STUDY PROJECT
MATH F266

Report On

SOLAR ENERGY DATA MODELLING

Prepared by-

RAJ SHREE SINGH

2017B4A70808P

Under Guidance of-

Prof. SUMANTA PASARI

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CONTENTS

1 Introduction

India holds the 2nd position for the most populated country globally and holds the 7th position for the largest country (by area). To cater for the economic development plans that are being implemented, India has an increasing energy demand. The National Electricity Plan [NEP] [1], framed by the Ministry of Power (MoP), has developed a 10-year detailed action plan to provide electricity across the country and has prepared a further plan to ensure that power is supplied to the citizens efficiently and at a reasonable cost. A large portion of this energy demand since the early 1990s is produced by fossil fuels (Fig 1). In 2020, India produced 61.7 % of its total energy needs via Fossil Fuels (Coal, Lignite, Gas, Diesel), 12.2% by Hydro Power sources, 1.8% by Nuclear Power Sources and the rest 24.3% by Renewable Energy Sources, according to Ministry of Power, GoI [2].

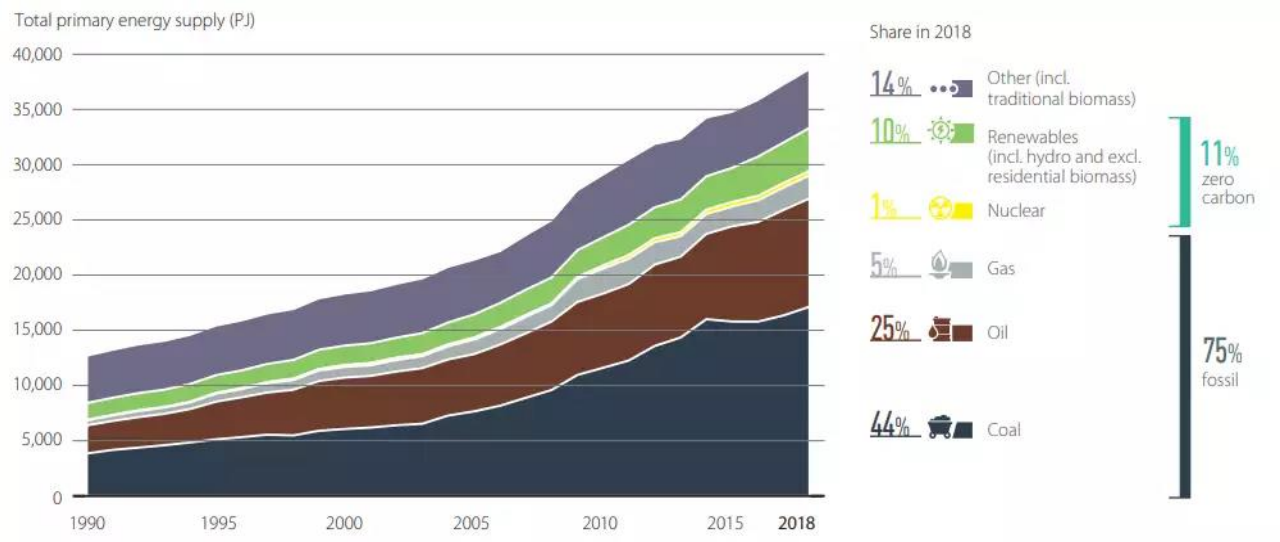


Figure 1 India's Energy Mix (1990-2018) [3]

According to the Union of Concerned Scientists Report 2020 [4], India (7 per cent) trails only behind China (28 per cent) and United States (15 per cent) when it comes to CO2 emissions. According to a World Energy Council [5] prediction, global electricity demand is expected to peak in 2030. Therefore, there is an urgent need to find alternate sources for generating electricity. In an article by Nature Communications [6], Wind and Solar energy can meet 80 per cent of India's electricity demands by 2040. The article states that achieving that level of reliance on renewable energy would reduce CO2 emissions by 85 per cent and reducing the overall costs for power generation by $50 billion. Thus, investing in renewable energy would be an essential asset in reducing India's energy costs and moving towards a low-carbon future.

When traditional sources like fossil fuels are used for power generation, the generation is predominantly controlled by the plant's generation capacity. But when renewable sources like Solar and Wind energy are used for power generation, the generation also depends on weather conditions apart from the machines' capacity. Thus, there is a need to build forecast models for the generation to have a better generation scheduling. This study project primarily focuses on Solar Energy. Considering the advantages of Solar energy over other non-renewable sources, the development and research on solar power has been rising year by year.

As the power generation from Solar Energy has heavy dependence on weather, seasonal changes, geographical location and time, the forecasting methods may not give uniformly efficient results across all regions. From Figure 2, we can observe that the production of Solar Energy varies across various states of India, with Karnataka being the largest, followed by Rajasthan, Tamil Nadu, Telangana, Andhra Pradesh and Gujrat. According to the Ministry of New and Renewable Energy progress report, its total installed solar power capacity was 36,910 MW as of December 31, 2020 [8].

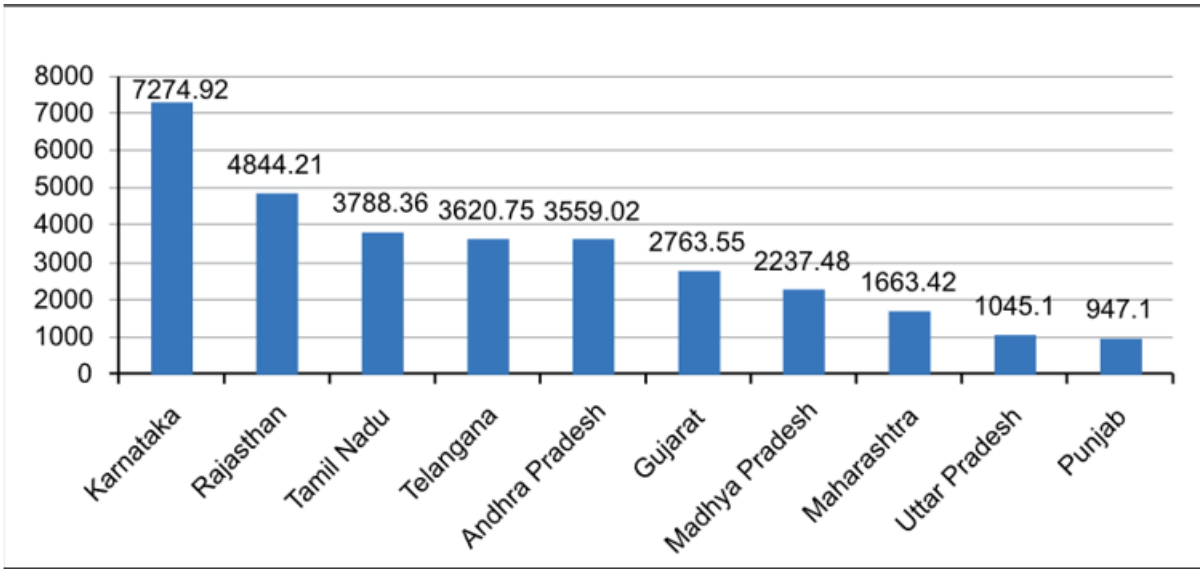


Figure 2 Top 10 states by Solar Installation (as of 31-12-2019)

This study project aims to build time series models (using ARMA, ARIMA, SARIMA and other variants) across 6 states of India to use those models for forecasting and see how these perform in comparison to ML models.

1.1 Motivation

With the increasing concern for rising amounts of CO2 in the atmosphere, there is a global need for widespread adoption of renewable energy sources. Fossil fuels have been the primary source of electricity in India for the past several decades. Now there is a need for much cleaner and sustainable renewable energy sources such as solar and wind energy.

The fundamental basis of managing existing and newly constructed power systems is Power generation forecasting; failing to do so may lead to inappropriate operational practices and inadequate energy transactions. Higher penetrations of solar energy significantly increase the uncertainties of power systems, leading to complications in system operations and planning. Thus, accurately forecasting the amount of Solar Energy received provides a valuable tool to ease the complications and enable independent system operators (ISOSs) to function more efficiently and reliably run power systems.

There are three solar energy forecasting methods – classical statistical techniques like time series forecasting, intelligent computational methods, and hybrid algorithms. Traditionally time series forecasting methods were used due to their less computational complexities, but with recent developments in Machine Learning algorithms, intelligent algorithms are extensively used for more accurate forecasting. Hence, there is a need for hybrid models that combine both these methods, and this area needs research and development.

1.2 Objective

Spatio-temporal analysis of Solar Energy data for across Karnataka, Andhra Pradesh, Rajasthan, Gujarat, Tamil Nadu, Telangana.

2 Literature Review

In this section, the concepts required for a better understanding of the project are presented. The power output from solar photovoltaic systems depends upon the incoming radiation and on the solar panel characteristics. The project is mainly concerned with forecasting solar irradiance, specifically Global Horizontal Irradiance (GHI) which is of particular interest to photovoltaic installations as it includes both Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI). The amount of solar radiation received per unit area by a surface that is always held normal to the direction of the sun at its current position in the sky is called Direct Normal Irradiance (DNI). The amount of radiation received per unit area by a surface that does not arrive on a direct path from the sun, but has been scattered by molecules and particles in the atmosphere and comes equally from all directions is called Diffuse Horizontal Irradiance (DHI). Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation received from above by a surface horizontal to the ground. [9]

Global Horizontal (GHI) = Direct Normal (DNI) * cos(θ) + Diffuse Horizontal (DHI)

Figure 3 DHI, DNI and GHI

The methods used for solar irradiance forecasting are generally divided into 4 categories –

1. Meteorological Methods - These methods are typically indirect as they do not use the GHI data rather they use satellite image processing and Numerical Weather Prediction (NWP) techniques the forecast the solar radiation intensity.
2. Statistical Methods – These methods use the time series analysis models such as Auto Regressive Moving Average (ARMA), Auto Regressive Integrated Moving Average (ARIMA) and its variants on the GHI data.
3. Machine Learning Methods – These methods use machine learning and deep learning algorithms such as Support Vector Machine (SVM), Artificial Neural Networks (ANN) etc, on the GHI data. These algorithms tend to learn patterns from the GHI data and build models according to its observations. There are two ways of utilizing machine learning techniques:
    a. By building a single prediction model.
    b. By grouping an ensemble of several prediction models together
4. Hybrid Methods – These methods combine different components from the three categories mentioned above.

The section below reviews the related literature about Statistical Methods, Machine Learning Methods and Hybrid Methods.

2.1 STATISTICAL METHODS

A series of observations measured on a variable at successive intervals is called a time series. These observations may be taken every hour, day, week, month, or year, or at any other regular interval. The pattern of the data is an important factor in understanding how the time series has behaved in the past. The past pattern can be used to select an appropriate forecasting method. The aims of time series analysis are to describe and summarize time series data, fit low-dimensional models, and make forecasts [10]. Many different models are used to predict the solar radiation time series like the classic Auto Regressive (AR) Model, the Autoregressive and Moving Average (ARMA) model, Auto Regressive Integrated Moving Average (ARIMA) and the Markov Chains.

2.1.1 ARMA Model

The Autoregressive Moving Average (ARMA) model is usually applied to auto correlated time series data. ARMA has two parts: autoregressive (AR) part and moving average (MA) part. Also, this model is usually referred as ARMA (p, q). In which p and q are the order of AR and MA respectively. Consider the ARMA(p,q) model-

$$Z_t=\varphi_1 Z_{t-1}+\cdots+\varphi_p Z_{t-p}+a_t-\theta_1 a_{t-1}-\cdots-\theta_q a_{t-q}$$

where

- $\varphi_1 \cdots \varphi_p$ are the autoregressive parameters to be estimated
- $\theta_1 \cdots \theta_q$ are the moving average parameters to be estimated
- Z_k is the value of Z at time k
- a_k is the random error of the process at time k

This model has process order(p,q), which indicates that the forecast is a linear combination of observations from previous steps and the errors from previous steps. Given the data, the process order can be determined by examining the autocorrelation function and partial autocorrelation function. It is also possible to automate the model building for online forecasting applications through calculating the information criterions

The popularity of the ARMA model is its ability to extract useful statistical properties and the adoption of the well-known Box–Jenkins methodology. ARMA models are very flexible since they can represent several different types of time series by using different order. It has been proved to be competent in prediction when there is an underlying linear correlation structure lying in the time series. One major requirement for ARMA model is that the time series must be stationary. However, from a stationary test for instance Augmented Dickey–Fuller (ADF) test, if the solar radiation series is found to be non-stationary, a stationary solar radiation time series can be easily obtained by removal of seasonality and trend.

To judge the goodness of different detrending models, Wu and Chan [12] use the Augmented Dickey–Fuller (ADF) test to measure the stationarity of the detrended series. The ADF test is a test for unit root in a time series. If there is a unit root in time series, the time series is not stationary; otherwise, it should be stationary. Wu and Chan (30) found that Al-Sadah's [13] model's performance is the best in both aspects detrending and fitting.

After the detrending phase, Wu and Chan [12] applied a classical ARMA model to the stationary series, and checked its order according to auto correlation and partial correlation. They concluded that the best order of the model is an ARMA (1, 1). The TDNN model was also used to predict the trend series. It was found to be much more sensitive than the ARMA model, but not as stable.

2.1.2 ARIMA Model

ARIMA (Auto-Regressive Integrated Moving Average) is a stochastic process coupling autoregressive component (AR) to a moving average component (MA). ARIMA models allow to treat non-stationary series. The ARIMA forecasting equation for a stationary time series consists of a *linear* equation in which the predictors consist of lags of the dependent variable and lags of the forecast errors.

Predicted value of Z = a constant and/or a weighted sum of one or more recent values of Z and/or a weighted sum of one or more recent values of the errors.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.

In terms of y, the general forecasting equation is:

$$\hat{z}_t = \mu + \phi_1 z_{t-1} + \ldots + \phi_p z_{t-p} - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$$

where

- $\phi_1 \cdots \phi_p$ are the autoregressive parameters to be estimated
- $\theta_1 \cdots \theta_q$ are the moving average parameters to be estimated
- a_k is the random error of the process at time k

Following the convention by Box and Jenkins the moving average parameters (θ's) are defined such that the signs are negative in the equation.

2.1.3 SARIMA Model

ARIMA models with a seasonal component make SARIMA models. As per the formula

SARIMA(*p*,*d*,*q*)x(*P*,*D*,*Q*,*s*),

the parameters for these types of models are as follows:

- ***p*** and seasonal ***P***: indicate number of autoregressive terms (lags of the stationarized series)
- ***d*** and seasonal ***D***: indicate differencing that must be done to stationarize series
- ***q*** and seasonal ***Q***: indicate number of moving average terms (lags of the forecast errors)
- ***s***: indicates seasonal length in the data

Reikard [14] applies a regression in log to the inputs of the ARIMA models to predict the solar radiation. He compares ARIMA models with other forecast methods such as ANN. At the 24-h horizon, he states that the ARIMA model captures the sharp transitions in irradiance associated with the diurnal cycle more accurately than other methods.

Pedro and Coimbra [15] evaluated five forecasting models with non-exogenous inputs. They compared ARIMA with a persistent model, standard k-NN, standard NN and a NN optimized by Genetic Algorithms (GA-NN) and tested the accuracy of these models using the data for eight months. Even though the results showed that GA-NN outperformed the other methods used for comparison, ARIMA also showed satisfactory accuracy.

Yang et al. [16] proposed three forecasting methods to predict the next hour solar irradiance values and the cloud cover effects. The proposed three methods take different types of meteorological data as input. The first method takes in global horizontal irradiance (GHI) values and directly uses it to forecast the GHI

values at 1-hour intervals through additive seasonal decomposition, followed by an ARIMA model. The second method forecasts diffuse horizontal irradiance (DHI) and direct normal irradiance (DNI) separately using additive seasonal decomposition, followed by an ARIMA model. The results of the two forecasts are then combined to predict GHI using an atmospheric model. The third method considers cloud cover effects and uses ARIMA to predict cloud transients. The final forecasts are made by non-linear regression techniques which uses GHI at different zenith angles and under different cloud cover conditions. Their results showed that the third method outperformed the other two, leading to MRE = 0.39 and 0.27, RMSE = 29.73 and 32.80 for the Miami and Orland test sets respectively. The results showed that the use of cloud cover techniques improves the performance of the forecasting models.

Li et al. [17] pointed out that the standard ARIMA for solar power forecasting considers only the solar power data and fails to take into account the weather information. Hence, they proposed a generalized model, ARIMAX, which allows for exogenous inputs for forecasting power output. The exogeneous inputs of the model are temperature, precipitation amount, insolation duration and humidity, which can be easily accessed. They also indicated that the proposed model is more general and flexible for practical use than the standard ARIMA and improves the performance of the latter based on the experiment results. Their results showed a 36.46% improvement in RMSE, showing that weather information can be used to enhance the performance of ARIMA for solar power forecasting.

Despite the benefits of ANN, a study by Reikard [20] compared the performance of the time series ARIMA with ANN models and found that the ARIMA generally performs better than the latter due to the effect of weather conditions, such as clouds. Solar radiation concentration is partially dependent on various weather, location, and time factors; thus, it displays a type of serial correlation, which suggests that time series forecasting is appropriate for solar radiation forecasting. ARIMA is applicable when the data is reasonably long and the correlation between past observations is stable.

2.2 MACHINE LEARNING METHODS

Machine learning methods can be used in several domains and the advantage of this method is that a model can solve problems which are impossible to be represented by explicit algorithms. The machine learning models find relations between inputs and outputs even if the representation is difficult to comprehend; this characteristic allow the use of machine learning models in many cases, for example in pattern recognition, classification problems, spam filtering, and also in data mining and forecasting problems. Machine learning concerns with the construction and study of systems that are capable of learning from data sets without being explicitly programmed.

In this project, SVM/SVR, NN and NLM are used for analysis.

2.2.1 Support Vector Regression (SVR) Model

Support vector machine is a kernel-based machine learning technique used in classification tasks and regression problems introduced by Vapnik in 1986. Support vector regression (SVR) is based on the application of support vector machines to regression problems [18]. In a similar manner as for the Gaussian Processes (GPs), the prediction calculated by a SVR machine for an input test case x is given by-

$$\hat{y} = \sum_{i=1}^{n} \alpha_i k_{rbf}(x_i, \ x_*) + b$$

With the commonly used Radial Basis Function (RBF) kernel defined by:

$$k_{rbf}(x_p, x_q) = exp\left[\frac{-(x_p - x_q)^2}{2\sigma^2}\right]$$

The parameter b is the bias parameter. In the case of SVR, the coefficients α_i are related to the difference of two Lagrange multipliers, which are the solutions of a quadratic programming (QP) problem.

