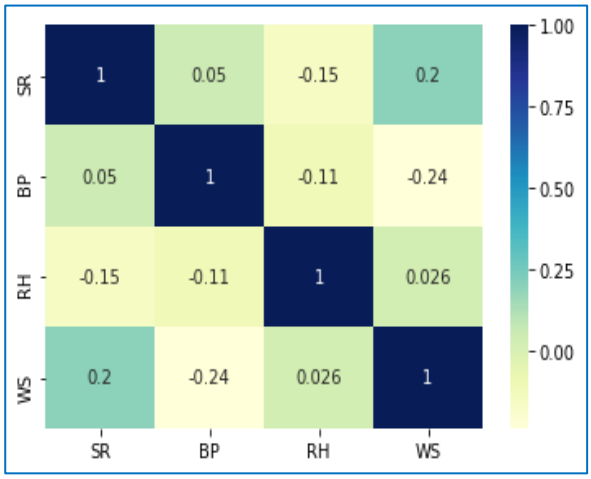


Figure 5: Heatmap of Andhra Pradesh Data

### 3.4.2 Data Cleaning and Preprocessing

Both datasets were loaded to Rstudio using xlsx package(Dragulescu and Arendt, 2020), as the datasets were of the same time duration, both xlsx files were initially containing 26304 records. Other variables of the datasets like wind speed, temperature, relative humidity, bar pressure were analyzed using scatterplot in python and it was found that there was a weak correlation between these parameters and solar irradiance, hence only solar irradiance was considered for analysis. This forecasting process is known as univariate time series forecasting[2] due to the involvement of a single variable in the whole process.

These 26304 hourly records were converted to daily data, so 1096 records were used for further analysis. Out of these 1096 records, 12 records were having negative values that were imputed by locally smoothed values using the tsclean() function of package tseries (Hornik and Trapletti, 2019). Andhra Pradesh and Rajasthan datasets had 52 and 32 null values respectively and few outliers which were also replaced by locally smooth values using the same function tsclean(). The null values, outliers, and negative values were not replaced by conventional mean, median because mean/median imputation causes loss in the uniformity of data.

Time-series was decomposed into components using decompose() function and it was observed that this data contains a huge amount of seasonality and trend. These components need to be removed from datasets for better modelling. Hence seasonal component (RJcomponents$seasonal and APcomponents$seasonal) was subtracted from the original time series datasets of Rajasthan and Andhra Pradesh respectively.

After removing the seasonality it can be seen from Figure 7 and Figure 8 that, the time series does not look seasonal anymore. The trend was also removed from both state datasets using moving average[3] function SMA() of the TTR package. After removing both seasonality and trend, various tests were performed on time series data to check the stationarity of time series, to check whether the data is white noise or contains information. Linearity check was also performed on the datasets to check whether the dataset follows normality in data or not.

[2] https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc44.htm
[3] https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html

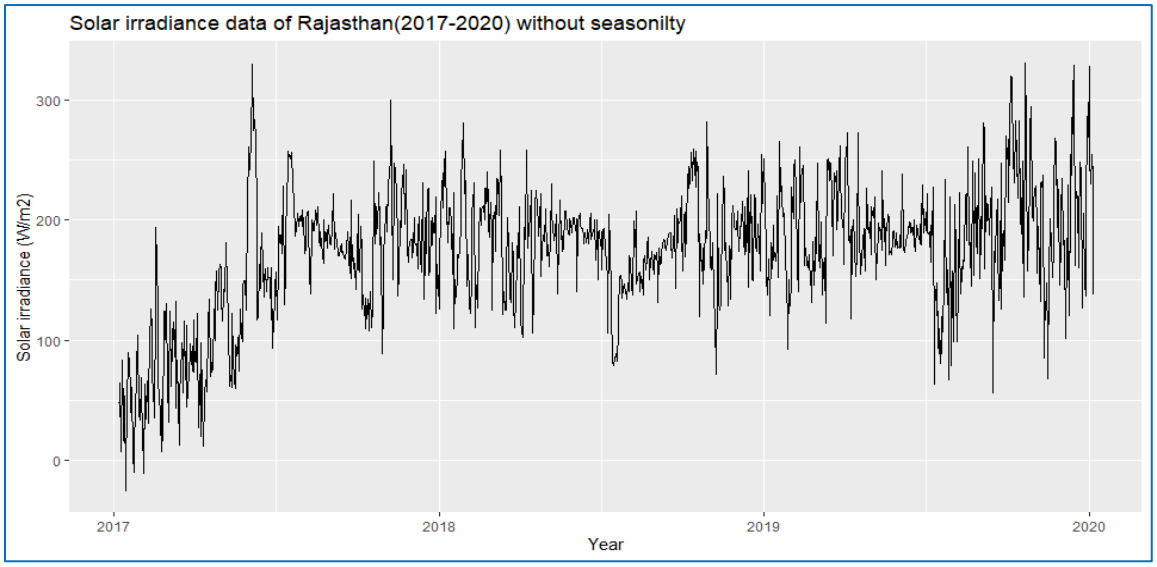


Figure 7: Solar irradiance data of Rajasthan after removing seasonality

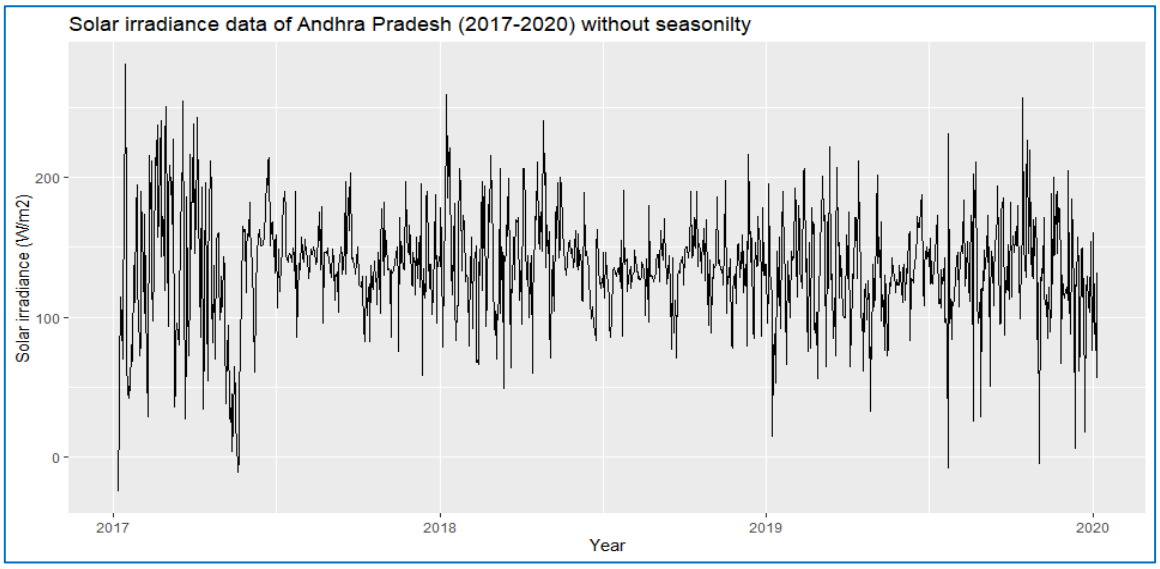


Figure 8: Solar irradiance data of Andhra Pradesh after removing seasonality

## 3.5 Data Interpretation

Augmented Dickey-Fuller test was conducted to check the stationarity of both datasets using adf.test() in Rstudio and the p-value of this test was found below 0.05, which implies that datasets are stationary [4].

The Shapiro Wilk test was performed on time series datasets to check the normality of datasets, the p-value of this test was very low which concludes that the data points are normally distributed. For being more sure the Q-Q plot shown in Figure 9 was also plotted, and it was confirmed that the data points are linear which suggests that there is no need for Box-Cox transformation. The datasets were tested against the Ljung-Box test[5] to verify the worthiness of data, as the p-value of this test was below 0.05 which gave the assurance that data is not white noise and contains valid information.