One way to use the SVR in prediction problem is related to the fact that given the training dataset $D = \{x_i, y_i\}_{i=1}^{n}$ and a test input vector x_*, the forecasted clear sky index can be computed for a specific horizon, h, like:

$$\widehat{k^*}(t+h) = \sum_{i=1}^{n} \alpha_i k_{rbf}(\mathbf{x_i}, \mathbf{x_*}) + b$$

Shi et al. [25] labelled the days as sunny, foggy, cloudy and rainy based on the weather report from a meteorological station and then trained a separate SVR model for each type of day, that predicts the PV power for the next day. As input, they used the PV power output of the nearest day in the training data with the same label, and also the average daily temperature forecast for the next day. The highest accuracy was achieved for sunny days (RMSE = 1.57MW) and the lowest for foggy days (RMSE=2.52MW).

Mellit et al. [27] proposed a LS-SVM model to make short-term forecasts for meteorological time series. As input variables they used the wind speed, wind direction, air temperature, relative humidity, atmospheric pressure and solar irradiance. The SVR model was compared with several NN models (MLP, RBF, RNN and PNN), and the results showed that the LS-SVM model provided more accurate forecasts than the NN models.

Ramli et al. [28] compared SVM and NN for solar irradiance forecasts using data from Jeddah and Qassim in Saudi Aribia. They used direct diffuse and global solar irradiation on the horizontal surface as input data, and evaluated the models in terms of RMSE, MRE, correlation coefficient and computation speed. The results showed that the SVM models provided higher accuracy and more robust computation, achieving MRE = 0.33 and 0.51 for the two cities, and faster forecasting speed of 2.15s.

Ekici [29] proposed a LS-SVM model using RBF kernel to forecast the solar radiation values for the next day. The model used as inputs daily mean and maximum temperature, sunshine duration, and historical solar radiation of the day. The results showed that the proposed model was effective and feasible for the task.

In [30] Mohammadi et al. integrated SVM with a wavelet transform and pro-posed SVM-WT model to forecast horizontal global radiation for an Iranian coastal city. They combined different input parameters such as daily global radiation on a horizontal surface, relative sunshine duration, minimum ambient temperature, relative humidity, water vapour pressure and extra-terrestrial global solar radiation on a horizontal surface. The performance was compared with ARMA, NN and Genetic Programming (GP) models. The results showed that the proposed model outperformed the other models used for comparison.

Olatomiwa et al. [31] proposed the Support Vector Machine Firefly Algorithm (SVM-FFA) to forecast the mean horizontal global solar radiation values. They used sunshine duration, maximum temperature and minimum temperature as inputs. The proposed model was compared with GP and NN models, and the results showed that the proposed model achieved the best RMSE, MAPE, r and r2.

2.2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are widely used in a variety of practical tasks from process monitoring, fault diagnosis and adaptive human interference to natural events and artificial intelligence such as atmospheric processes and computers. ANN process can be considered as a black-box modeling

with a set of input factors and output variables which are results of input factors treatment through a systematic neural network.

In regression applications, inputs are mapped to outputs in a nonlinear manner. Historical data are used as inputs to an ANN and irradiance of the immediate time steps are the outputs. ANN models therefore take two steps, the training and the forecast. The training phase determines the weights of the artificial neurons and the forecasts are computed based on the trained weights. Similar to regression applications, pattern recognition applications also involve training and testing. Instead of outputting the forecast irradiance, the ANN outputs a natural number representing the classification of objects. For example, we can use ANN to classify the cloud cover index, which is a discrete measure of the opaqueness of the cloud on a scale of 0 (clear) to 10 (opaque). There is a rich literature on ANN models in solar irradiance forecasting.

.• Hybrid ANN with other techniques. Combining the ANN models with other techniques may improve the forecasting results. Cao and Cao [32] combined are current ANN with wavelet analysis for the forecast. Dong et al. [33] combined self–organized map with exponential smoothing to forecast hourly–ahead irradiance in Singapore.

•Compare various ANN architectures. Yona et al. [34] compared MLPs, RNNs, and RBNNs for 24–h–ahead power output forecasting for PV systems. Wu et al. [35] compared SVM, ANN, and GA for one–hour–ahead PV power forecast.

An evident disadvantage of ANN is its “black–box” nature. The network may determine a desired mapping between input (historical data) and output (forecasts) vectors, but does not provide any information of why a particular input is mapped to a particular output. Despite this criticism, ANN represents a significant bulk in solar irradiance forecasting. The applications of other AI–based methods such as k–nearest neighbours methods , genetic programming and fuzzy inference are less common than those of the ANN method.

In a paper by A. Gensler, J. Henze, B. Sick and N. Raabe, titled "Deep Learning for solar power forecasting — An approach using Auto Encoder and LSTM Neural Networks,". They compared performances of P-PVFM, MLP, LSTM, DBN and Auto-LSTM on a German Solar power dataset. All analysed ANN and DNN models outperform the P-PVFM. The best DNN model is the Auto-LSTM. and DNN models outperform the P-PVFM. The best DNN model is the Auto-LSTM. It combines the feature extraction ability of the Auto Encoder with the forecasting ability of the LSTM. The best performing model is the Auto-LSTM with an RMSEof0.0713, closely followed by the DBN with an RMSEof0.0714. This shows the feature extraction capability of these models, which enables a good solar power forecast. Both models without this capability, the MLP and the LSTM without an AE, perform worse.

2.3 HYBRID METHODS

These models are designed when either of the models specified above fail to perform which are supplemented by another model. Hybrid Models combine the predictions of any types of models unlike ensemble models which combines predictions of machine learning models only. Most of the literature in this field are models designed for specific regions.

Wu and Chan [19] use a hybrid model of ARMA and TDNN to improve the prediction accuracy. They suppose that the daily solar radiation series is composed by linear and nonlinear component and used the ARMA model to fit the linear component and the TDNN model to find the nonlinear pattern lying in the residual. This hybrid model has the potential to harness the unique feature and strength of both models. It is more accurate than using the ARMA or TDNN model separately.

Ref. [22] used SARIMA model for solar PV power time series forecast. The forecast accuracy was further increased by forecasting the residuals using support vector machine (SVM). In the proposed hybrid model, SARIMA model captured the linearity in the data and SVM model captured non-linearity in the residuals.

Ref. [23] used ARIMA, SVM, ANN and ANFIS for solar PV power generation forecast. The forecasts of individual models were linearly combined after optimizing the combination parameters using GA. The hybrid model was found to perform better than individual models.

3 Methodology

The dataset used in this project is obtained from National Solar Radiation Dataset (NSRDB) maintained by US Department of Energy. This dataset is used to collect data across 6 states (Karnataka, Rajasthan, Tamil Nadu, Telangana, Andhra Pradesh and Gujrat) from a period of January 1 2000 to December 31 2014.

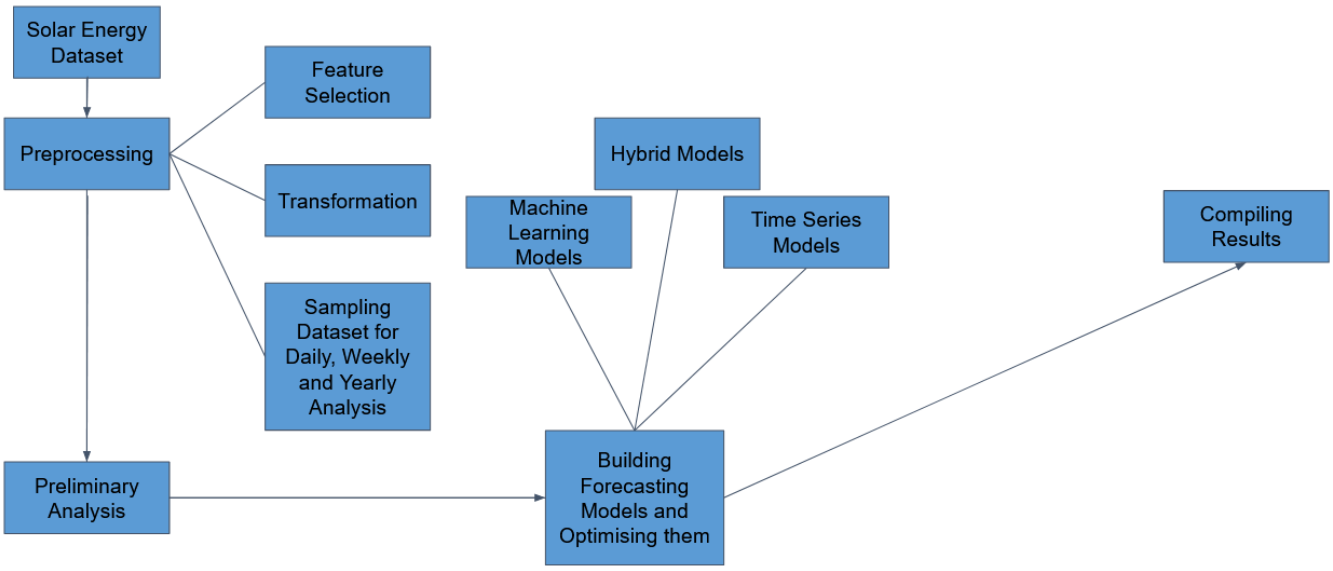
The features in the dataset are-

1. DHI and Clearsky DHI
2. DNI and Clearsky DNI
3. GHI and Clearsky GHI
4. Other environmental variables such as Temperature, Pressure, Relative Humidity, Solar Zenith etc.

The project primarily focuses on forecasting and analysing GHI and Clearsky index.

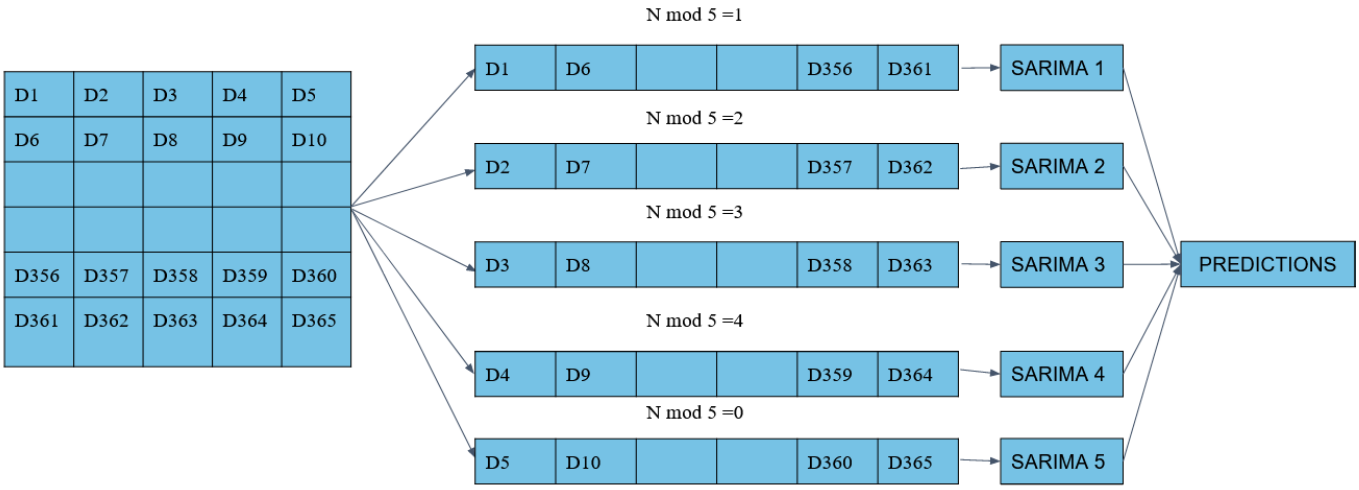
A brief outline of the project is as follows-

1. Dataset Pre-processing- This includes feature selection, choosing appropriate transformation, sampling the dataset for daily, weekly and yearly analysis and choosing an appropriate split for training and testing.
2. Preliminary Analysis – This includes analysing the training dataset to get to know the mean, median, their quartiles, standard deviation and coefficient of variance.
3. Building models for forecasting- The training dataset from each state is used to construct time series models, machine learning models and hybrid models. These models are evaluated on the test dataset and depending on the performance, the models are optimised for better results. For time series ARIMA, SARIMA and its variants will be evaluated. In machine learning models, SVR, ANN and MLP models will be evaluated. Based on their performances, Hybrid models will be considered for forecasting if the previous models fail to give satisfactory results.
4. Spatial Analysis – Models from temporal analysis are compared with other states and their differences are analysed.



Flowchart of Methodology

4 Preprocessing techniques



Flowchart explaining Daily SARIMA model

The SARIMA model for the daily dataset uses the seasonality order as 365. Due to high order of seasonality, the model consumes large memory and takes long time for training even on the Google Collab Engine. To optimize the model, each year is split into 5 datasets of 73 days computed using the modulo function (shown above). This process is repeated for every year, thus reducing the order of seasonality from 365 to 73 for each dataset. Now 5 SARIMA models were used for predictions and their results were combined together to obtain final predictions.

	Input Matrix	**Output**
Monthly Dataset	GHI values for the days in that month (input dim -31,28,30)	Mean GHI value for the month
Weekly Dataset	GHI values for the days in that week (input dim -7)	Mean GHI value for the week
Daily Dataset	GHI values for the 9 hours in that day (input dim -9)	Mean GHI value for that day

The above table shows how the input and output training vectors are obtained for the Monthly, Weekly and Daily Datasets.

5 Preliminary Analysis

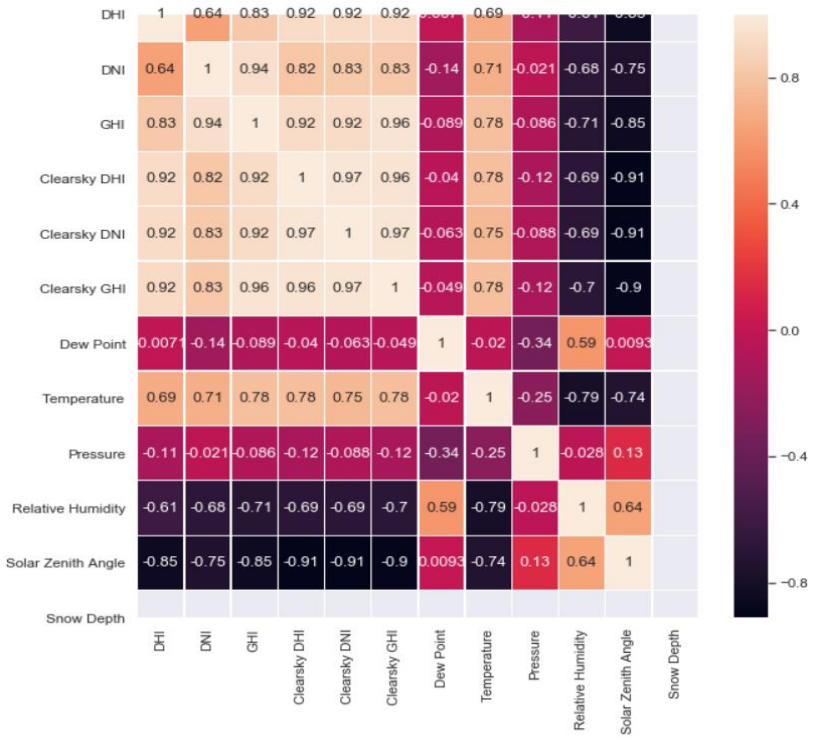


Figure 4 Correlation Heatmap

The above figure show Correlation Heatmap for various features available in the dataset. We can see that the GHI index is highly correlated to DHI, DNI, Clearsky GHI, Clearsky DNI, Clearsky DHI and Temperature. Thus, the project will be focussing on GHI and Clearsky GHI mainly for analysis.

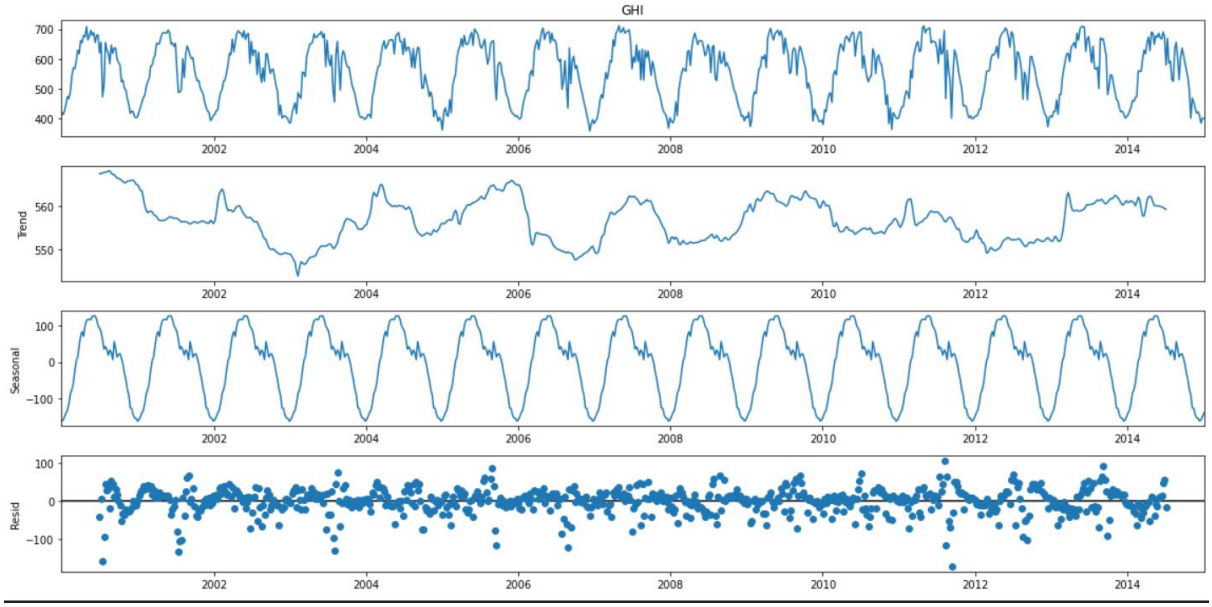
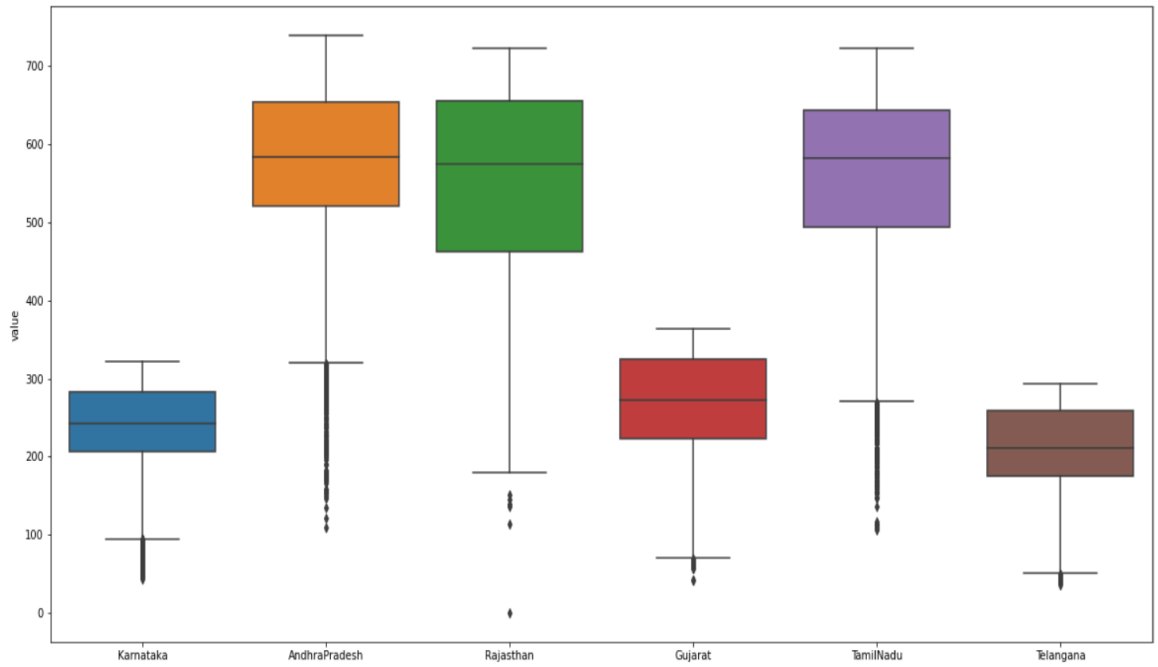


Figure 5 Time Series Decomposition (Additive) for a site near Pokhran, Rajasthan

The above figure shows the additive time series decomposition for a site near Pokhran, Rajasthan.