[4] https://rpubs.com/richkt/269797
[5] https://www.statisticshowto.com/ljung-box-test/

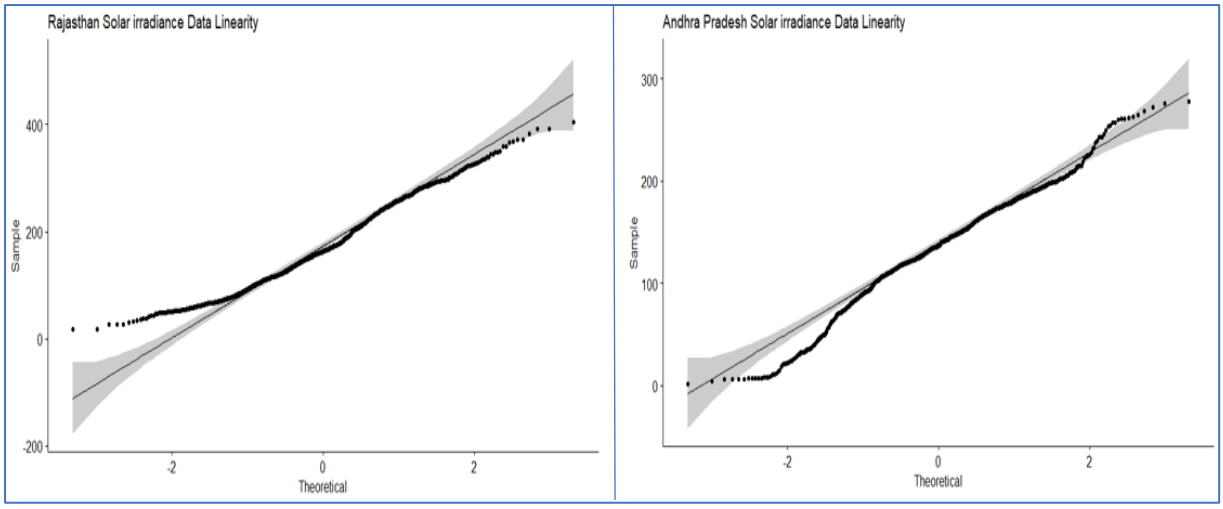


Figure 9: Q-Q Plot to check the linearity

## 3.6 Data Mining and Evaluation

Six machine learning models ARIMA, SES, DHR, TBATS, Neural Network, and Prophet were applied on datasets of Rajasthan and Andhra Pradesh.

ARIMA, DHR, and SES have been implemented by different researchers in the time series forecasting field and have given very accurate results. TBATS has been applied in areas such as electricity price forecasting and wind forecasting (Jifri, Hassan and Miswan, 2017) (Condeixa *et al.*, 2017) and provided good results over conventional time series models, but it has not been explored much, so this model is used in this research project for contributing to the area of knowledge. (Srivastava, Tiwari and Giri, 2018) have implemented Neural networks using various algorithms and forecasted solar power with less error. The prophet model has gained popularity in recent years, it was developed by Facebook for handling business records with irregular data. It has been used in cryptocurrency forecasting, but not yet used in solar power forecasting field. A detailed description of all implemented models is discussed in the next section 4.

All these models were then evaluated with the help of three evaluation assessment parameters: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). A thorough explanation of implemented models' performance using these evaluation parameters is discussed in section 4, which targets to achieve research objective 4.

## 3.7 Design Specification

To represent the understanding of this research project to other researchers, a diagram showing the project flow process is represented in Figure 10. As the dataset was gathered from monitoring systems of CPCB, then was preprocessed and converted in the format accepted by machine learning models, and at last visualized results were represented in front of the government authorities of India, the three-tier design has been followed in this research project.

All pre-processing and cleaning steps described earlier were only restricted to the Data Persistence Tier, model implementation, evaluation, and results were under Business Logic Tier, and the visualization to help stakeholders in making decisions were under Presentation Tier.

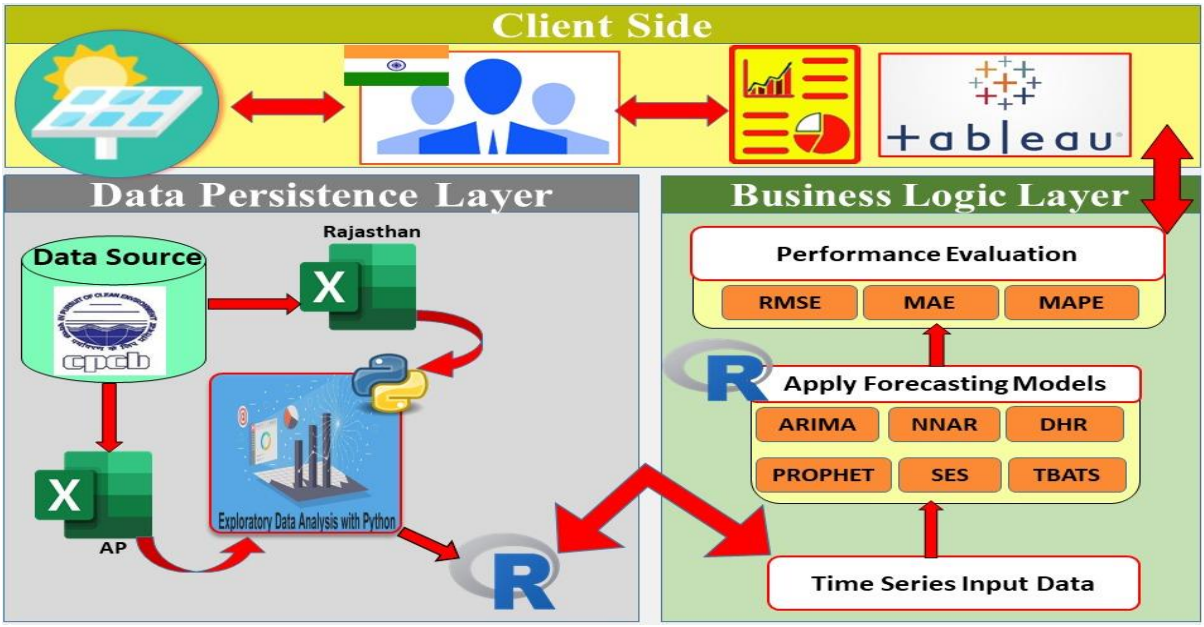


Figure 10: Architectural Project Flow

## 3.8 Tools and software

- Microsoft Excel 2019 was used for the initial pre-processing of data (Eliot, 2011) such as dropping unnecessary columns and rows and initial data type check.
- Python Jupyter Notebook was used for exploratory data analysis (EDA), such as scatterplot, heatmap. For EDA visualizations matplotlib and seaborn libraries were used.
- R programming language was used for machine learning modelling and plotting forecast results.
- Tableau 2020.2 was used as a visualization tool (Diamond and Mattia, 2017)for presenting the forecast results.

A detailed explanation of all tools and software used in this research project can be found in the configuration manual.

## 3.9 Conclusion

The CRISP-DM was selected as the preferred methodology for this research project and it was amended as per the project requirements. The datasets were obtained from the Central Pollution Control Board, data cleaning and preprocessing were performed on the datasets, so that chosen models ARIMA, SES, DHR, TBATS, Neural Networks, and Prophet can be implemented on it. In this way, the research objective 2 of Table 1 was achieved.

In the following section the implementation, evaluation, and results of applied models will be discussed that will help in achieving research objective 3 and 4.

# 4 Implementation, Evaluation, and Results of Solar Power Forecasting Models

## 4.1 Introduction

This section discusses the implementation procedure, evaluation techniques, and results of the six forecasting models selected for this research project as discussed in section 3.6. The formulas of evaluation parameters on which the performance of models was evaluated are also discussed in this section.

The preprocessed data were converted to time series data with the help of zoo variable of tseries package of R. Datasets need to be split in training and testing sets to apply any machine learning model, so dataset was converted into two sets of training and testing: one with a split of 75-25 (training set 1) and other with 80-20 (training set 2) using splitTrainTest() function from package CombMSC[6]. Various R packages (base as well as an add-on) were used during the implementation process, a list of those can be found in the configuration manual. A definition and formula of evaluation metrics discussed in section 3.6 are provided below:

**Root Mean Square Error (RMSE):** This is the measure of the deviation of predicted value from actual value and also defined as the standard deviation of residuals. This term measures the closeness of the model fitted line with actual data points. Mathematically, it can be calculated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

**Mean Absolute Error:** This term is the measure of error between the forecasted and actual value of any data point. As the name suggests, it can be calculated as the mean of the magnitude of error (ignoring the direction) and expressed as below:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

**Mean Absolute Percentage Error:** This matrix is a measure of the accuracy of any forecasting model and can be represented as the percentage of mean absolute error for the observed value as shown below:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

In all these equations, $y_i$ represents the actual value, $\hat{y}_i$ represents the forecasted/predicted value, and N as the number of observations.