Boxplot of the 6 states on the Daily Dataset

The above figure shows the boxplots of GHI values for the Daily Dataset across 6 states. We can clearly observe that Rajasthan, Andhra Pradesh and Tamil Nadu have the highest median GHI values. The standard deviation and their coefficient of Variance are summarized below-

States	Karnataka	Andhra Pradesh	Rajasthan	Gujarat	Tamil Nadu	Telangana
Std. Deviation	54.31	106.34	110.04	61.75	117.59	54.96
Coeff. of Variance	0.23	0.186	0.197	0.229	0.212	0.262

From the table we observe that the states with highest coefficient of Variance is Telangana followed by Karnataka, Gujarat, Tamil Nadu, Rajasthan and Andhra Pradesh.

The results from the SARIMA, MLP , LSTM models for the monthly, weekly and daily datasets are discussed below.

6 Results

6.1 Andhra Pradesh

6.1.1 Monthly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled monthly.

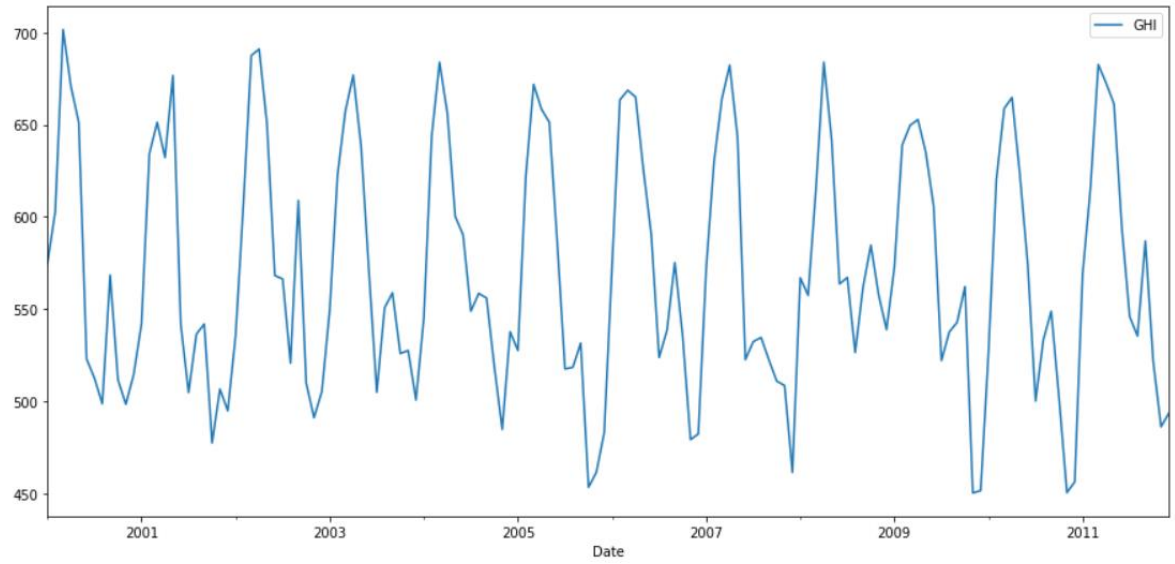


Figure 7 Monthly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,0,1)[12] gives us the best results with MAPE = 2.53 per cent and RMSE =16.82. The graph below shows the results from MLP model with MAPE = 4.96 per cent and RMSE = 34.62. The LSTM model predictions have RMSE = 130.47 and MAPE = 16.44 (Since the LSTM model gives significantly worse results,it is not shown in the graph below). Thus for the monthly dataset of Andhra Pradesh, SARIMA model performs best.

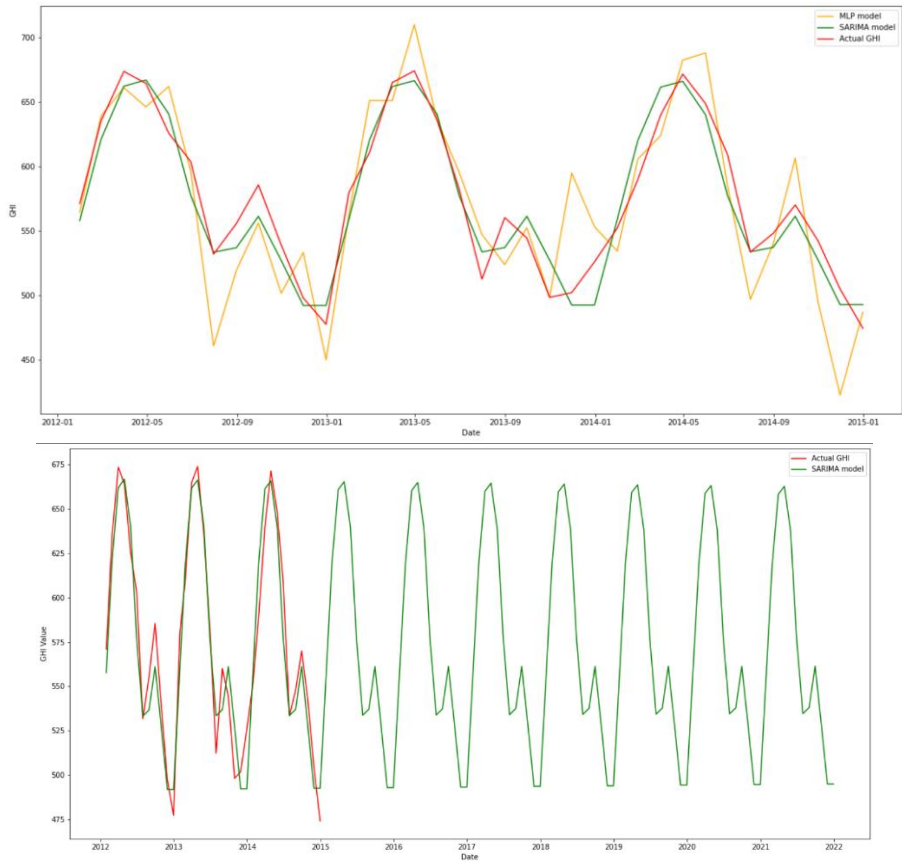


Figure 8 Results from the prediction models for the training dataset and extrapolation of SARIMA model till 2022

6.1.2 Weekly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled weekly.

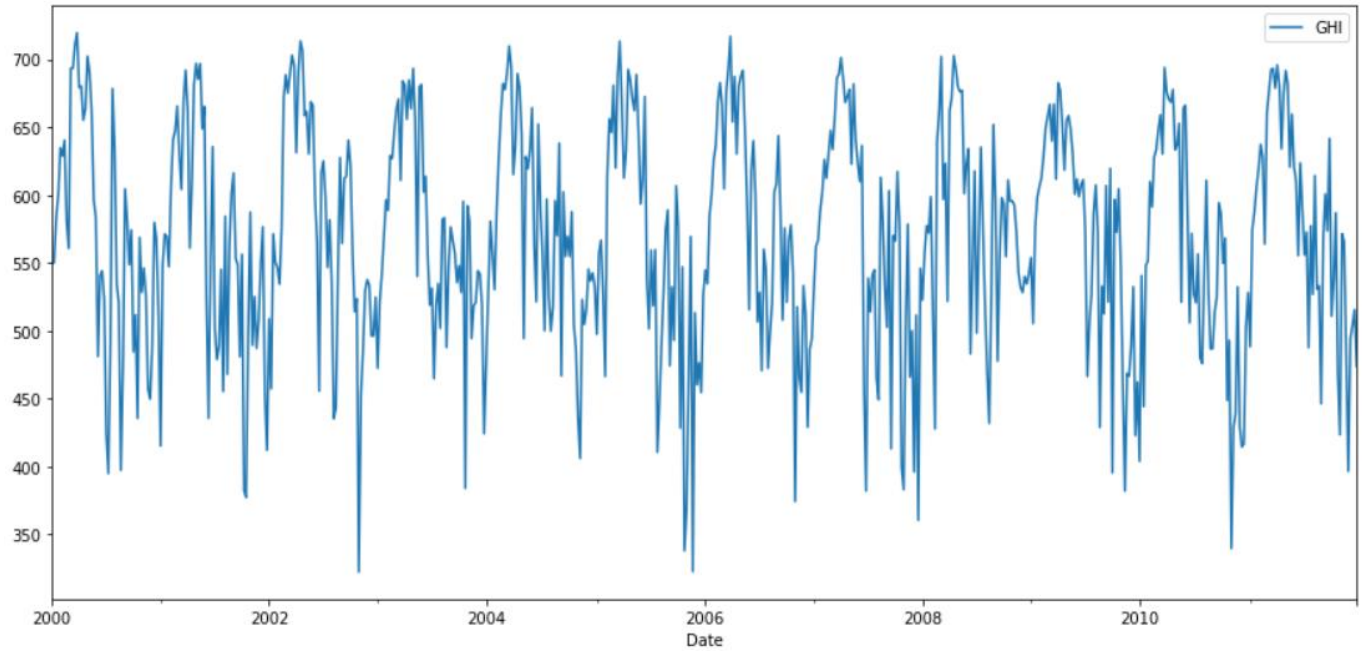


Figure 9 Weekly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[52] gives us the best results with MAPE = 7.92 per cent and RMSE = 57.08. The graph below shows the results from MLP model with MAPE = 4.78 per cent and RMSE = 35.97. The LSTM model predictions have MAPE = 4.60 and RMSE = 35.089. Thus for the weekly dataset of Andhra Pradesh, LSTM model gives the best predictions.

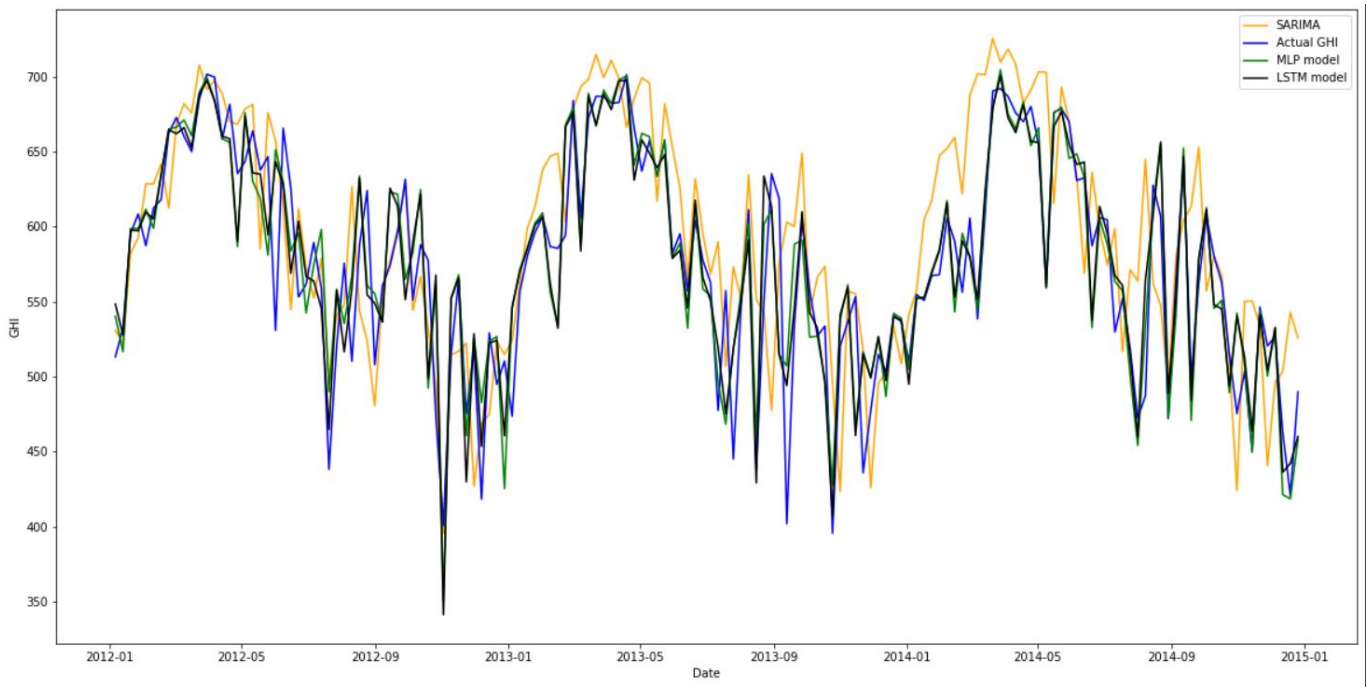


Figure 10 Results from the prediction models for the training dataset

6.1.3 Daily dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled daily.

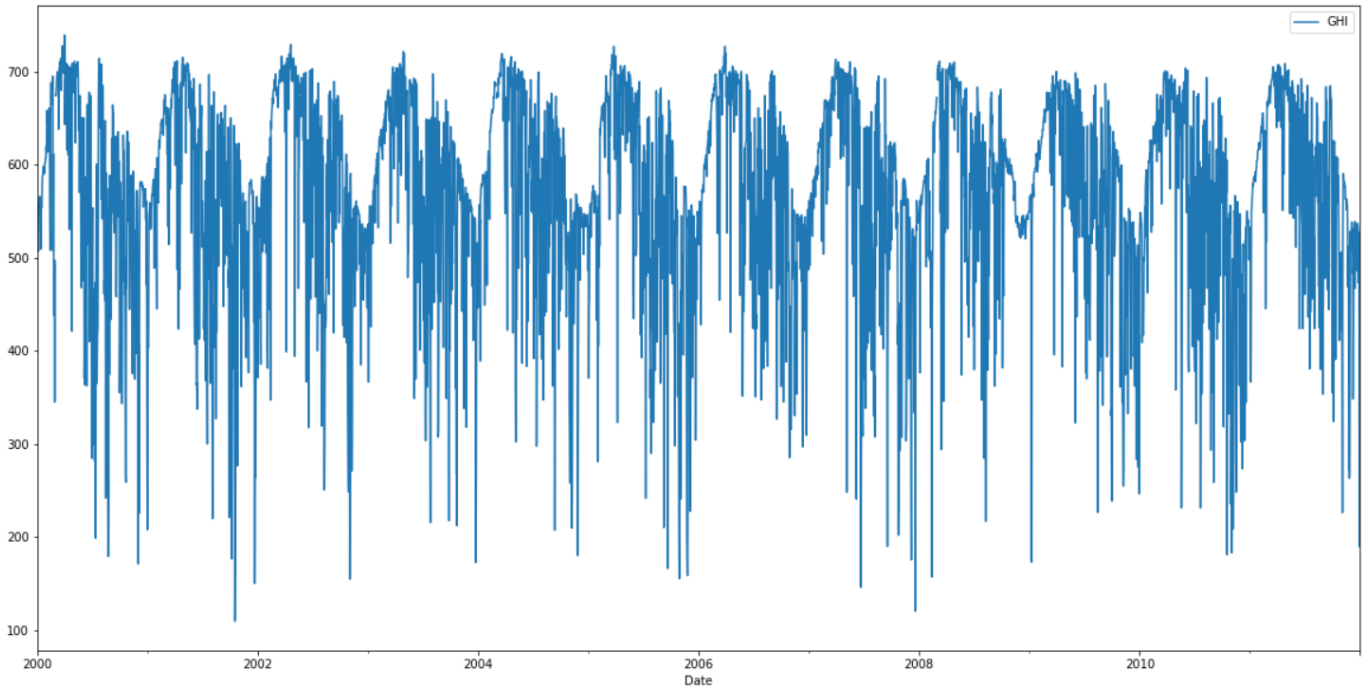


Figure 11 Daily Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[365] gives us the best results with MAPE = 26.73 per cent and RMSE = 178.59 (The predictions from ARIMA model gives significantly worse results, hence it is not shown in the graph below). The graph below shows the results from MLP model with MAPE = 0.0054 per cent and RMSE = 0.0741. The LSTM model predictions have MAPE = 0.01 and RMSE = 0.072. Thus for the Daily dataset of Andhra Pradesh, MLP model performs best.

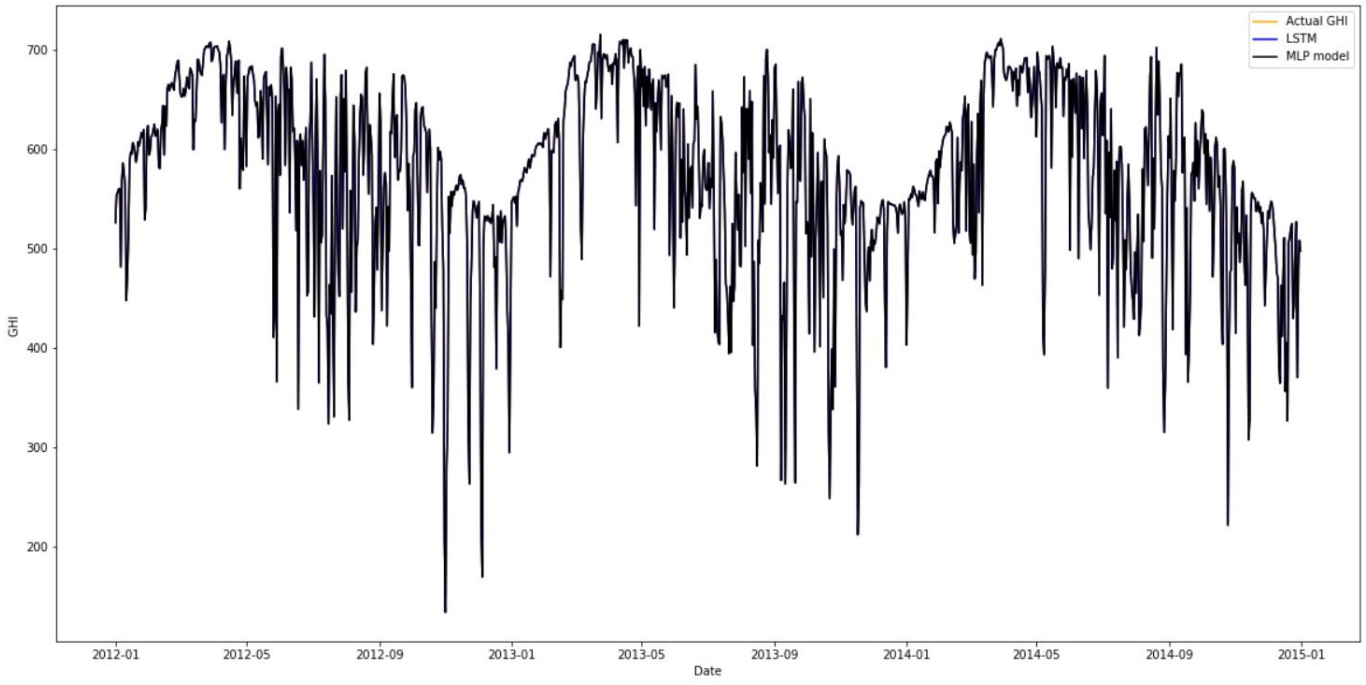


Figure 12 Results from the prediction models for the training dataset

The LSTM model, MLP model and the actual GHI values in the above graph appear to overlap because the predictions and the actual values differ significantly less.

6.2 Rajasthan

6.2.1 Monthly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled monthly.

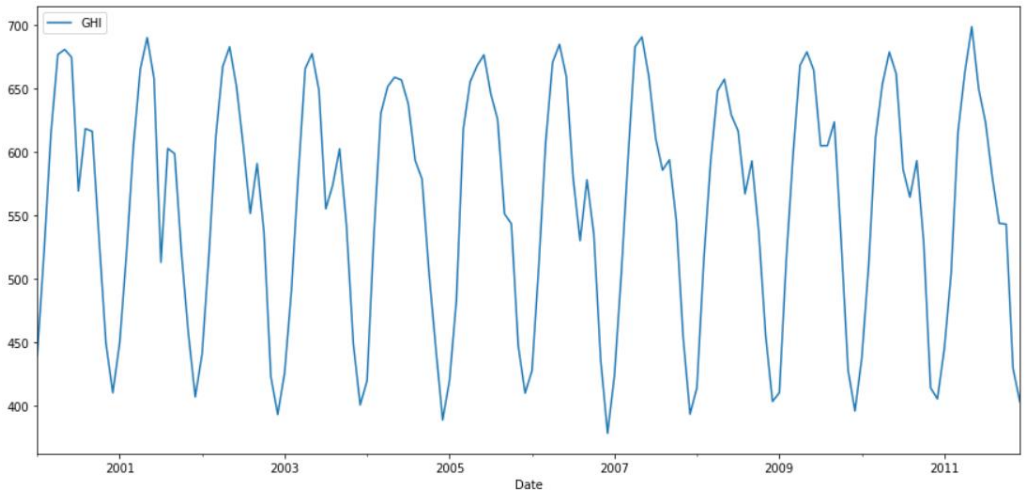


Figure 13 Monthly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 1)(1,0,1)[12] gives us the best results with MAPE = 2.1 per cent and RMSE =14.713. The graph below shows the results from MLP model with MAPE = 3.11 per cent and RMSE = 23.65. The LSTM model predictions have RMSE = 83.43 and MAPE = 6.81 (Since the LSTM model gives significantly worse results,it is not shown in the graph below). Thus for the monthly dataset of Rajasthan, SARIMA model performs best.

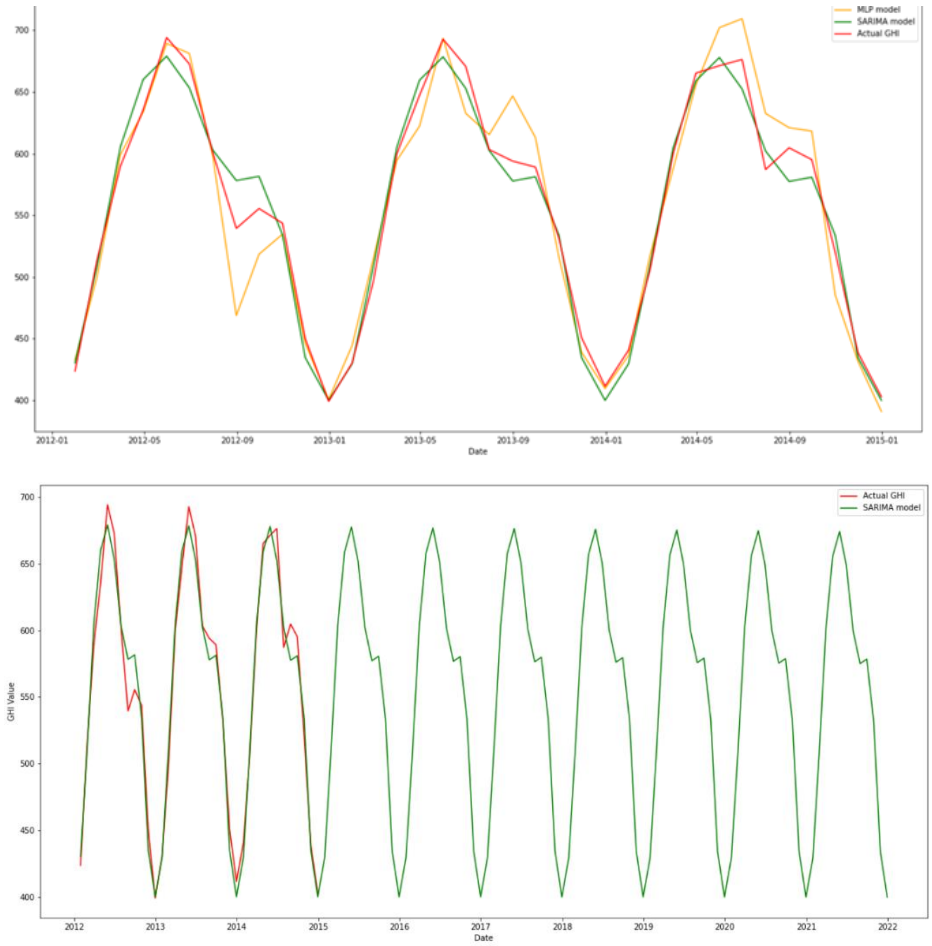


Figure 14 Results from the prediction models for the training dataset and extrapolation of SARIMA model till 2022

6.2.2 Weekly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled weekly.

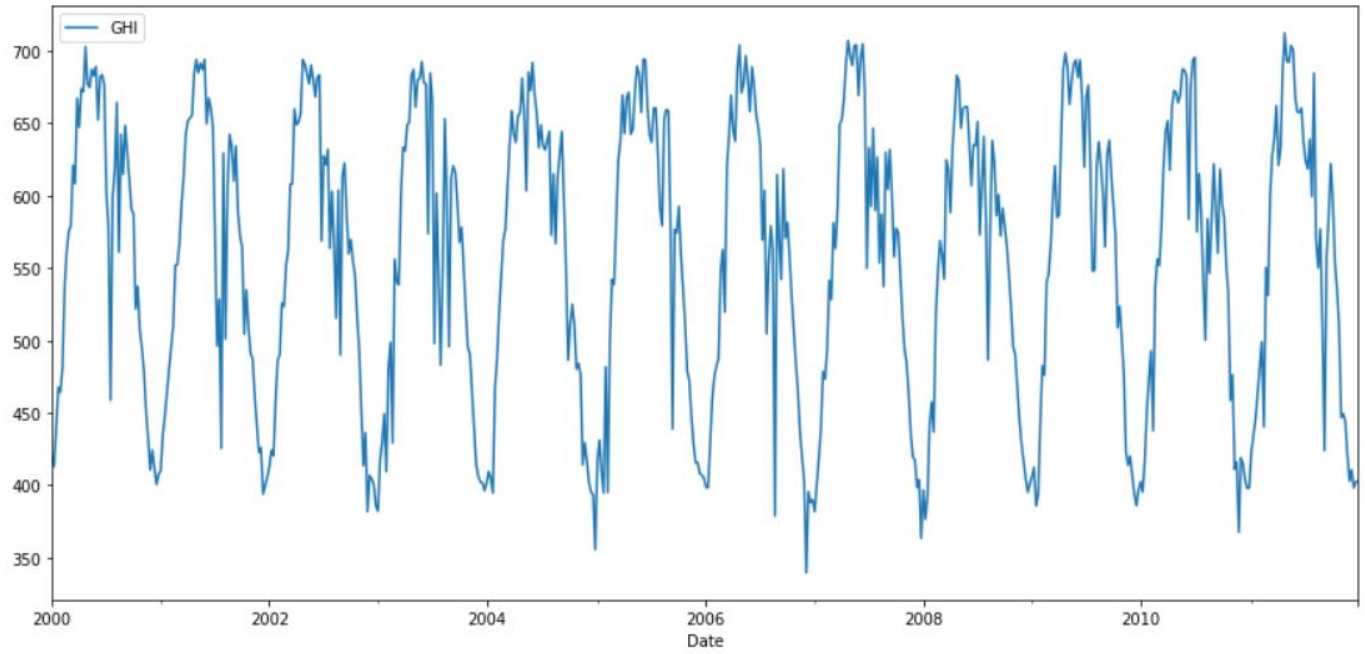


Figure 15 Weekly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[52] gives us the best results with MAPE = 5.52 per cent and RMSE = 42.717. The graph below shows the results from MLP model with MAPE = 3.20 per cent and RMSE = 28.47. The LSTM model predictions have MAPE = 2.98 and RMSE = 23.99. Thus for the weekly dataset of Rajasthan, LSTM model performs best.

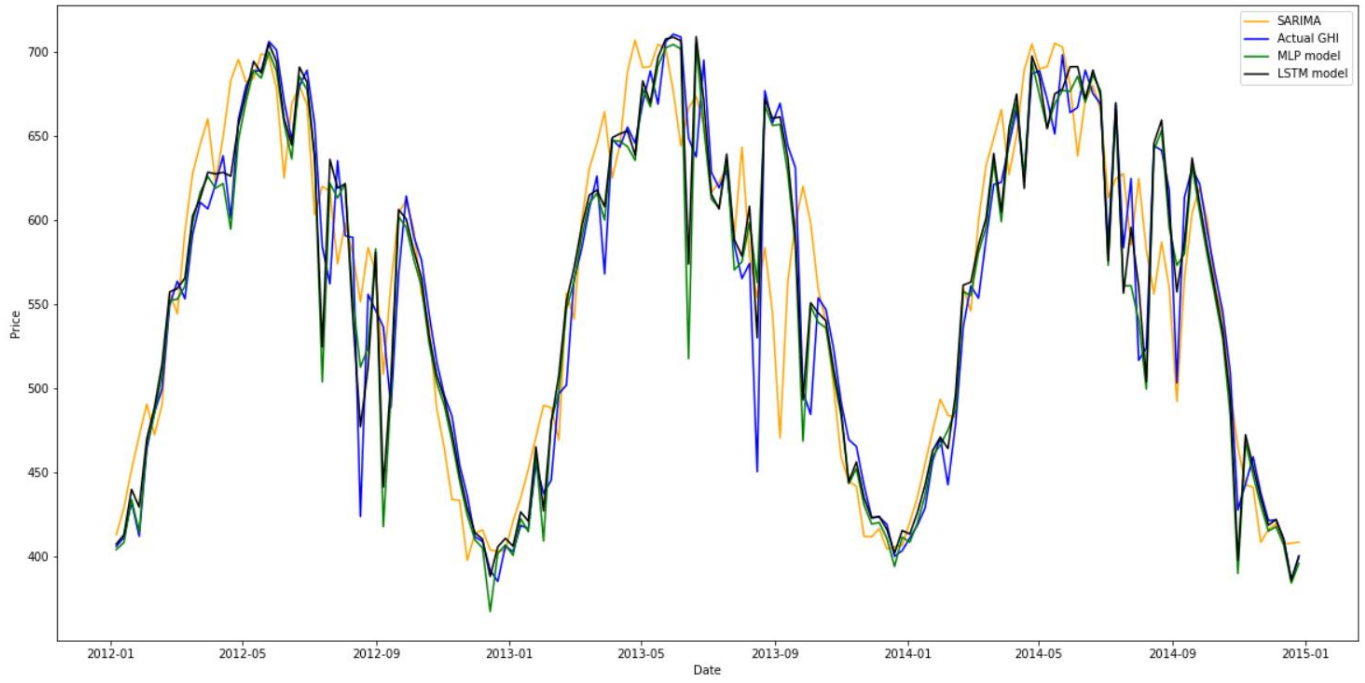


Figure 66 Results from the prediction models for the training dataset

6.2.3 Daily dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled daily.

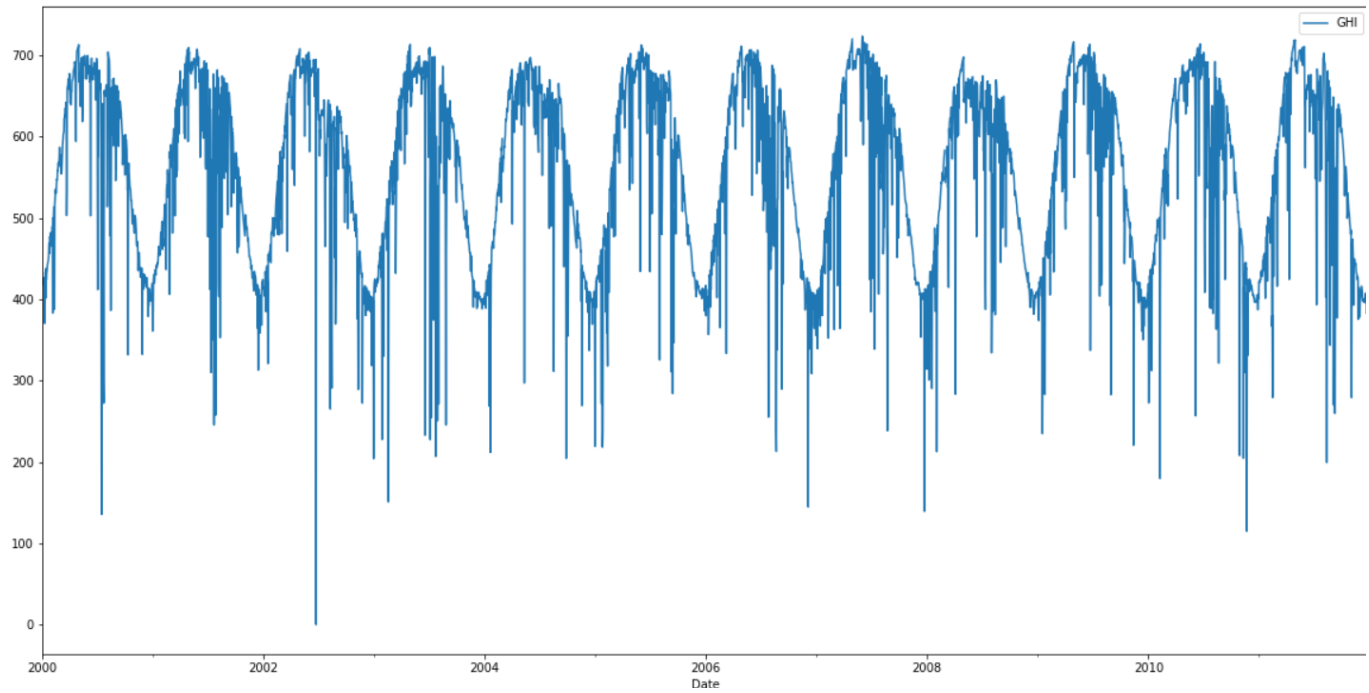


Figure 17 Daily Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[365] gives us the best results with MAPE = 9.38 per cent and RMSE = 76.13. The graph below shows the results from MLP model with MAPE = 0.0021 per cent and RMSE = 0.0207. The LSTM model predictions have MAPE = 0.0097 and RMSE = 0.078. Thus for the Daily dataset of Rajasthan, MLP model performs best.

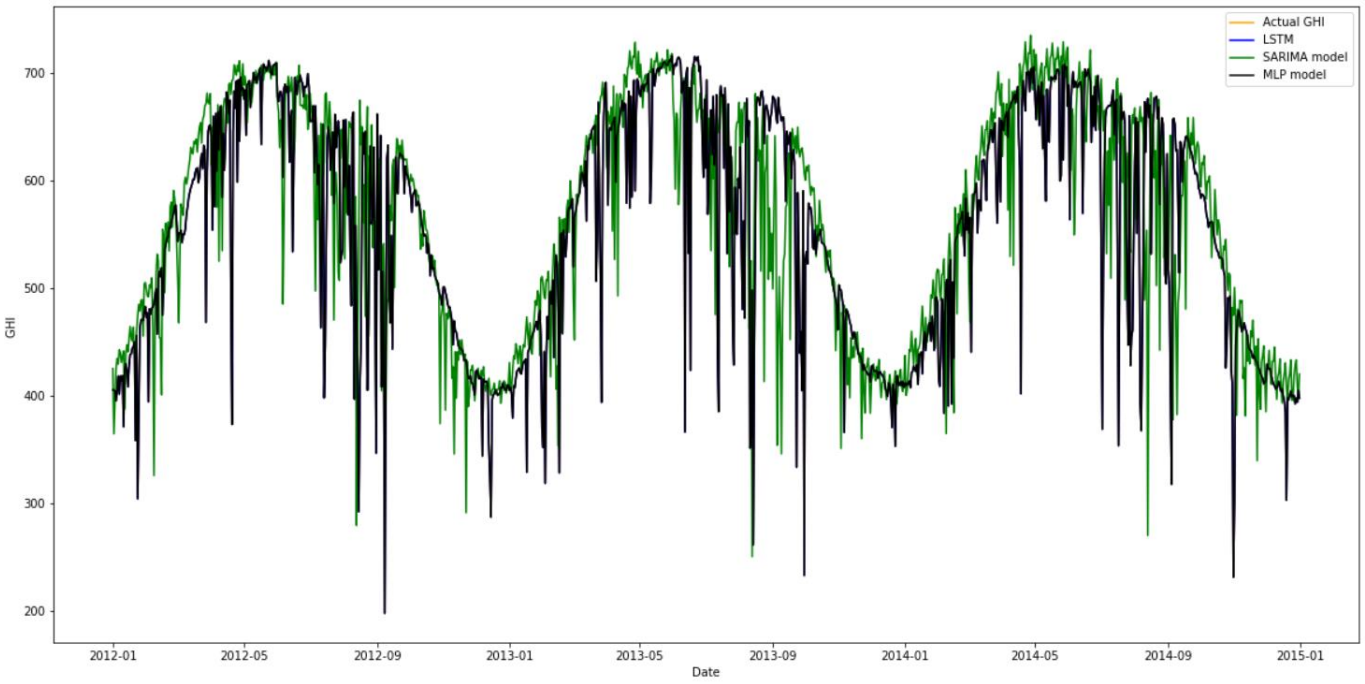


Figure 78 Results from the prediction models for the training dataset

The LSTM model, MLP model and the actual GHI values in the above graph appear to overlap because the predictions and the actual values differ significantly less. The SARIMA model's predictions (in green) differs significantly from the other models.