## 4.2 ARIMA Model

AutoRegressive Integrated Moving Average (ARIMA) is a commonly used method used for fitting the time series data and forecasting future trends. ARIMA is a combination of both autoregression (AR) and moving average (MA) models and provides a better result than these models. ARIMA is typically represented as ARIMA (p,d,q), where p represents the autoregressive part, d stands for the number of times data needs to be differenced to make time-series stationary and q explains the moving average part of the model. ARIMA models possess the excellent ability to perform well with both stationary and non-stationary data. (Reikard, 2009), (Wan *et al.*, 2015) have used this model for predicting solar radiation and global

[6] http://finzi.psych.upenn.edu/library/CombMSC/html/splitTrainTest.html

horizontal radiation in their research and it has provided a good forecast, so this model was implemented in this research project.

### 4.2.1 Implementation

The auto.arima() function from forecast package (Hyndman and Khandakar, 2008) looks for all possible ARIMA models and selects the best model based on AICc value. AICc value is a technique to check how well mode fits data points with respect to complexity. Lower the AICc value, better the model fit.

As time-series need to be stationary for better forecasting, both datasets were tested against the Augmented Dickey-Fuller test and were found as stationary. So the parameter Stationary in auto.arima() was passed as True for both datasets. As the datasets were not so huge and searching for the best model will not cause a huge computation time, the approximation parameter was kept as off. ARIMA model was applied to both training sets (set 1 and set 2) of Rajasthan and Andhra Pradesh. ARIMA (3,0,0) and ARIMA (3,0,1) were selected for Rajasthan and Andhra Pradesh solar radiation data for both training sets (75% and 80%), and it was noticed that small change in training size did not impact the model specification. This implementation has fulfilled research objective 3.1 of Table 1.

### 4.2.2 Evaluation and Results

The implemented models ARIMA (3,0,0) and ARIMA (3,0,1) were used to forecast solar irradiance of 274 and 220 days of both Rajasthan and Andhra Pradesh. These 274 and 220 days are the testing sets which were kept aside for testing. If a model is a good fit, there should no residual left, to ensure the same checkresiduals() function was applied on forecast object and it was observed that there is no information left only white noise is left in residual.

Autoplot() function was used to plot the forecasting results of both testing sets in both states. Figure 11 and 12 shows the forecasted value for the test horizon of Rajasthan and Andhra

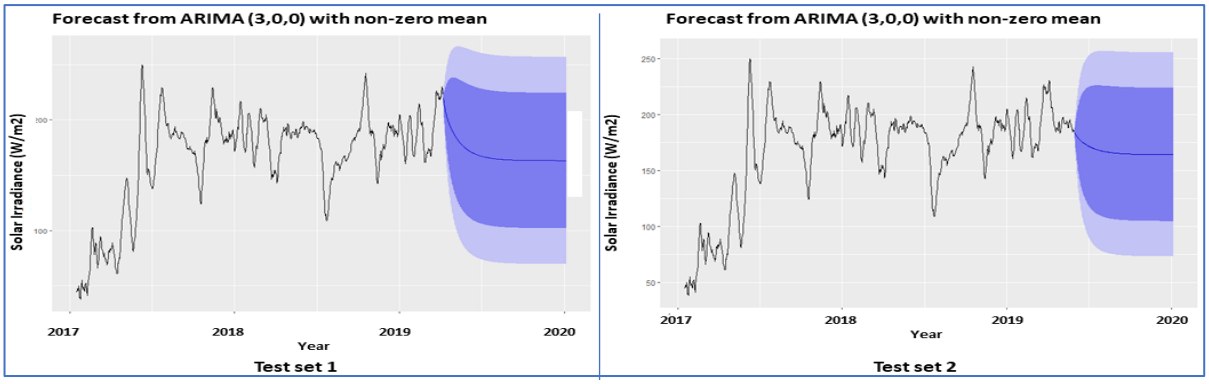


Figure 11: Rajasthan's Forecast plot using ARIMA Model

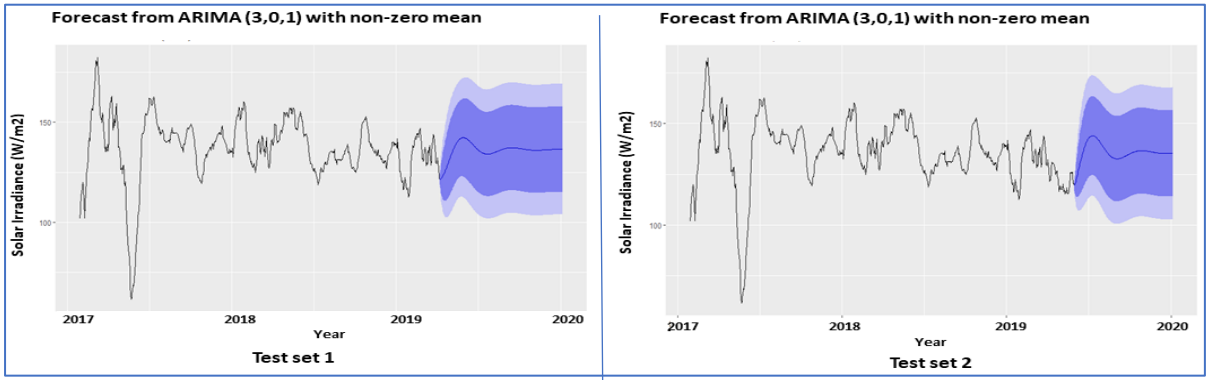


Figure 12: Andhra Pradesh's Forecast plot using ARIMA Model

Pradesh for 274 days (test set 1) and 220 days (test set 2). The blue line in plots represents the forecasted value. The model with both testing sets on both datasets was evaluated using error metrics RMSE, MAE, and MAPE, and results for the same are discussed in Table 3.

Table 3: Evaluation Metrics of ARIMA model

| Data | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|
| Testing Sets | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh |
| Test Set 1 (25%) | 3.447163 | 1.93236 | 2.658186 | 1.39816 | 1.912769 | 1.044843 |
| Test Set 2 (20%) | 3.429134 | 1.92607 | 2.64023 | 1.39542 | 1.870284 | 1.049723 |

## 4.3 Simple Exponential Smoothing Model

Simple Exponential smoothing known as SES is a univariate forecasting model that can handle only stationary data, i.e. without trend and seasonality. This time series model uses exponential window function for smoothing, a smoothing parameter(α), and gives the most weightage to last observation and less to the past observations. Mathematically it can be represented as :

$$Y_{t+1} = \alpha y_t + (1-\alpha) Y_t$$

where $Y_{t+1}$ is the forecast of next period (t+1), $y_t$ is the actual value of previous observation, $Y_t$ is forecasted value at time t.

### 4.3.1 Implementation

This model was implemented by applying ses() function from forecast package on both datasets using training set 1 and set 2. Due to the prerequisite of the SES model, the time series was already tested through adf test for stationarity in section 3.5. While training the model the initial parameter was passed as optimal, so that initial values of the forecast are optimized through ets model. Also, the data was found as normalized in the data interpretation section, so there is no need for Box-cox transformation. Hence lambda, alpha was passed as Null and biasadj as False. The R code used for implementing this model can be found in the configuration manual. The target of implementation was to fulfil research objective 3.2 which was achieved successfully.

### 4.3.2 Evaluation and Results

Simple exponential smoothing uses the method of giving the most weightage to the last observation which is the main reason for providing an accurate forecast. Plots for forecasted outcomes along with original data of both datasets with two testing sets are shown in Figure 13

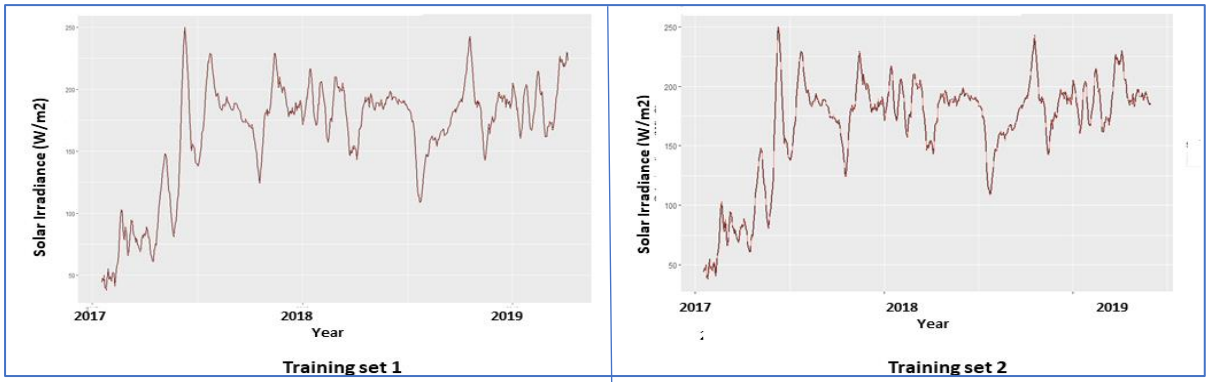


Figure 11: Rajasthan's Forecast plot using SES Model

and 14. It can be seen from the plots that forecasted value fits the actual data very closely, the difference between the fitted line and actual observation is very low in both the cases. The red line in the plots shows the predicted data points and black as the actual value.

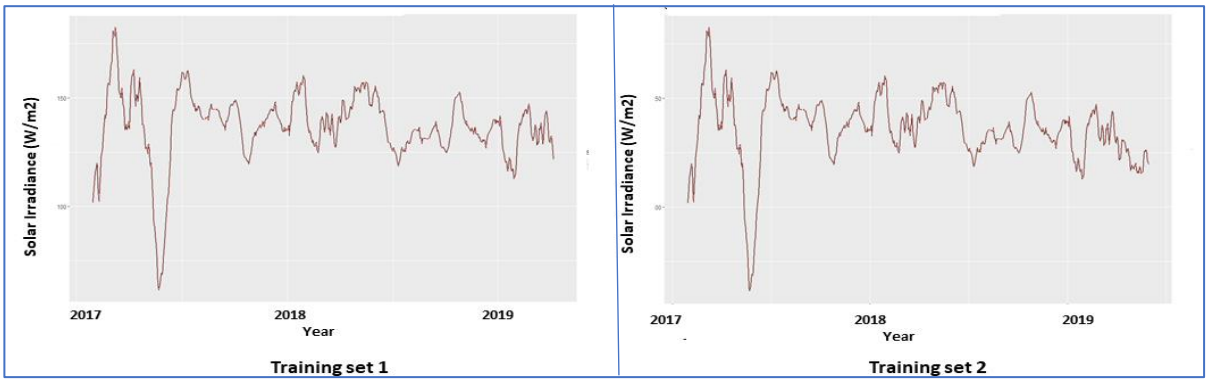


Figure 12: Andhra Pradesh's Forecast plot using SES Model

Table 4 shows the calculated values of evaluation parameters, to examine the performance of the model. It can be seen from the below table that set 2 (80-20 split) gives a better result as compared to set 1. The results are significantly good but not so accurate as the ARIMA model.

Table 4: Evaluation Metrics of SES model

| Data | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|
| Testing Sets | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh |
| Test Set 1 (25%) | 4.413052 | 2.306446 | 3.364791 | 1.668519 | 2.334902 | 1.270832 |
| Test Set 2 (20%) | 4.346164 | 2.276351 | 3.297340 | 1.642982 | 2.26291 | 1.256257 |

## 4.4 Dynamic Harmonic Regression Model

Dynamic Harmonic Regression is a forecasting model that can handle any length seasonality (Hyndman, 2018), unlike the ARIMA model which performs well with shorter length seasonal periods. DHR possesses this property due to the Fourier term included in this model. For handling time series with more than one seasonal period, Fourier terms of different frequencies need to be added while training the model. A factor K is used in this model for handling seasonality.

### 4.4.1 Implementation

This model also uses auto.arima() function of forecast package but with an added Fourier term, xreg, and a factor K to handle different seasons. Different value of K was tried to get the best model on the basis of AICc value, as discussed earlier lower the AICc, better the model. K =2 was found as the best parameter, as above and below K = 2, the AICc was higher. The model was implemented with K = 2, stationary as true, and biasadj as true, hence the research objective 3.3 of Table 1 was completed. To double-check the model fitness, residuals were checked and found as white noise. More details of this model implementation and K-value selection have been provided in the configuration manual. Research objective 3.3 was achieved successfully with the help of this implementation.

### 4.4.2 Evaluation and Results

Forecasts of DHR model using the training set 1 and 2 on datasets of both states is shown in Figure 15 and 16 respectively. These forecast plots look smoother as compared to the ARIMA model forecasts discussed in section 4.2.2 due to the extra Fourier term added in the processing.

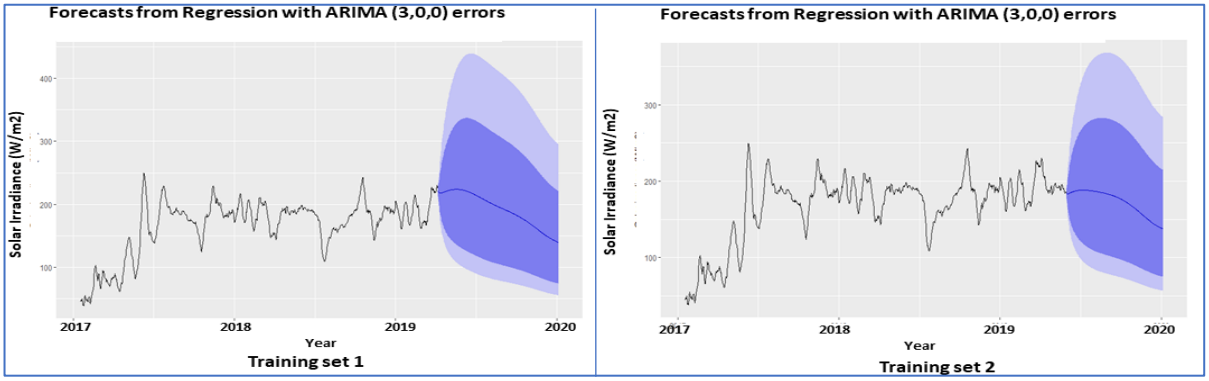


Figure 14: Rajasthan's Forecast plot using DHR Model

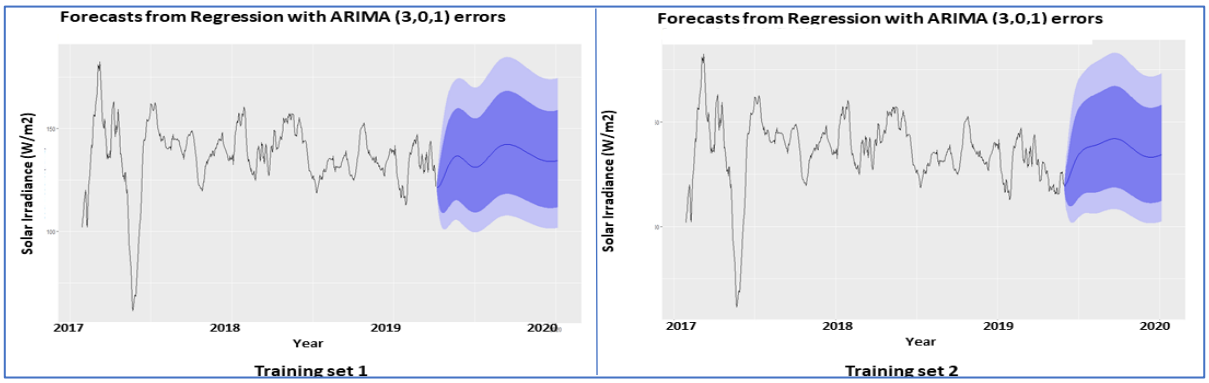


Figure 13: Andhra Pradesh's Forecast plot using DHR Model

It can also be seen from Table 5 that all evaluation parameters have nearly the same value due to the same algorithm behind DHR and ARIMA models. The same hyperparameters for the model were selected by DHR as by ARIMA in the earlier section of ARIMA implementation.