6.3 Karnataka

6.3.1 Monthly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled monthly.

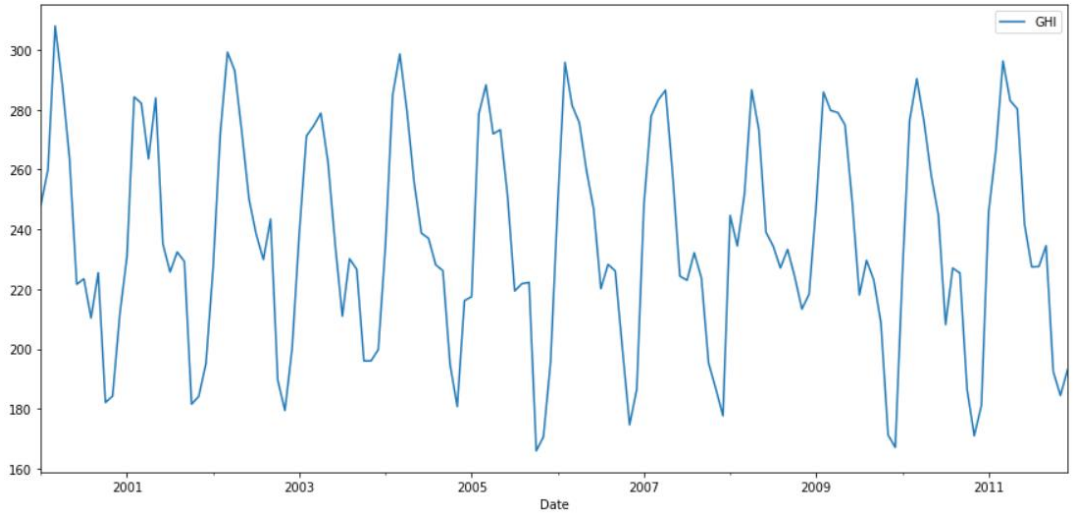


Figure 19 Monthly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 1)(1,0,1)[12] gives us the best results with MAPE = 2.83 per cent and RMSE =8.57. The graph below shows the results from MLP model with MAPE = 5.94 per cent and RMSE = 17.76. The LSTM model predictions have RMSE = 43.66 and MAPE = 10.17 (Since the LSTM model gives significantly worse results, it is not shown in the graph below). Thus, for the monthly dataset of Karnataka, SARIMA model performs best.

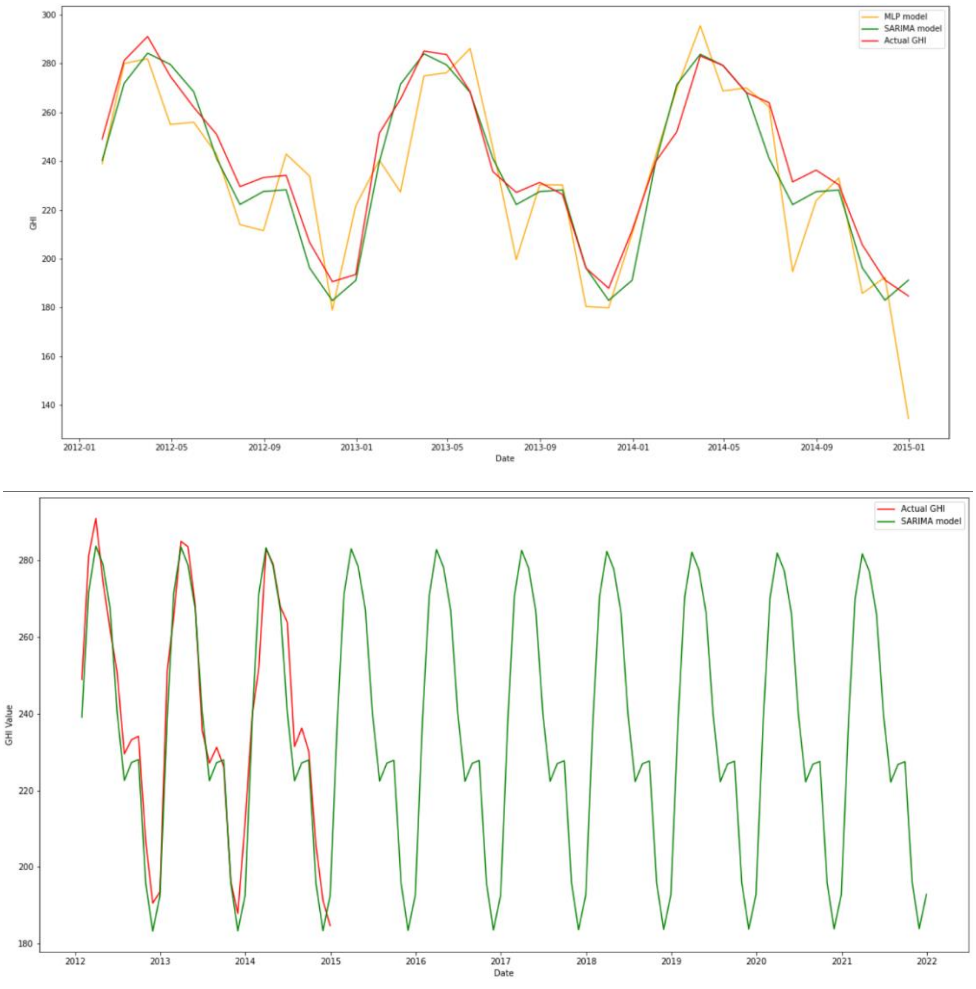


Figure 20 Results from the prediction models for the training dataset and extrapolation of SARIMA model till 2022

6.3.2 Weekly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled weekly.

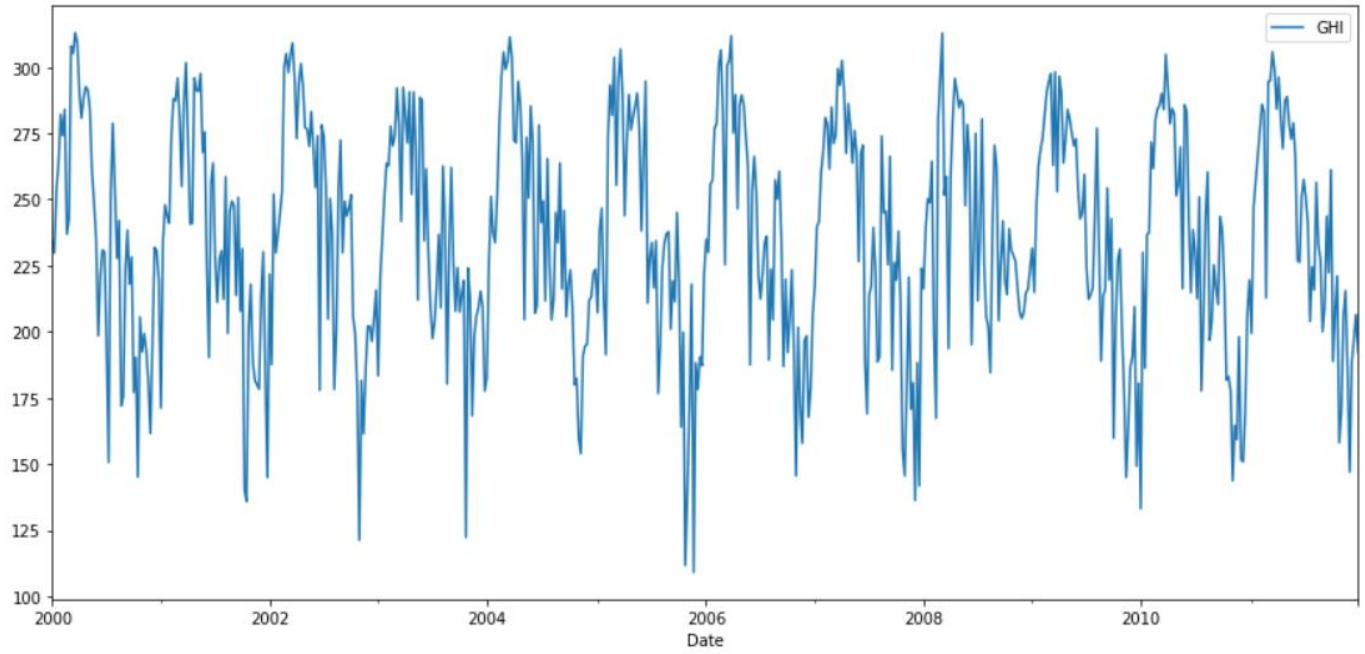


Figure 21 Weekly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[52] gives us the best results with MAPE = 11.23 per cent and RMSE = 32.03. The graph below shows the results from MLP model with MAPE = 5.69 per cent and RMSE = 17.31. The LSTM model predictions have MAPE = 5.78 and RMSE = 17.14. Thus for the weekly dataset of Karnataka, both MLP and LSTM model perform equally good but LSTM model has slightly lesser RMSE while their MAPE values are approximately same.

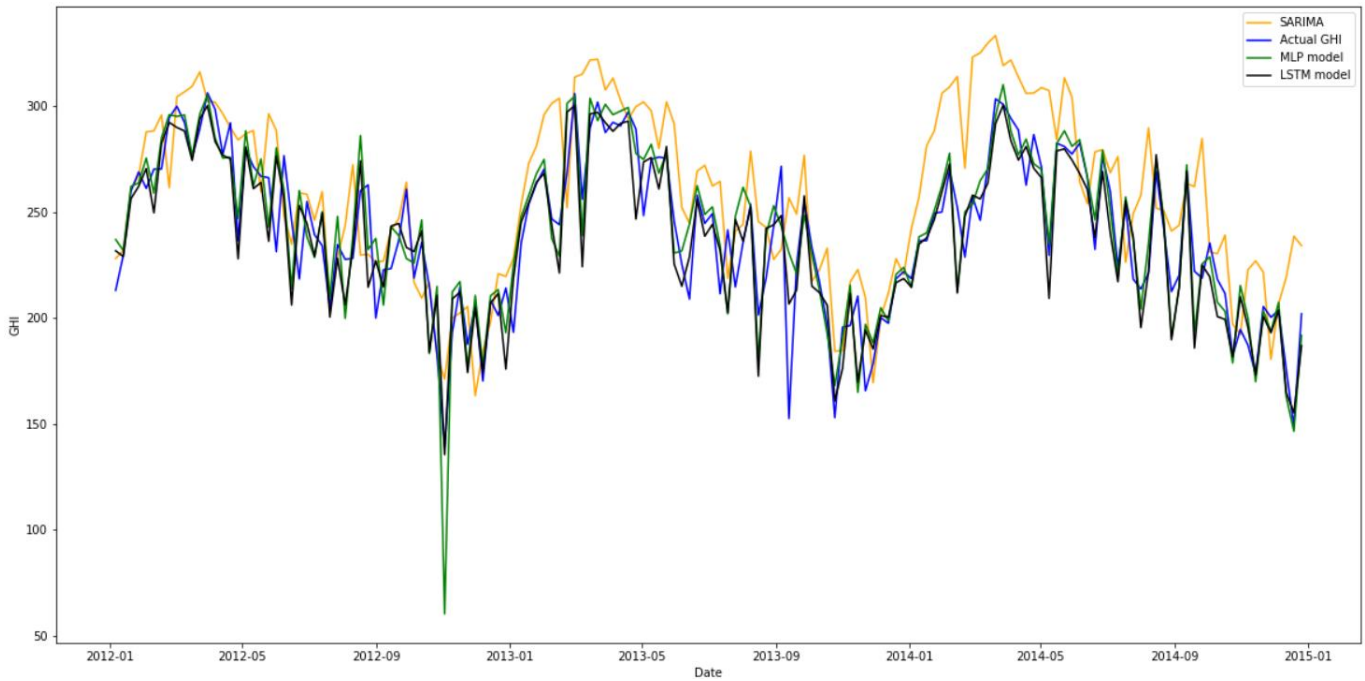
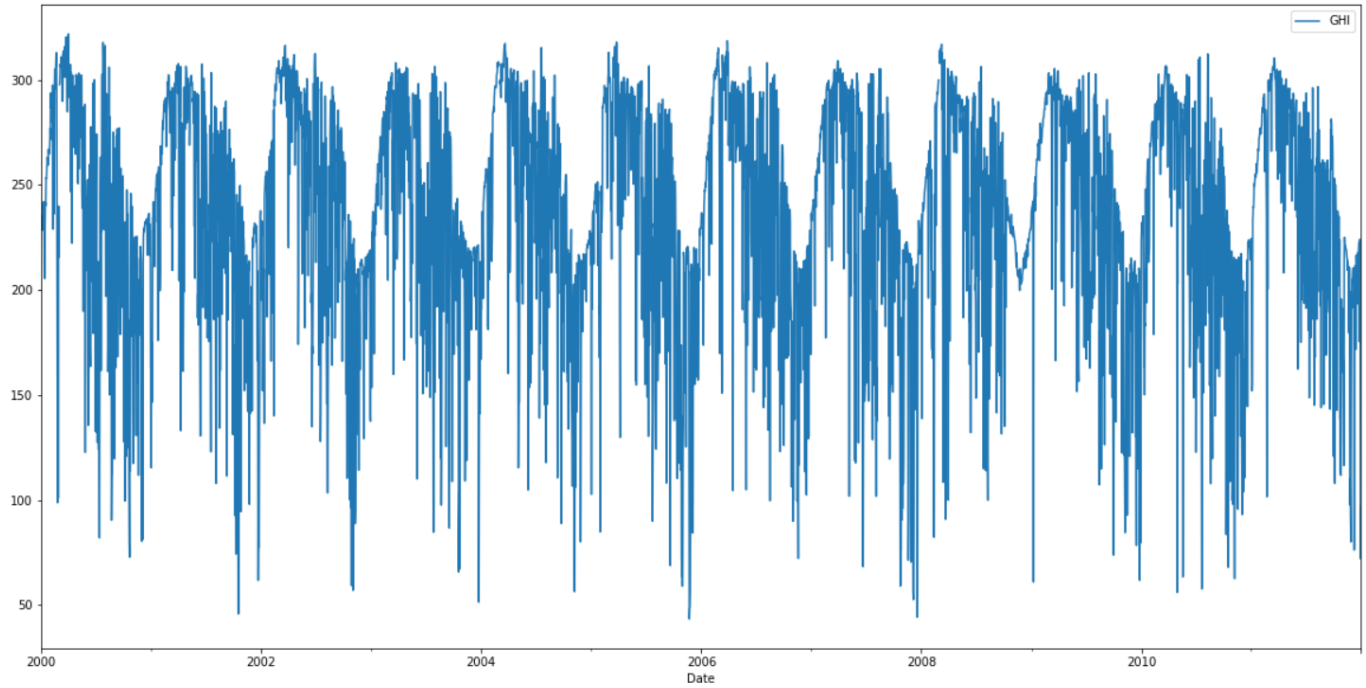


Figure 22 Results from the prediction models for the training dataset

6.3.3 Daily dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled daily.



s

Figure 23 Daily Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[365] gives us the best results with MAPE = 34.03 per cent and RMSE = 89.064 (The predictions from ARIMA model gives significantly worse results, hence it is not shown in the graph below). The graph below shows the results from MLP model with MAPE = 0.16 per cent and RMSE = 0.399. The LSTM model predictions have MAPE = 0.16 and RMSE = 0.443. Thus for the Daily dataset of Karnataka, MLP model performs best.

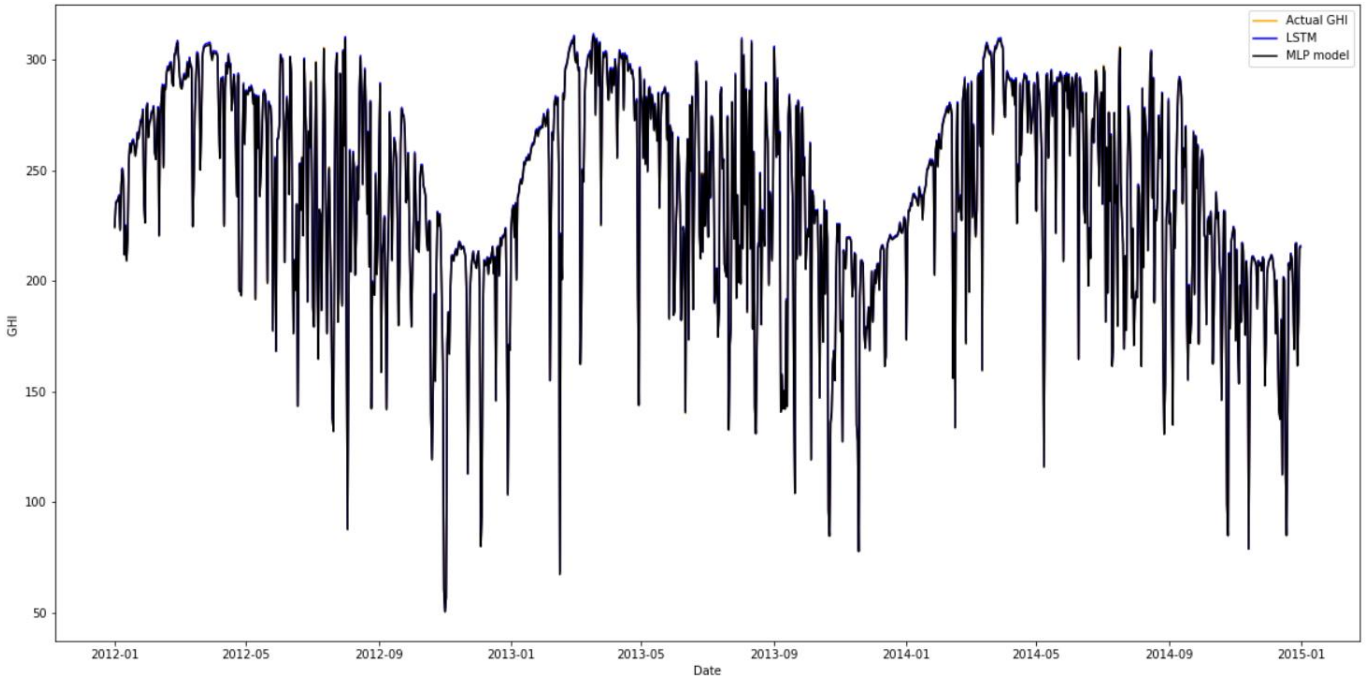


Figure 24 Results from the prediction models for the training dataset

The LSTM model, MLP model and the actual GHI values in the above graph appear to overlap because the predictions and the actual values differ significantly less.

6.4 Gujarat

6.4.1 Monthly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled monthly.

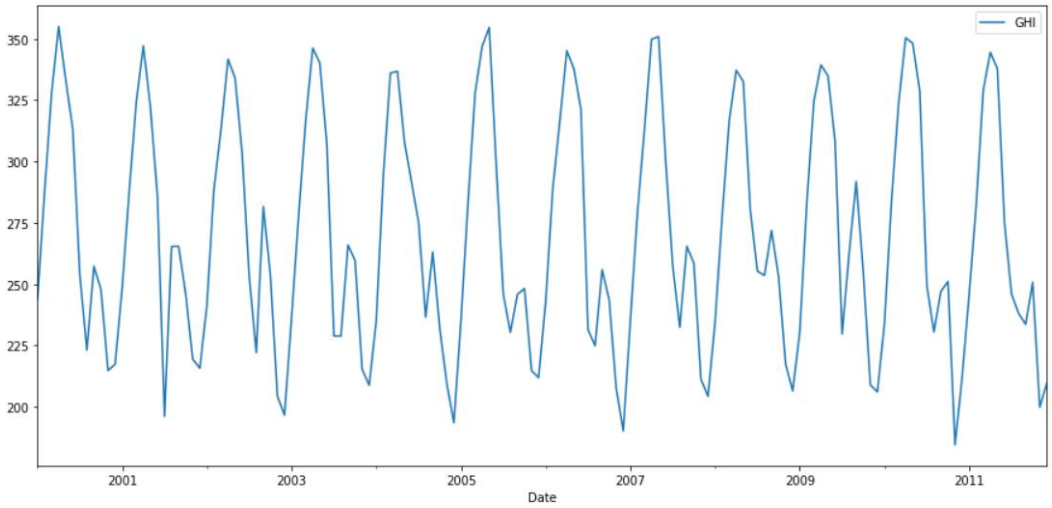


Figure 25 Monthly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 1)(1,0,0)[12] gives us the best results with MAPE = 7.2 per cent and RMSE =28.31. The graph below shows the results from MLP model with MAPE = 4.12 per cent and RMSE = 21.24. The LSTM model predictions have RMSE = 44.89 and MAPE = 9.83 (Since the LSTM model gives significantly worse results, it is not shown in the graph below). Thus, for the monthly dataset of Gujarat, MLP model performs best.

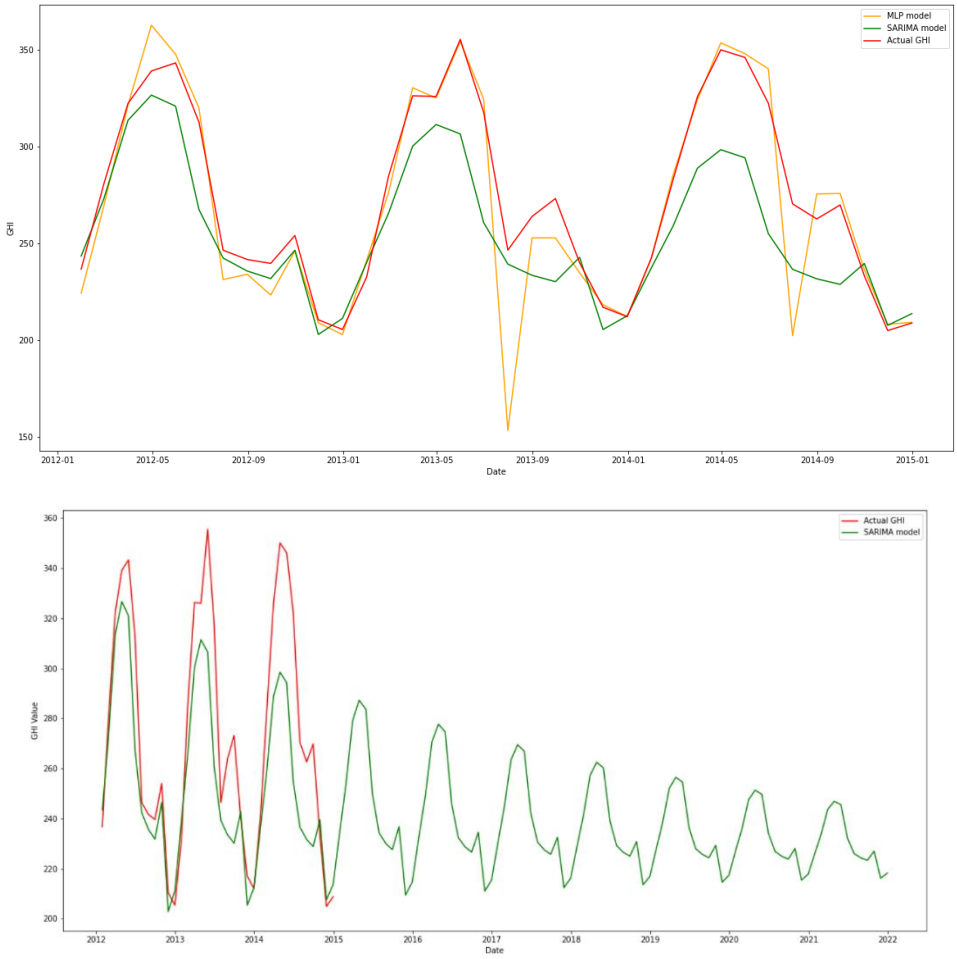


Figure 268 Results from the prediction models for the training dataset and extrapolation of SARIMA model till 2022

6.4.2 Weekly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled weekly.

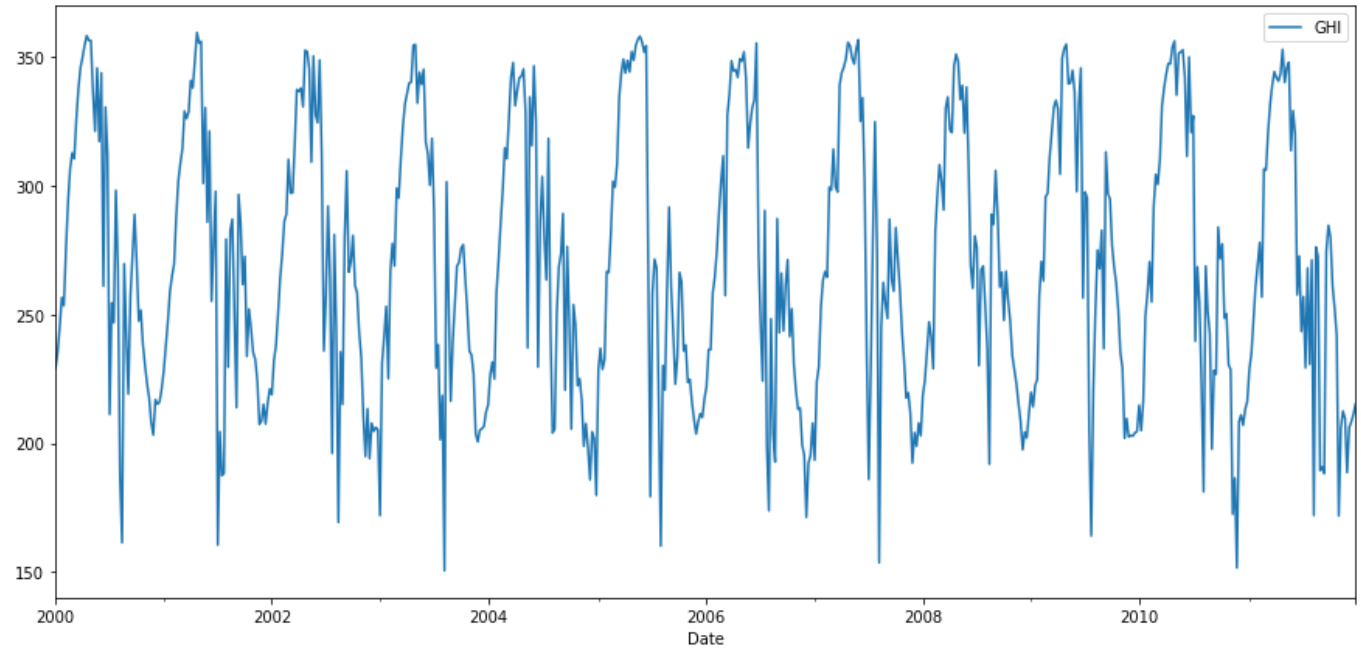


Figure 27 Weekly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[52] gives us the best results with MAPE = 6.84 per cent and RMSE = 27.02. The graph below shows the results from MLP model with MAPE = 3.99 per cent and RMSE = 16.79. The LSTM model predictions have MAPE = 4.06 and RMSE = 16.12. Thus for the weekly dataset of Gujarat, both MLP and LSTM model perform equally good but LSTM model has slightly lesser RMSE while their MAPE values are approximately same.

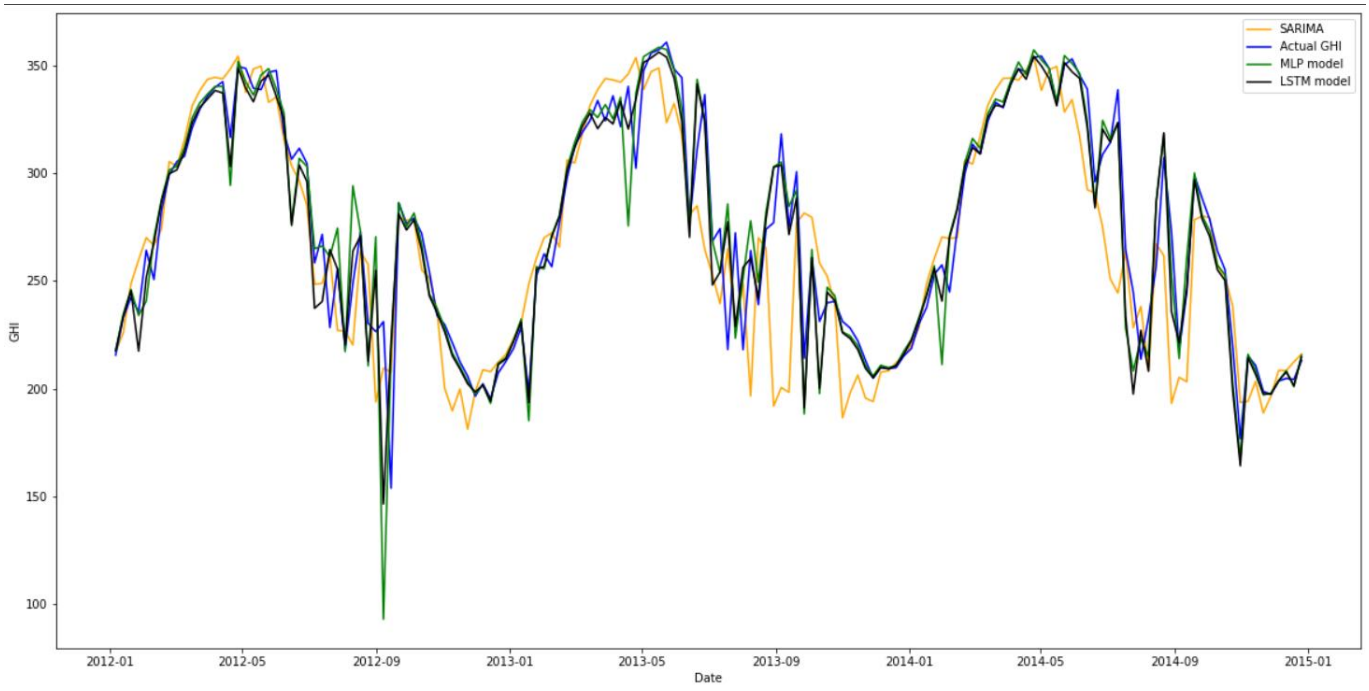


Figure 28 Results from the prediction models for the training dataset

6.4.3 Daily dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled daily.

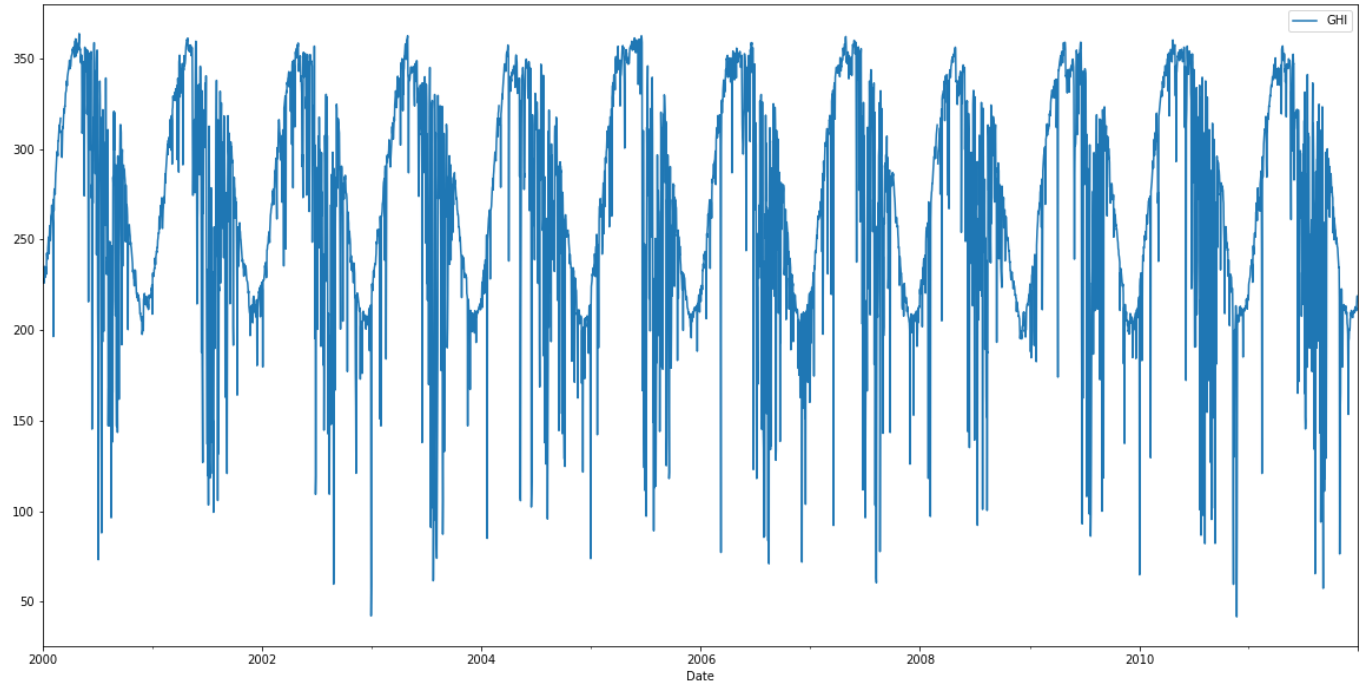


Figure 29 Daily Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[365] gives us the best results with MAPE = 12.84 per cent and RMSE = 48.34. The graph below shows the results from MLP model with MAPE = 0.015 per cent and RMSE = 0.045. The LSTM model predictions have MAPE = 0.015 and RMSE = 0.0444. Thus for the Daily dataset of Gujarat, LSTM model performs silghtly better than MLP.

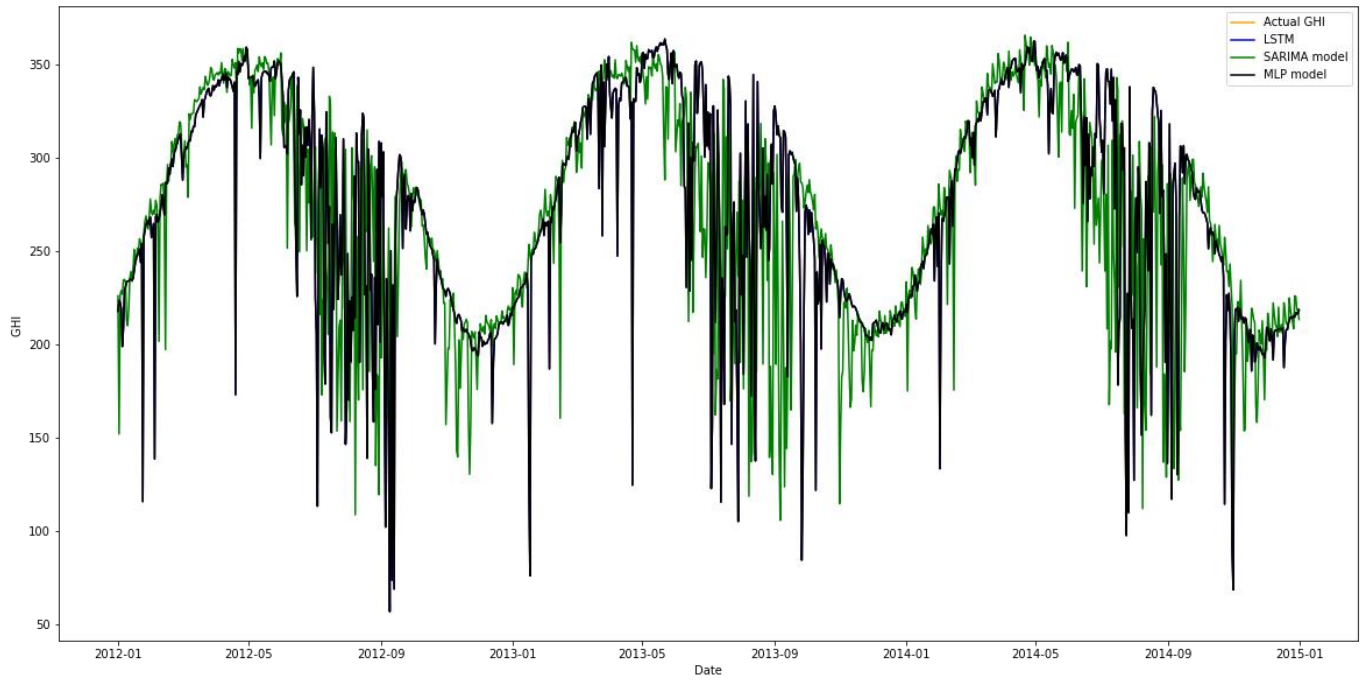


Figure 94 Results from the prediction models for the training dataset

The LSTM model, MLP model and the actual GHI values in the above graph appear to overlap because the predictions and the actual values differ significantly less. The SARIMA model differ significantly from the predictions than the other models.