Table 5: Evaluation Metrics of DHR model

| Data | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|
| Testing Sets | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh |
| Test Set 1 (25%) | 3.465390 | 1.936291 | 2.684694 | 1.400135 | 1.905461 | 1.045884 |
| Test Set 2 (20%) | 3.443069 | 1.927582 | 2.660631 | 1.396146 | 1.861069 | 1.050017 |

## 4.5 TBATS Model

TBATS, a time series forecasting model, originated by De Livera, Hyndman & Snyder is built to handle the multiple complex seasonalities in the time series dataset. It is the acronym of the features of time series, such as T for Trignomertic, B for Box-Cox transformation, A for ARMA errors, T for trend, and S for seasonality. (Condeixa *et al.*, 2017) have used this model in wind forecasting and achieved good forecast results, as the wind also has seasonality like solar irradiance, hence this model was chosen for this project.

### 4.5.1 Implementation

This model was implemented using tbats() function from forecast package, as discussed earlier the datasets contain normality in their distribution, so the box-cox transformation was set as null and biasadj as false. The trend was removed from series in section 3.4.3, hence use. trend was also passed as null in this case. The model was applied to training sets of 822 and 876 records of both Rajasthan and Andhra Pradesh. Solar irradiance contains multiple seasonalities, so this model was a good choice for analyzing it. A detailed description along with R code has been provided in the configuration manual. In this manner, research objective 3.4 was fulfilled.

### 4.5.2 Evaluation and Results

The plots of forecasted results were plotted using autoplot() function and it can be observed from Figure 17 and 18 that there is a very minute difference between the forecasted future values of training sets of Rajasthan in set 1 and set 2, but same does not apply for Andhra Pradesh data. These results are a little contradicting, this may be due to incorrect data entries entered by the automated machine of CPCB. Due to the capability of handling complex

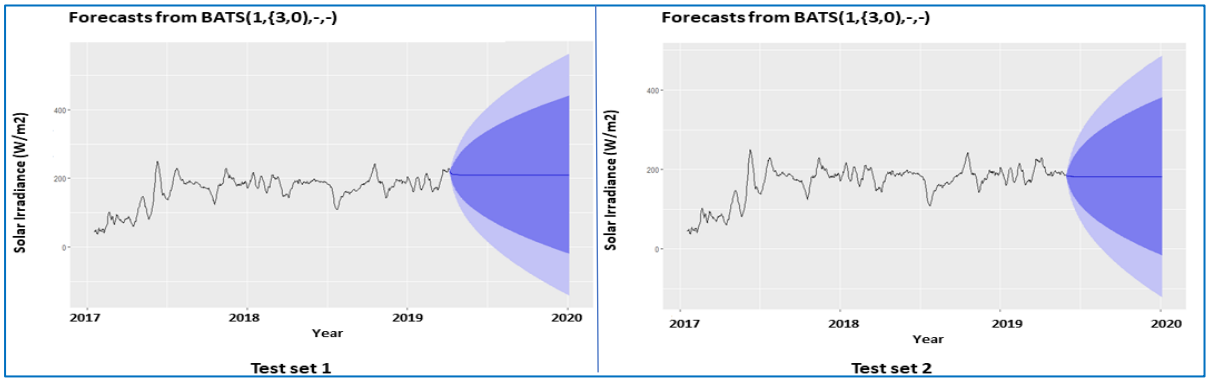


Figure 15: Rajasthan's Forecast plot using TBATS Model

multiple seasonalities, it has outperformed ARIMA model for Rajasthan's data. Also, from Andhra Pradesh's plot, it can be concluded that TBATS has given good results for a short horizon.

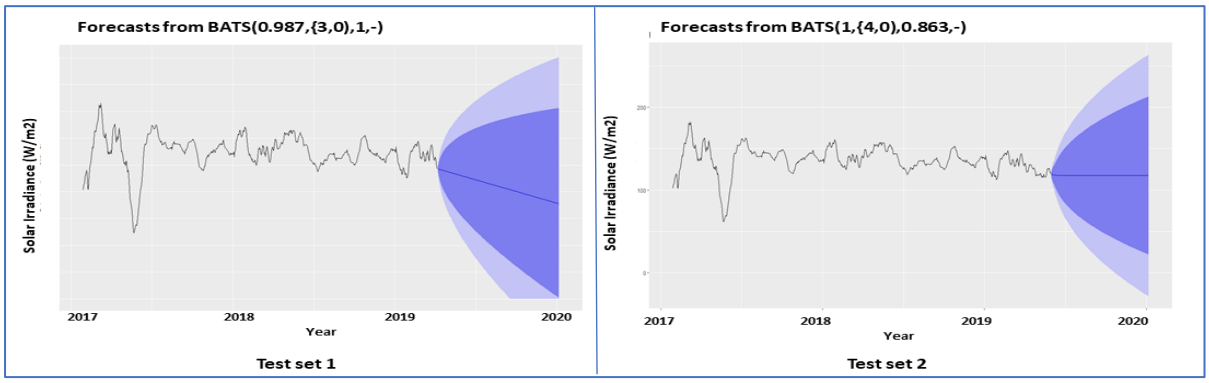


Figure 16: Andhra Pradesh's Forecast plot using TBATS Model

From Table 6 it can be noted that all the error metrics are significantly low for both states, so this model can be used further for forecasting future value.TBATS outperformed ARIMA model but the AICc was much higher, which means a little forecast accuracy has caused huge complexity. Hence TBATS was not considered better than ARIMA in this case.

Table 6: Evaluation Metrics of TBATS model

| Data | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|
| Testing Sets | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh |
| Test Set 1 (25%) | 3.415193 | 2.007695 | 2.645722 | 1.464390 | 1.879771 | 1.100587 |
| Test Set 2 (20%) | 3.395737 | 1.978833 | 2.625447 | 1.451111 | 1.837209 | 1.097205 |

## 4.6 Neural Network model

Neural Networks can be imagined as a network of neurons arranged in layers. A simple neural network consists of an input layer, a hidden layer, and an output layer. The predictors or independent variable form the input layer, forecast results form the output layer, and the hidden layer contains neurons. The neural networks have already shown its outstanding performance in classification and regression and have been giving excellent results in forecasting time series data. In time series analysis, lagged value of time series are used as inputs in the neural network, just like linear autoregression model, so the neural network is known as NNAR in time series. NNAR model is represented as NNAR(p,P,k), where p represents the number of lagged inputs, P as the number of seasons in data, and k as the number of neurons in the hidden layer.

### 4.6.1 Implementation

This model was implemented using the nnetar() function from the forecast package. Due to the normal distribution of datasets, lambda was passed as Null and scale.inputs as False. The value of p, P, and size was decided by a different combination of values, and combination giving the least AICc was selected. This process is known as hyperparameter tuning. P was selected as 2 because solar radiation data contains two major seasons. Different values for repeats were tried to choose the most accurate weights and value 100 gave the least AICc value hence the value of repeats was selected as 100. NNAR(10,2,6) and NNAR(8,2,5) model was chosen as the best NNAR model for Rajasthan and Andhra Pradesh datasets respectively. In this way, the research objective 3.5 of chapter 1 was fulfilled.

### 4.6.2 Evaluation and Results

Both models NNAR(10,2,6) and NNAR(8,2,5) were applied to testing sets of 274 and 220 days of both datasets to compare the forecasted value with the actual value. In order to understand the model performance in a better way, forecast plots were drawn using autoplot() function as shown in Figure 19 and 20. It can be observed from figures that NNAR(8,2,5) was able to forecast the values with less error as compared to NNAR(10,2,6) of Rajasthan's data. NN model on both datasets has provided almost a constant value which is the cause of less accurate results as compared to ARIMA, SES, DHR, and TBATS.