6.5 Tamil Nadu

6.5.1 Monthly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled monthly.

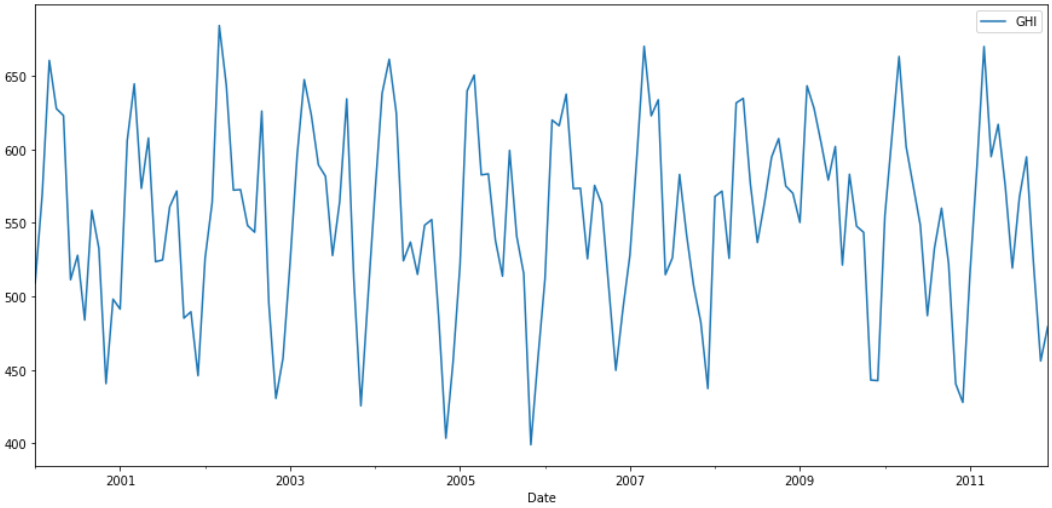


Figure 30 Monthly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 1)(1,0,0)[12] gives us the best results with MAPE = 4.59 per cent and RMSE =29.46. The graph below shows the results from MLP model with MAPE = 6.67 per cent and RMSE = 48.80. The LSTM model predictions have RMSE = 60.76 and MAPE = 8.91 (Since the LSTM model gives significantly worse results, it is not shown in the graph below). Thus, for the monthly dataset of Tamil Nadu, SARIMA model performs best.

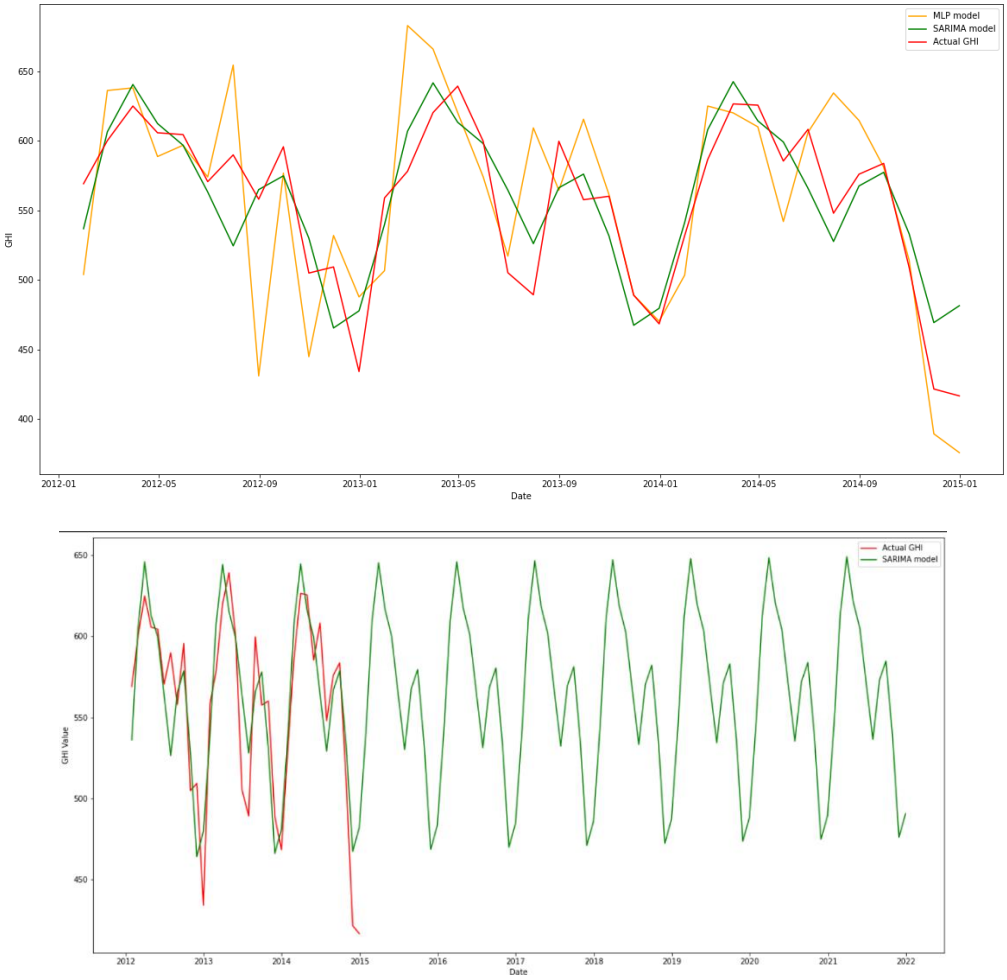


Figure 31 Results from the prediction models for the training dataset and extrapolation of SARIMA model till 2022

6.5.2 Weekly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled weekly.

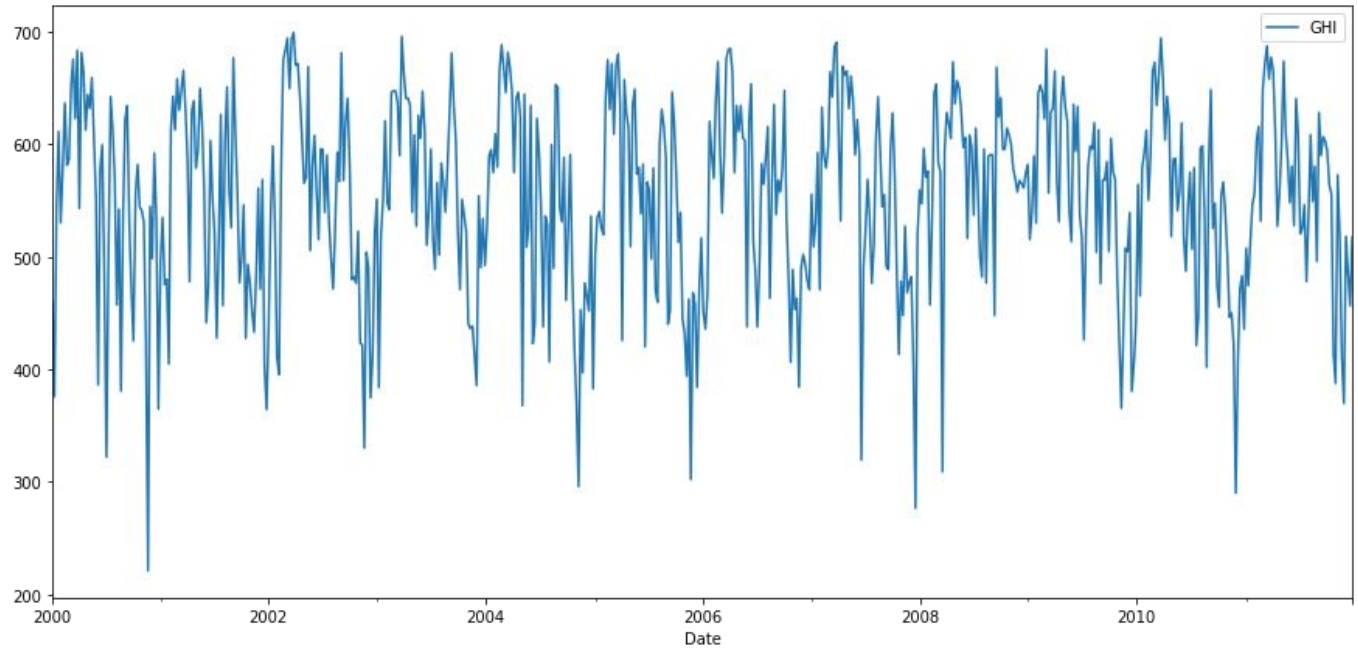


Figure 32 Weekly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (0,1,1)(1,1,0)[52] gives us the best results with MAPE = 10.9 per cent and RMSE = 70.87. The graph below shows the results from MLP model with MAPE = 6.77 per cent and RMSE = 45.23. The LSTM model predictions have MAPE = 6.61 and RMSE = 43.74. Thus for the weekly dataset of Tamil Nadu LSTM model performs better.

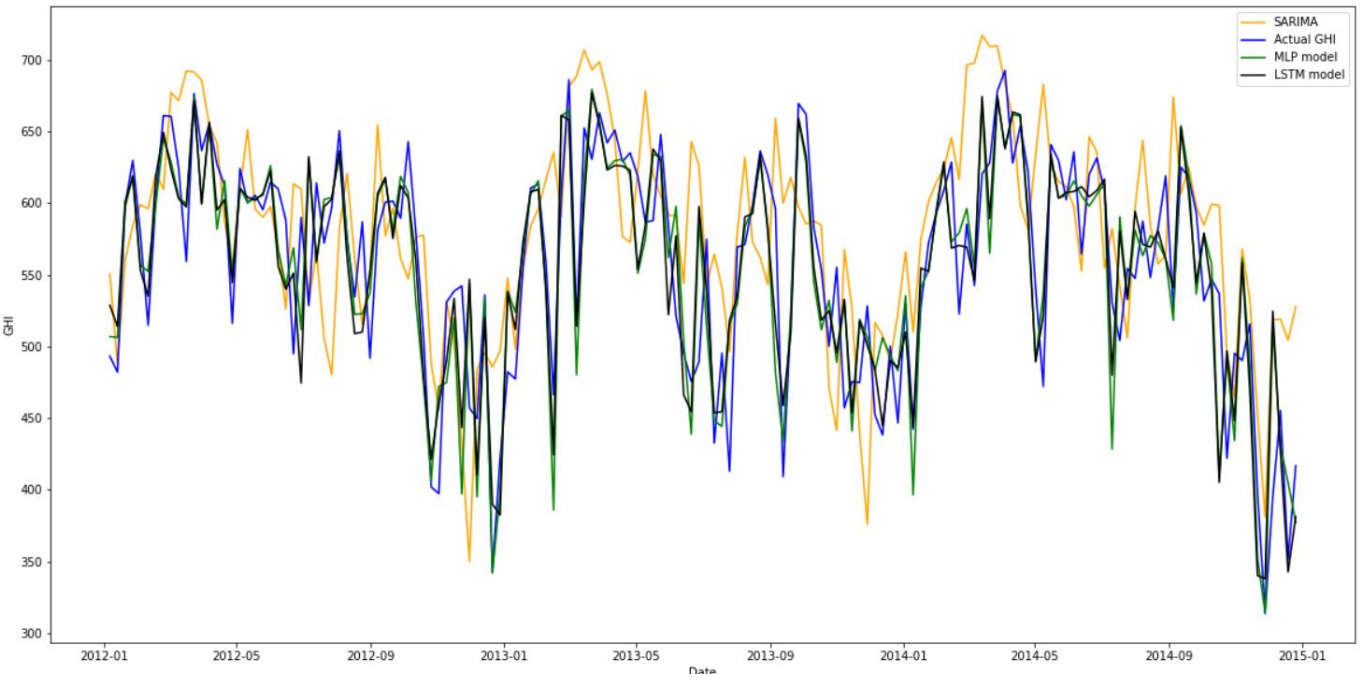


Figure 33 Results from the prediction models for the training dataset

6.5.3 Daily dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled daily.

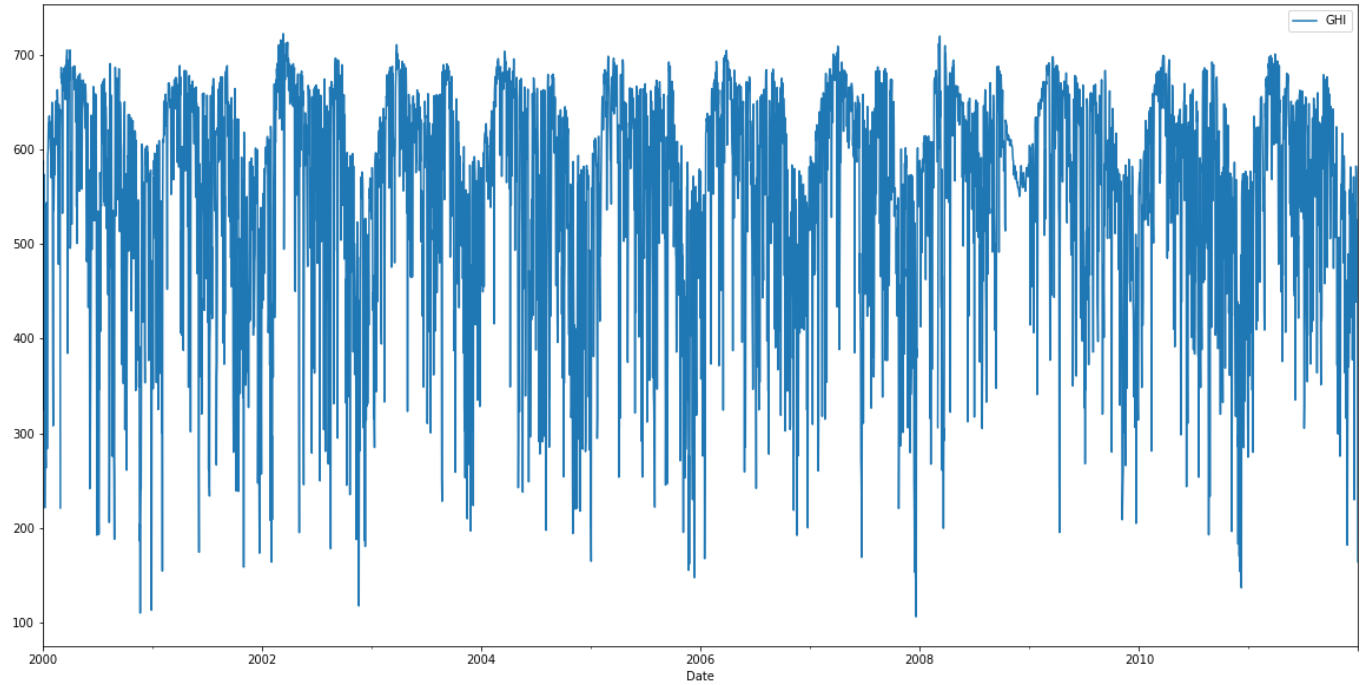


Figure 34 Daily Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[365] gives us the best results with MAPE = 48.33 per cent and RMSE = 308.57(Due to significantly worse results the predictions from SARIMA models are not shown in the graph). The graph below shows the results from MLP model with MAPE = 0.012 per cent and RMSE = 0.203. The LSTM model predictions have MAPE = 0.011 and RMSE = 0.0718. Thus for the Daily dataset of Tamil Nadu, LSTM model performs better.

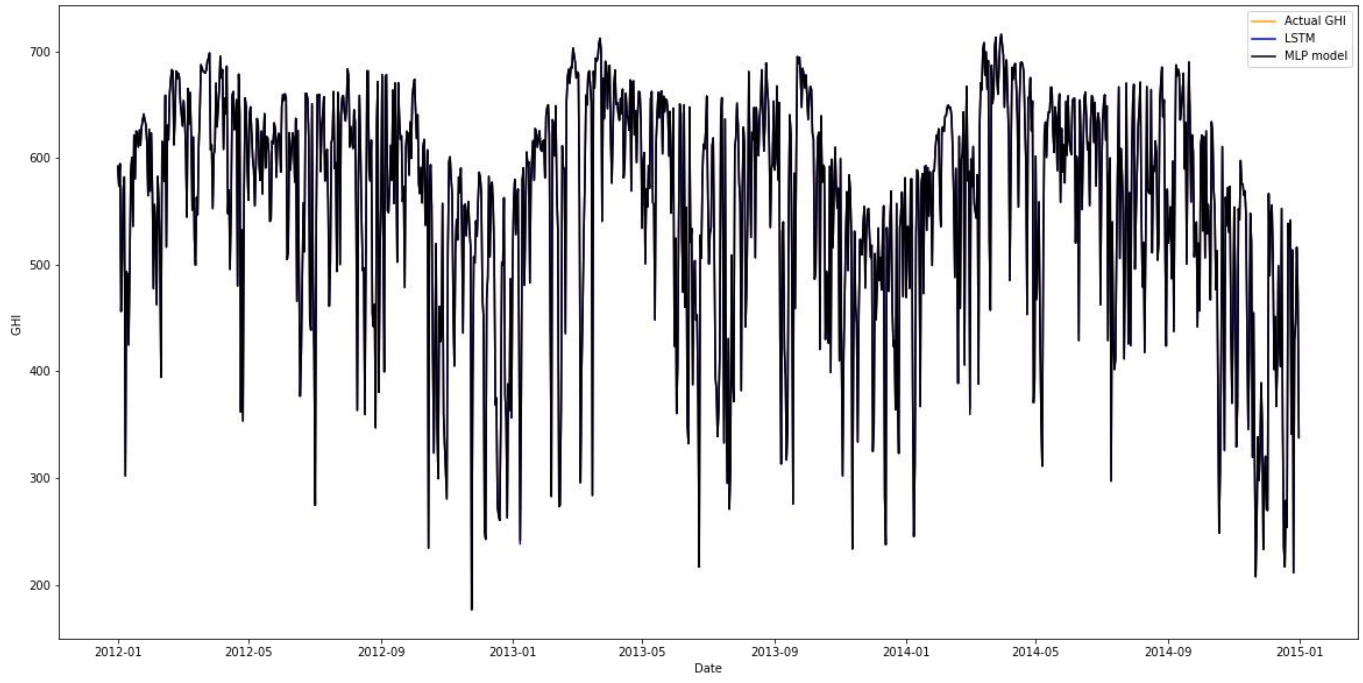


Figure 35 Results from the prediction models for the training dataset

The LSTM model, MLP model and the actual GHI values in the above graph appear to overlap because the predictions and the actual values differ significantly less.

6.6 Telangana

6.6.1 Monthly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled monthly.

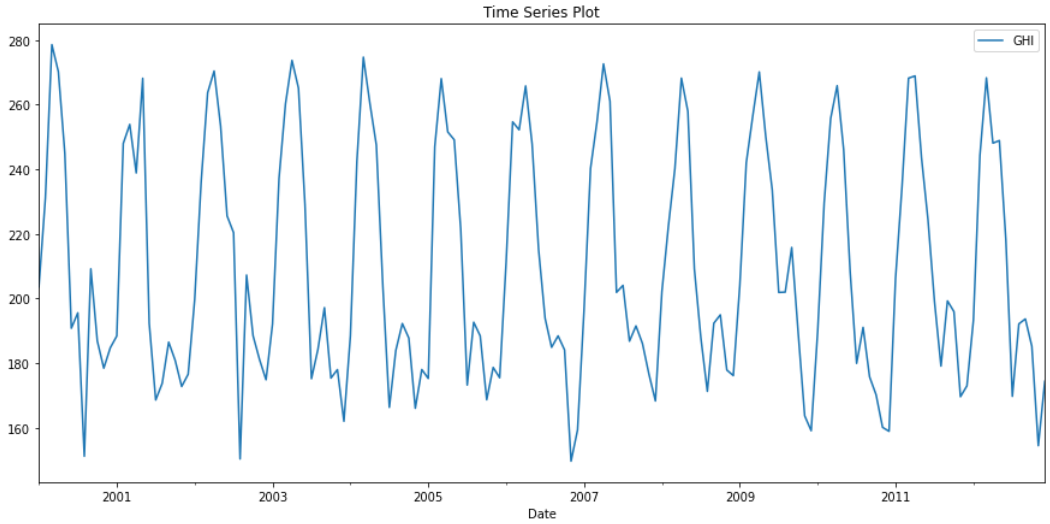


Figure 36 Monthly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 1)(1,0,1)[12] gives us the best results with MAPE = 5.45 per cent and RMSE =13.28. The graph below shows the results from MLP model with MAPE = 6.92 per cent and RMSE = 17.14. The LSTM model predictions have RMSE = 43.808 and MAPE = 14.58 (Since the LSTM model gives significantly worse results, it is not shown in the graph below). Thus, for the monthly dataset of Telangana, SARIMA model performs best.