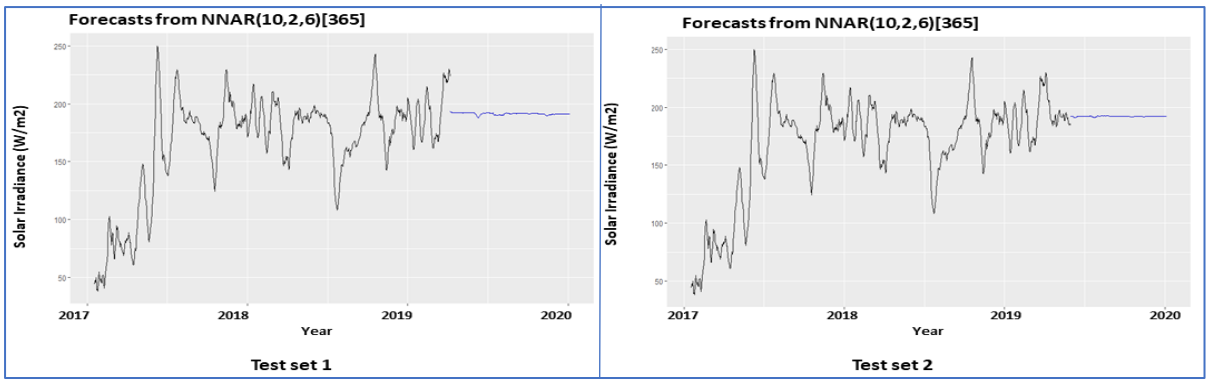


Figure 17: Rajasthan's Forecast plot using Neural Network Model

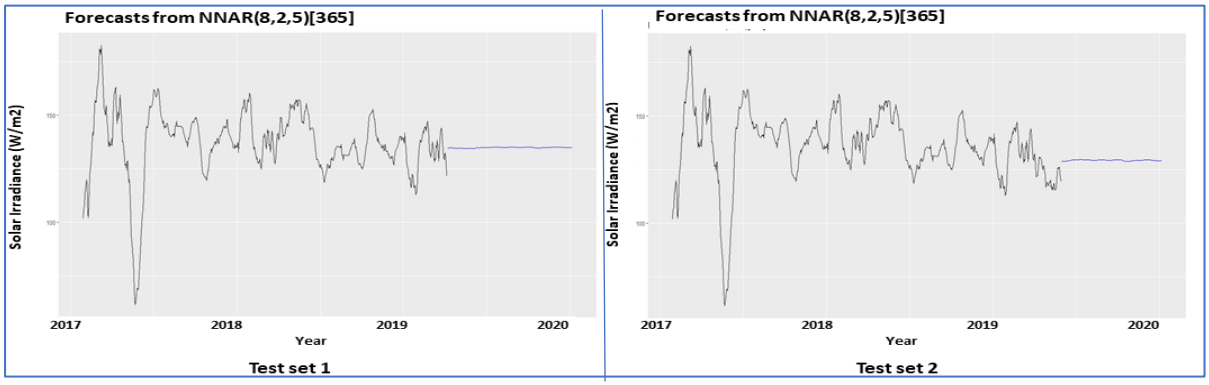


Figure 18: Andhra Pradesh's Forecast plot using Neural Network Model

Table 7 explains the error metrics of the NN model and it can be inferred that Andhra Pradesh's data has outperformed Rajasthan's performance for both testing sets.

Table 7: Evaluation Metrics of Neural Network model

| Data | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|
| Testing Sets | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh | Rajasthan | Andhra Pradesh |
| Test Set 1 (25%) | 19.81523 | 7.291134 | 17.33042 | 5.767962 | 9.063925 | 4.367968 |
| Test Set 2 (20%) | 16.38944 | 8.938446 | 12.70494 | 7.521864 | 6.622763 | 5.851798 |

## 4.7 Prophet model: Novelty of Research

PROPHET is open-source software for forecasting time series data available in both R and python. This model has strong capabilities of handling missing values and trends in data. It is based on a model that fits non-linear trends along with seasonality and holidays. The prophet model has not been explored much in forecasting area, as it was developed by Facebook in 2017, its implementation to bitcoin forecasting area(Duvodq *et al.*, 2018) has outperformed ARIMA model, so this model was considered for analysis in this research. The implementation of this model is the novelty of this research project as this model has never been applied for solar power forecasting.

### 4.7.1 Implementation

PROPHET model uses its framework for modelling, it requires a data frame with two columns named “ds” and “y” for forecasting the results more accurately by handling seasonality and trend. The columns “ds” stores the datestamp of series and “y” stores its corresponding value in time series. The time series was converted into a data frame as per requirement for modelling using ts_df() function from tsbox library. As seasonality was removed in section 3.4.2, so seasonality parameters were passed as False and changepoints as Null due to nonseasonal data while training the model to create a forecast object. This implementation of the prophet has fulfilled research objective 3.6 of Table 1.

### 4.7.2 Evaluation and Results

Forecast object obtained after model implementation was applied on test data sets 1 and 2 of both states and forecast results were plotted using dyplot.prophet() function of prophet package. The plot obtained has been shown below in Figure 21 and 22. Black dots represent the actual value of time series data whereas blue represents the predicted outcome of this model.

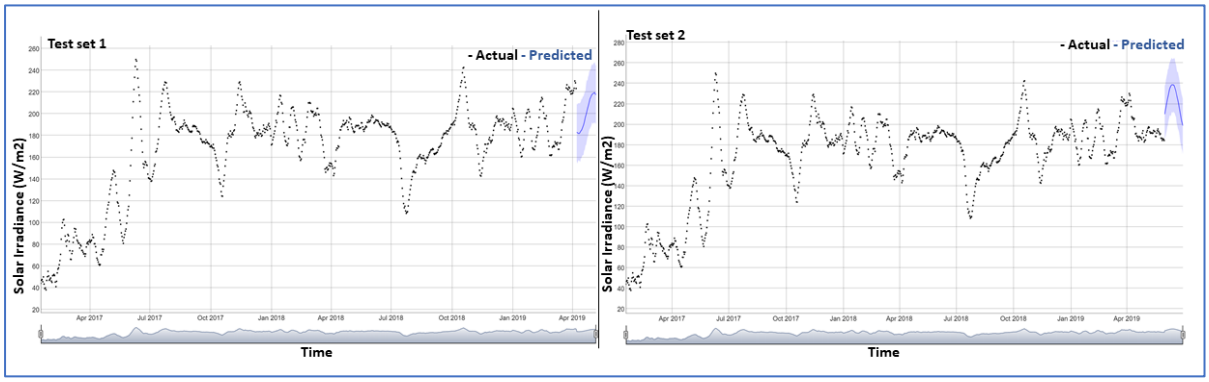


Figure 21: Rajasthan's Forecast plot using Prophet Model

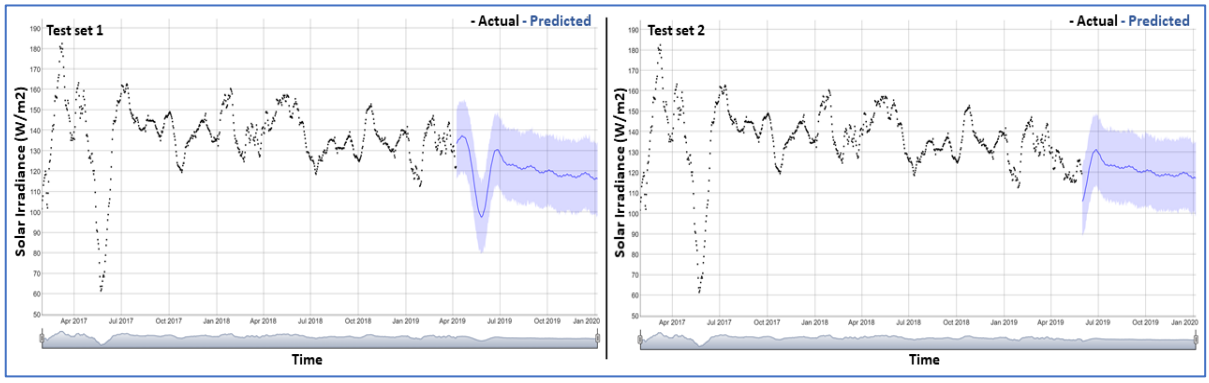


Figure 22: Andhra Pradesh's Forecast plot using Prophet Model

For evaluating the performance of this model, Metrics() package was used and RMSE, MAE, and MAPE values were calculated. RMSE of 21.23 was obtained as shown in Table 8.

Table 8: Evaluation Metrics for Prophet Model

| Data | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|
| **Testing Sets** | **Rajasthan** | **Andhra Pradesh** | **Rajasthan** | **Andhra Pradesh** | **Rajasthan** | **Andhra Pradesh** |
| **Test Set 1 (25%)** | 21.23669 | 16.95289 | 19.01291 | 13.98698 | 0.09786661 | 0.1025259 |
| **Test Set 2 (20%)** | 39.02077 | 16.73597 | 36.62748 | 13.60640 | 0.1958492 | 0.0971253 |

# 5 Discussion and Comparison of Developed Models

## 5.1 Result Comparison

**Comparison with Implemented Models:** In this research project, six implemented models were evaluated over three different evaluation metrics RMSE, MAE, MAPE. From section 4 it can be concluded that forecast results of each model using the training set 2 (80-20) are more accurate, so the performance of implemented models was compared using the results of set 2.

With the help of error metrics of these models shown in Table 9, it can be concluded that DHR gave almost the same forecasts as ARIMA for both datasets. The Prophet model was never used before in the solar forecasting area but has gained popularity in handling time series data with irregular intervals, so implementation of this model was the novelty of this research project. From the error metrics, it can be seen the performance of the prophet model was not significantly good in this field. TBATS outperformed ARIMA and all other time series models applied for only Rajasthan's data, but the complexity of the model was increased drastically with a little increase in the forecast results, hence ARIMA was considered as best performing model. Finding the best performing model by comparing error metrics fulfills the research objective 5 of Table 1.

Table 9: Performance comparison table

| Applied Models | Rajasthan | | | Andhra Pradesh | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | MAPE | RMSE | MAE | MAPE |
| ARIMA | 3.429134 | 2.64023 | 1.870284 | 1.92607 | 1.39542 | 1.049723 |
| SES | 4.346164 | 3.29734 | 2.26291 | 2.276351 | 1.642982 | 1.256257 |
| DHR | 3.443069 | 2.660631 | 1.861069 | 1.927582 | 1.396146 | 1.050017 |
| TBATS | 3.395737 | 2.625447 | 1.837209 | 1.978833 | 1.451111 | 1.097205 |
| Neural | 16.38944 | 12.70494 | 6.622763 | 8.938446 | 7.521864 | 5.851798 |
| Prophet | 21.23669 | 19.01291 | 0.097867 | 16.95289 | 13.98698 | 0.102526 |

**Comparison with Existing Models:** The performance of existing models discussed in Table 2 can be compared with the performance of implemented models of Chapter 4. It can be concluded that the implemented ARIMA model with RMSE of 3.429 ( Table 3) has performed better than the existing model with RMSE of 33.18, but the performance of the existing neural network (RMSE = 6.68) was better than the performance of the implemented neural network model (RMSE = 16.38944) discussed in Table 7. This comparison achieves objective 5 of Table 1.

## 5.2 Generation of Solar Power

As discussed in section 5.1, ARIMA was the best performing model, so this model is applied to datasets of both states to forecast solar irradiance for one year (365 days). Solar energy can be calculated with the help of solar irradiance using below mathematical expression[7]

$$E = A * r * H * PR$$

where,

E = Solar energy
A = Solar panel area
H = Solar irradiance
r = Solar panel's efficiency
PR = Coefficient of loss

[7] https://photovoltaic-software.com/principle-ressources/how-calculate-solar-energy-power-pv-systems

To answer the Sub RQ defined in section 1.3, parameters mentioned above other than solar irradiance are required. In various studies of solar power forecasting, it was noticed that the efficiency of solar panel lies between 10-15%, so in this research project, 15% efficiency (r=0.15) has been considered. The area of the solar panel has been fixed at 1.6 $m^2$, and the coefficient of loss has defaulted at 0.75. To replicate the same industrial photovoltaic power plants, the standard values of the parameters mentioned above were considered. By putting these constant parameters value along with forecasted solar irradiance, solar power generation was calculated for both states and represented using a line chart as shown in Figure 21.

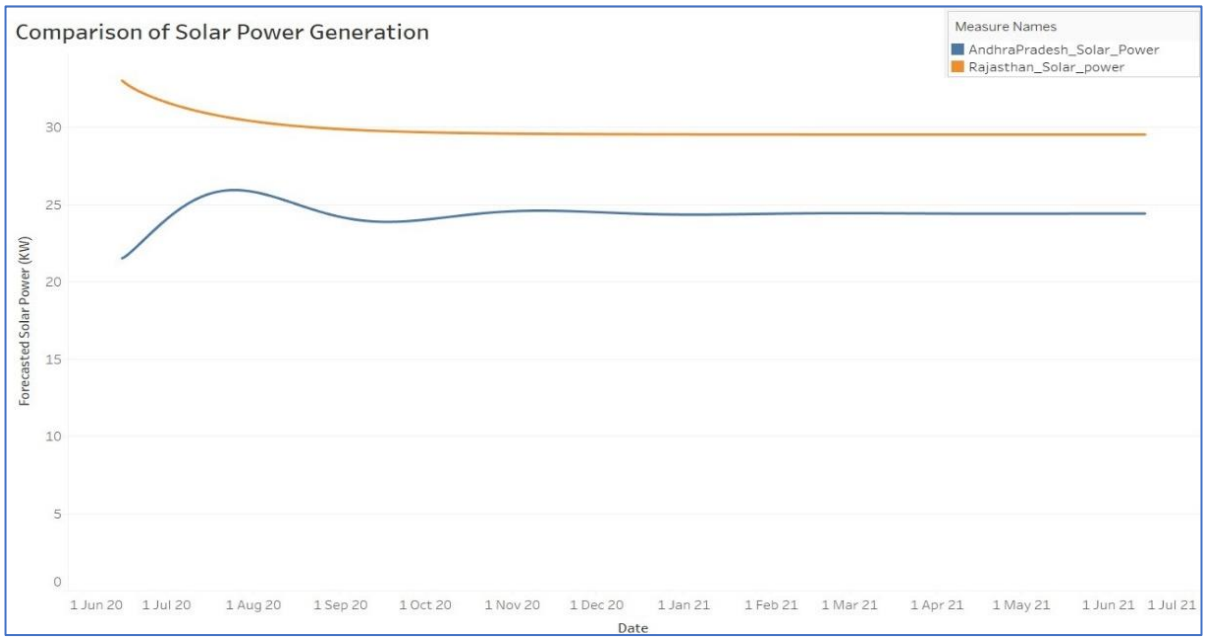


Figure 19: Forecasted Solar power generation of Rajasthan and Andhra Pradesh

It can be observed from the above chart that, Rajasthan's generated solar energy is consistently higher than the power generation of Andhra Pradesh. A single solar panel generates approx 30 KW of energy in a day, which can easily help in reducing the gap between energy production and requirement. From the graph, it is clear that Rajasthan has a higher potential for generating solar power, so Indian government authorities should invest in installing solar photovoltaic plants in Rajasthan as it can produce a higher amount of energy compared to Andhra Pradesh by investing an equal amount of money. Hence, the sub research question of this research project and research objective 6 of Table 1 was answered.

**Limitation:** The historical data of solar irradiance was getting inconsistent by going deeper into the past years, so the data 3 years was gathered from the source. Also, the data of a few cities were available at the source. Due to the limited size of the dataset, the research was not conducted on the whole state. If the historical data of at least ten years were available, then the long term forecasting could have been done with better forecasting results to help Indian authorities in reducing the energy crisis.

During this research project, a fair knowledge of python, Tableau, and deep knowledge of time series analysis in R was gained. The application of the prophet model in the solar power forecasting field was the major contribution in the area of knowledge and finding a location to install the photovoltaic plant to increase the energy generation in India was a minor contribution.

# 6 Conclusion and Future Work

The goal of this research was to find a location in India that can help the government in reducing the dependency on energy imports by installing photovoltaic plants. With the results discussed in chapter 5, it can be concluded that Rajasthan has more capacity to generate solar power. During the process of finding the location, the performance of ARIMA, SES, DHR, Neural network, TBATS, and Prophet was compared and it was found that ARIMA outperformed all other applied model. So, ARIMA model was used to forecast the solar energy generation of the next year (365 days). Thus, the research question discussed in (Chapter1, subsection 1.3) and objectives (Chapter1, subsection 1.4) has been successfully achieved.

The Prophet model did not perform well with both the datasets, but its implementation has opened some unseen corners in the forecasting domain, which has a possibility of improvement. This could have performed better if it was implemented using datasets with parameters such as holiday, seasonality, and trend. It was also seen that the TBATS model has given outstanding performance, hence it can be used in other researches where data has complex multiple seasonalities.

Due to the excellent results of research, the solar energy forecast details for the next one year can be provided to Indian government authorities to help them in deciding the installation of solar photovoltaic plants in Rajasthan which will further help in reducing the dependency on energy imports for supplying energy requirements.

In the future, the research can be enlarged by widening the variables used for solar irradiance forecasting. There are various factors that influence the solar irradiance forecasting, so multivariate forecasting techniques can be applied to forecast the solar irradiance. Also, there are three types of solar radiation: global horizontal irradiation, diffuse horizontal irradiation, and Direct normal irradiation. The forecasting of solar power by forecasting these three types of irradiation can be done in the future.

# Acknowledgment

I would like to thank my supervisor, Dr. Catherine Mulwa for her consistent support and guidance for helping me in completing the research successfully on time. She continuously encouraged and was enthusiastic to assist me in every possible way. I would also like to convey my gratitude to my parents Mr. and Mrs. Gupta for their unfailing support throughout the journey.

# References

Alsharif, M., Younes, M. and Kim, J. (2019) ‘Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea’, *Symmetry*, 11(2), p. 240. doi: 10.3390/sym11020240.

Amarasinghe, G. (2019) ‘An artificial neural network for solar power generation forecasting using weather parameters’, (January).

Condeixa, L. D. *et al.* (2017) ‘Wind speed time series analysis using TBATS decomposition and moving blocks bootstrap’, *International Journal of Energy and Statistics*, 05(02), p. 1750010. doi: 10.1142/S2335680417500107.

Dewangan, C. L., Singh, S. N. and Chakrabarti, S. (2018) ‘Solar irradiance forecasting using

wavelet neural network', *Asia-Pacific Power and Energy Engineering Conference, APPEEC*, 2017-Novem, pp. 1–6. doi: 10.1109/APPEEC.2017.8308987.

Diamond, M. and Mattia, A. (2017) 'Data Visualization: An Exploratory Study into the Software Tools Used by Businesses.', *Journal of Instructional Pedagogies*, 18, pp. 1–7. Available at: https://eric.ed.gov/?id=EJ1151731.

Dragulescu, A. and Arendt, C. (2020) 'Package " xlsx "'. Available at: https://cran.r-project.org/web/packages/xlsx/xlsx.pdf.

Duvodq, F. *et al.* (2018) 'ARIMA e ProphetFB', C, pp. 6–9.

Eliot, S. (2011) 'Using Excel for qualitative data analysis', *The Listening Resource*, (July), pp. 1–11. Available at: http://www.qualitative-researcher.com/qualitative-analysis/using-excel-for-qualitative-data-analysis/.

Gensler, A. *et al.* (2017) 'Deep Learning for solar power forecasting - An approach using AutoEncoder and LSTM Neural Networks', *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, pp. 2858–2865. doi: 10.1109/SMC.2016.7844673.

Heidari Kapourchali, M., Sepehry, M. and Aravinthan, V. (2019) 'Multivariate Spatio-temporal Solar Generation Forecasting: A Unified Approach to Deal with Communication Failure and Invisible Sites', *IEEE Systems Journal*, 13(2), pp. 1804–1812. doi: 10.1109/JSYST.2018.2869825.

Hornik, K. and Trapletti, A. (2019) 'tseries: Time Series Analysis and Computational Finance. R package version 0.10-47'.

Hyndman, R. J. (2018) *Forecasting : principles and practice*.

Hyndman, R. J. and Khandakar, Y. (2008) 'Automatic Time Series Forecasting: The forecast Package for R', *Journal of Statistical Software*, 27(3), pp. 1–22. doi: 10.18637/jss.v027.i03.

Jiang, H., Dong, Y. and Xiao, L. (2017) 'A multi-stage intelligent approach based on an ensemble of two-way interaction model for forecasting the global horizontal radiation of India', *Energy Conversion and Management*. Elsevier Ltd, 137, pp. 142–154. doi: 10.1016/j.enconman.2017.01.040.

Jifri, M. H., Hassan, E. E. and Miswan, N. H. (2017) 'Forecasting performance of time series and regression in modeling electricity load demand', *2017 7th IEEE International Conference on System Engineering and Technology, ICSET 2017 - Proceedings*, (October 2017), pp. 12–16. doi: 10.1109/ICSEngT.2017.8123412.

Li, Y. Z. and Niu, J. C. (2009) 'Forecast of power generation for grid-connected photovoltaic system based on Markov Chain', *Asia-Pacific Power and Energy Engineering Conference, APPEEC*. IEEE, pp. 1–4. doi: 10.1109/APPEEC.2009.4918386.

Marquez, R., Gueorguiev, V. G. and Coimbra, C. F. M. (2011) 'Es2011-54 Forecasting of Global Horizontal Irradiance', (209), pp. 1–7.

Prema, V. and Rao, K. U. (2015) 'Development of statistical time series models for solar

power prediction', *Renewable Energy*. Elsevier Ltd, 83, pp. 100–109. doi: 10.1016/j.renene.2015.03.038.

Premalatha, N. and Valan Arasu, A. (2016) 'Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms', *Journal of Applied Research and Technology*, 14(3), pp. 206–214. doi: 10.1016/j.jart.2016.05.001.

Purohit, I., Purohit, P. and Shekhar, S. (2013) 'Evaluating the potential of concentrating solar power generation in Northwestern India', *Energy Policy*. Elsevier, 62, pp. 157–175. doi: 10.1016/j.enpol.2013.06.069.

Ramachandra, T. V., Jain, R. and Krishnadas, G. (2011) 'Hotspots of solar potential in India', *Renewable and Sustainable Energy Reviews*. Elsevier Ltd, 15(6), pp. 3178–3186. doi: 10.1016/j.rser.2011.04.007.

Reikard, G. (2009) 'Predicting solar radiation at high resolutions: A comparison of time series forecasts', *Solar Energy*. Elsevier Ltd, 83(3), pp. 342–349. doi: 10.1016/j.solener.2008.08.007.

Sharma, N. *et al.* (2011) 'Predicting solar generation from weather forecasts using machine learning', in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, pp. 528–533. doi: 10.1109/SmartGridComm.2011.6102379.

Shearer, C. *et al.* (2000) 'The CRISP-DM model: The New Blueprint for Data Mining', *Journal of Data Warehousing*, 5(4), pp. 13–22. Available at: www.spss.com%5Cnwww.dw-institute.com.

Srivastava, R., Tiwari, A. N. and Giri, V. K. (2018) 'Forecasting of Solar Radiation in India Using Various ANN Models', in *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, pp. 1–6. doi: 10.1109/UPCON.2018.8597170.

Tiwari, S., Sabzchgar, R. and Rasouli, M. (2018) 'Short Term Solar Irradiance Forecast Using Numerical Weather Prediction (NWP) with Gradient Boost Regression', *2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems, PEDG 2018*. IEEE, pp. 1–8. doi: 10.1109/PEDG.2018.8447751.

Varma, R. and Sushil (2019) 'Bridging the electricity demand and supply gap using dynamic modeling in the Indian context', *Energy Policy*. Elsevier Ltd, 132(June), pp. 515–535. doi: 10.1016/j.enpol.2019.06.014.

Wade, A. and Nicholson, R. (2010) 'Improving Airplane Safety : Tableau and Bird Strikes', *VISWeek '10*, pp. 1–3.

Wan, C. *et al.* (2015) 'Photovoltaic and solar power forecasting for smart grid energy management', *CSEE Journal of Power and Energy Systems*, 1(4), pp. 38–46. doi: 10.17775/CSEEJPES.2015.00046.