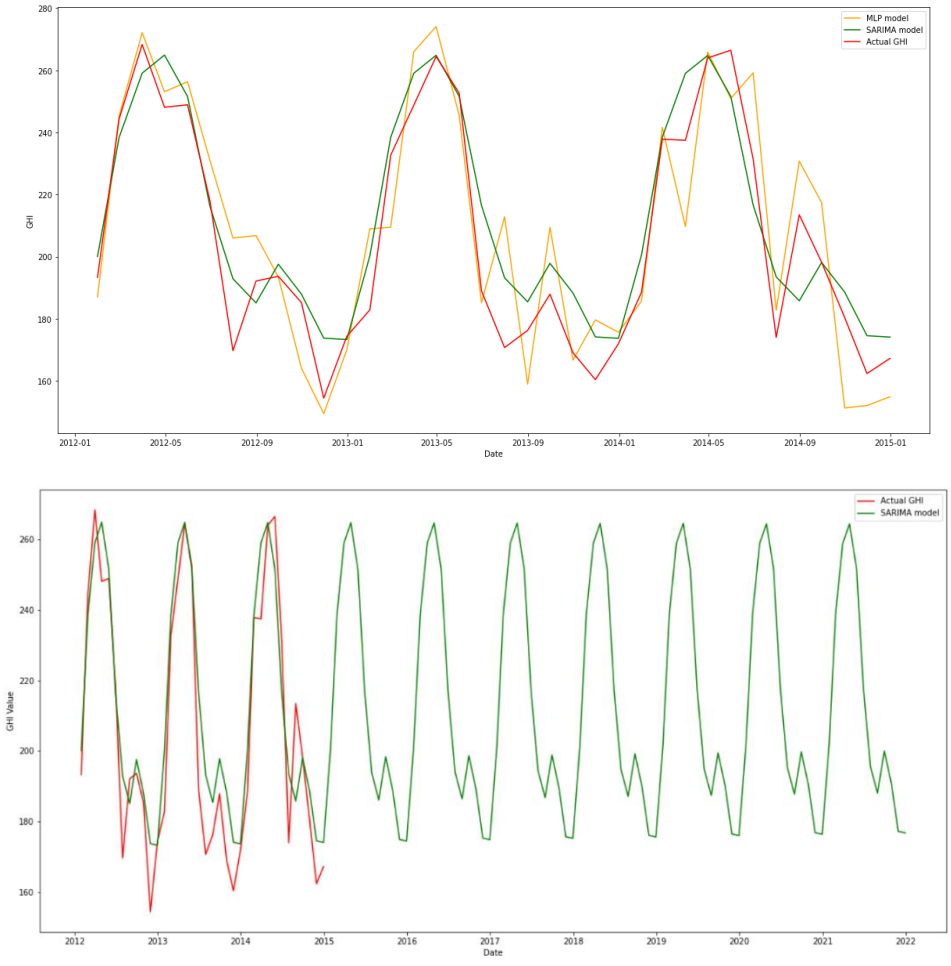


Figure 37 Results from the prediction models for the training dataset and extrapolation of SARIMA model till 2022

6.6.2 Weekly dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled weekly.

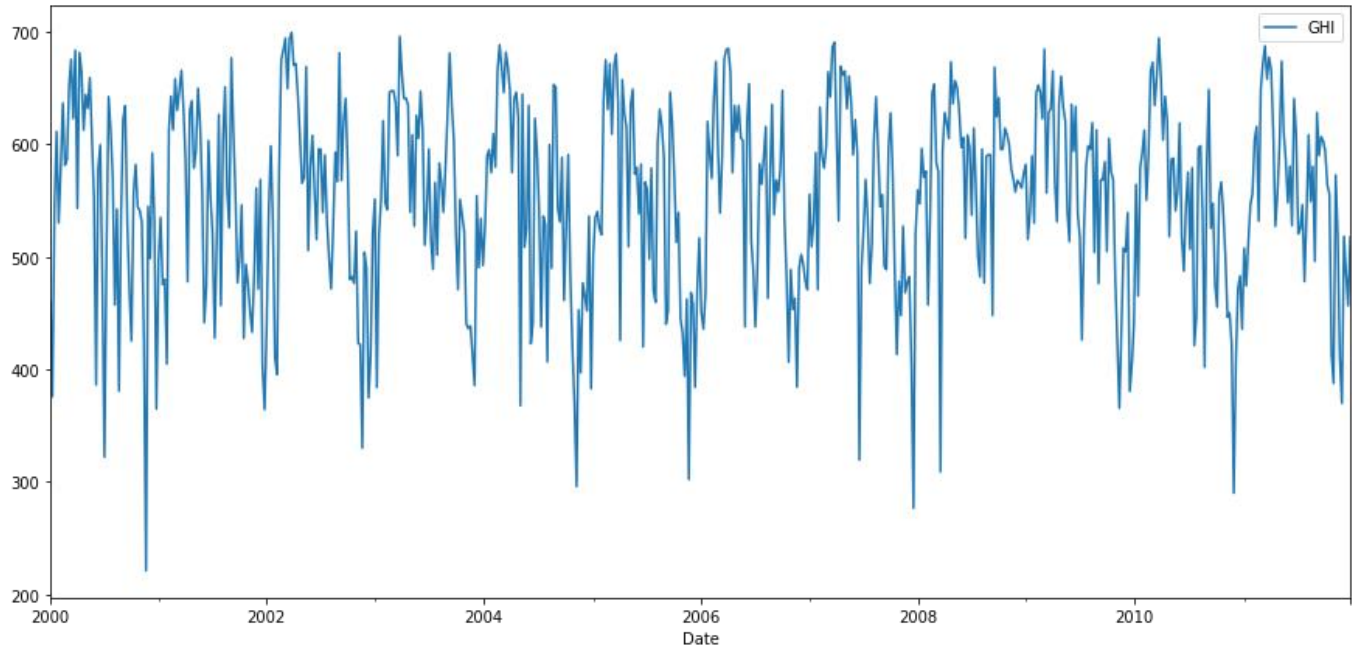


Figure 38 Weekly Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1,1,0)(1,1,0)[52] gives us the best results with MAPE = 17.28 per cent and RMSE = 38.87. The graph below shows the results from MLP model with MAPE = 9.26 per cent and RMSE = 24.73. The LSTM model predictions have MAPE = 8.54 and RMSE = 22.40. Thus for the weekly dataset of Telangana LSTM model performs better.

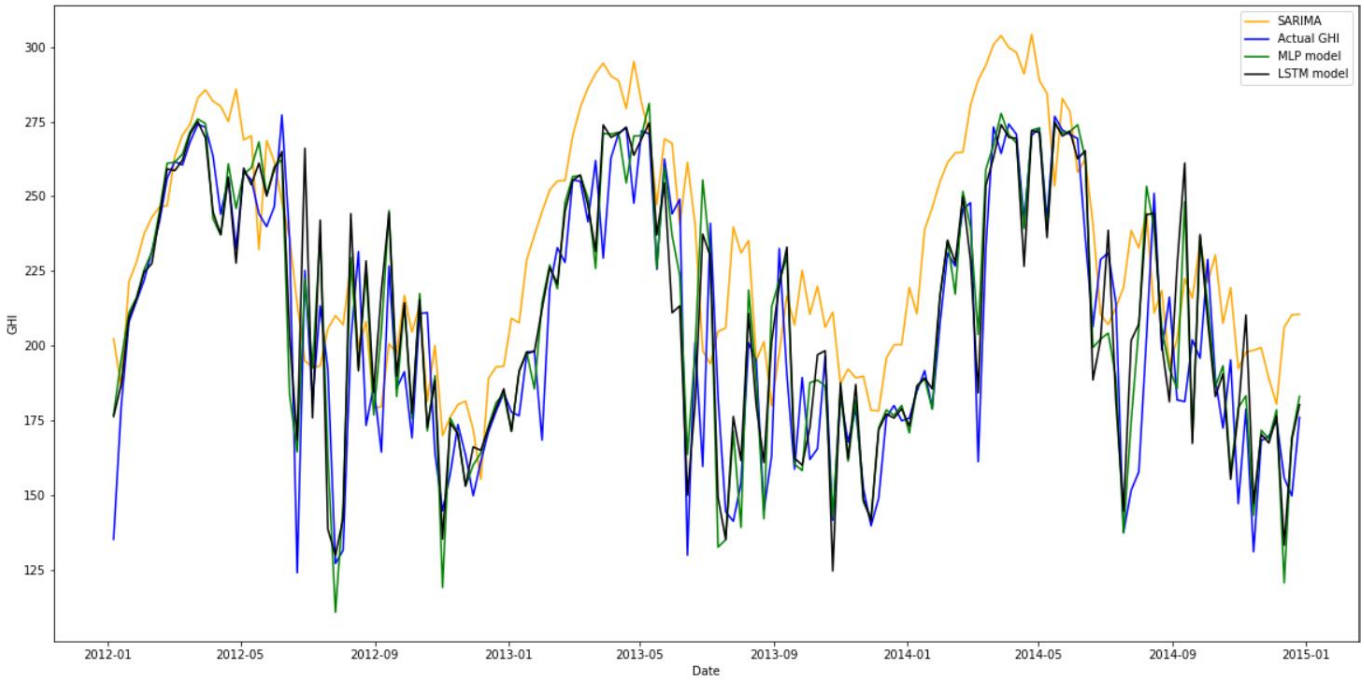


Figure 39 Results from the prediction models for the training dataset

6.6.3 Daily dataset

The train set used for analysis is from January 1, 2000- December 31, 2011, sampled daily.

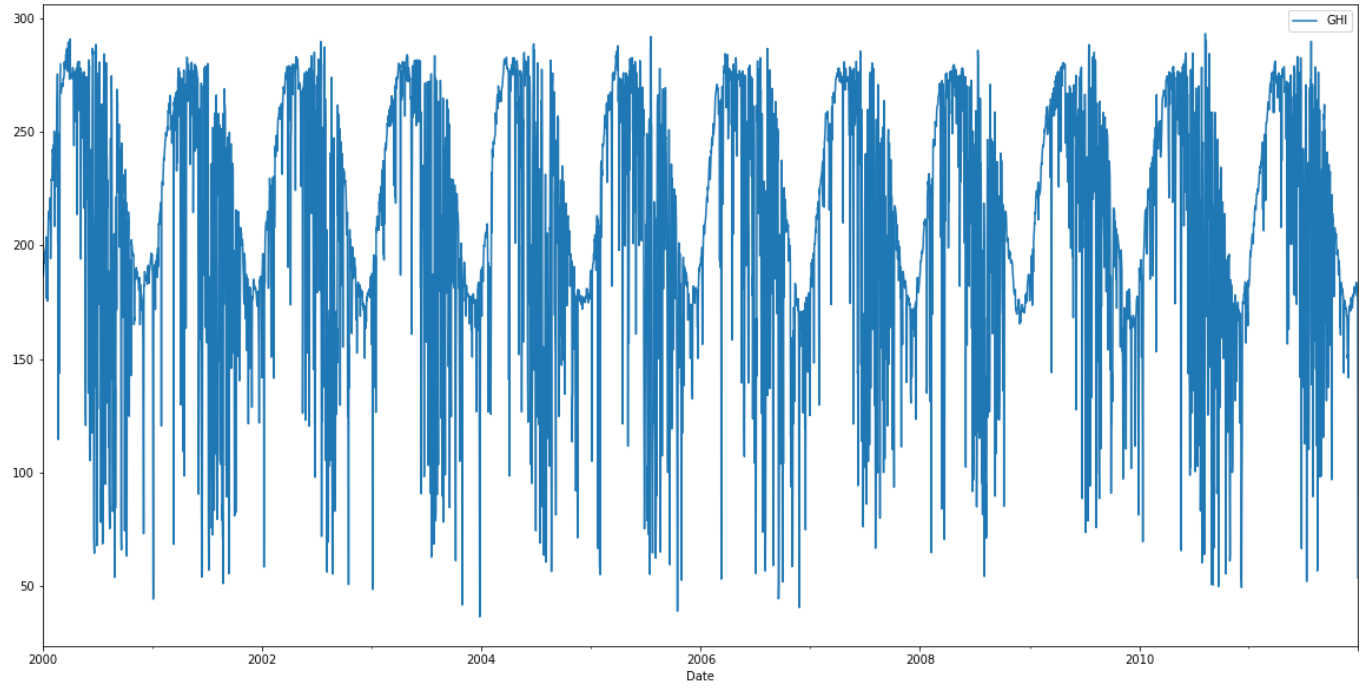


Figure 40 Daily Training Dataset

The testing dataset is from January 1, 2012 - December 31, 2014. The results from the Auto Arima function shows that the SARIMA (1, 1, 0)(1,1,0)[365] gives us the best results with MAPE = 41.65 per cent and RMSE = 90.49. The graph below shows the results from MLP model with MAPE = 0.019 per cent and RMSE = 0.0411. The LSTM model predictions have MAPE = 0.017 and RMSE = 0.036. Thus for the Daily dataset of Telangana, LSTM model performs.

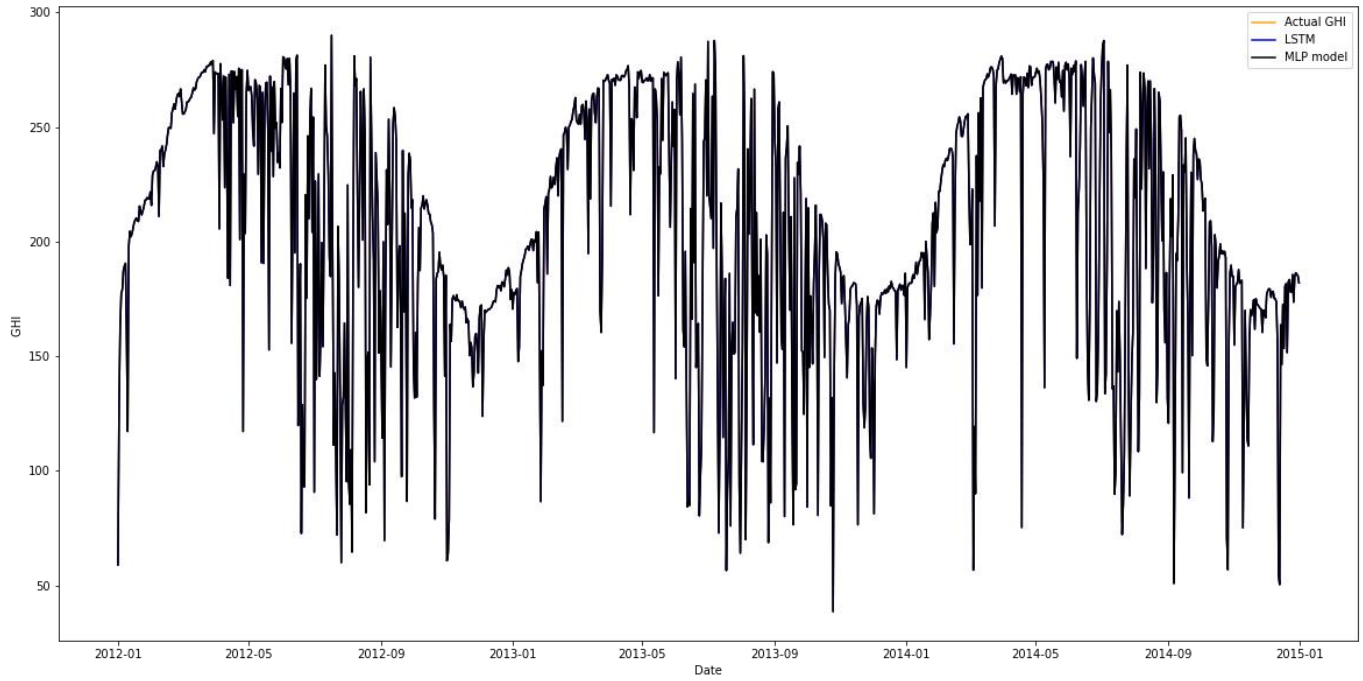


Figure 41 Results from the prediction models for the training dataset

The LSTM model, MLP model and the actual GHI values in the above graph appear to overlap because the predictions and the actual values differ significantly less.

The tables below summarize the best models and their MAPE and RMSE values for the 6 states.

	MONTHLY DATASET		
STATE	**MODEL**	**MAPE**	**RMSE**
KARNATAKA	SARIMA(1,1,1)(1,0,1)[12]	2.83	8.57
ANDHRA PRADESH	SARIMA(1,1,0)(1,0,1)[12]	2.53	16.82
RAJASTHAN	SARIMA(1,1,1)(1,0,1)[12]	2.1	14.713
GUJARAT	MLP	4.12	21.24
TAMIL NADU	SARIMA(1,1,1)(1,0,0)[12]	4.59	29.46
TELANGANA	SARIMA(1,1,1)(1,0,1)[12]	5.45	13.28

	WEEKLY DATASET		
STATE	**MODEL**	**MAPE**	**RMSE**
KARNATAKA	LSTM	5.78	17.14
ANDHRA PRADESH	LSTM	4.60	35.089
RAJASTHAN	LSTM	2.98	23.99
GUJARAT	LSTM	4.06	16.12
TAMIL NADU	LSTM	6.61	43.74
TELANGANA	LSTM	8.54	22.4

	DAILY DATASET		
STATE	**MODEL**	**MAPE**	**RMSE**
KARNATAKA	MLP, LSTM	0.16	0.399,0.443
ANDHRA PRADESH	MLP	0.0054	0.0741
RAJASTHAN	MLP	0.0097	0.078
GUJARAT	LSTM,MLP	0.015	0.0444,0.045
TAMIL NADU	LSTM	0.011	0.0718
TELANGANA	LSTM	0.017	0.036

6.7 SARIMA-MLP Hybrid Models

For Hybrid Models, the MLP model and the SARIMA model have been hybridized on the Weekly dataset. The best predictions from both Models are manually picked to construct the Hybrid Model. Their results are discussed below:

6.7.1 Karnataka

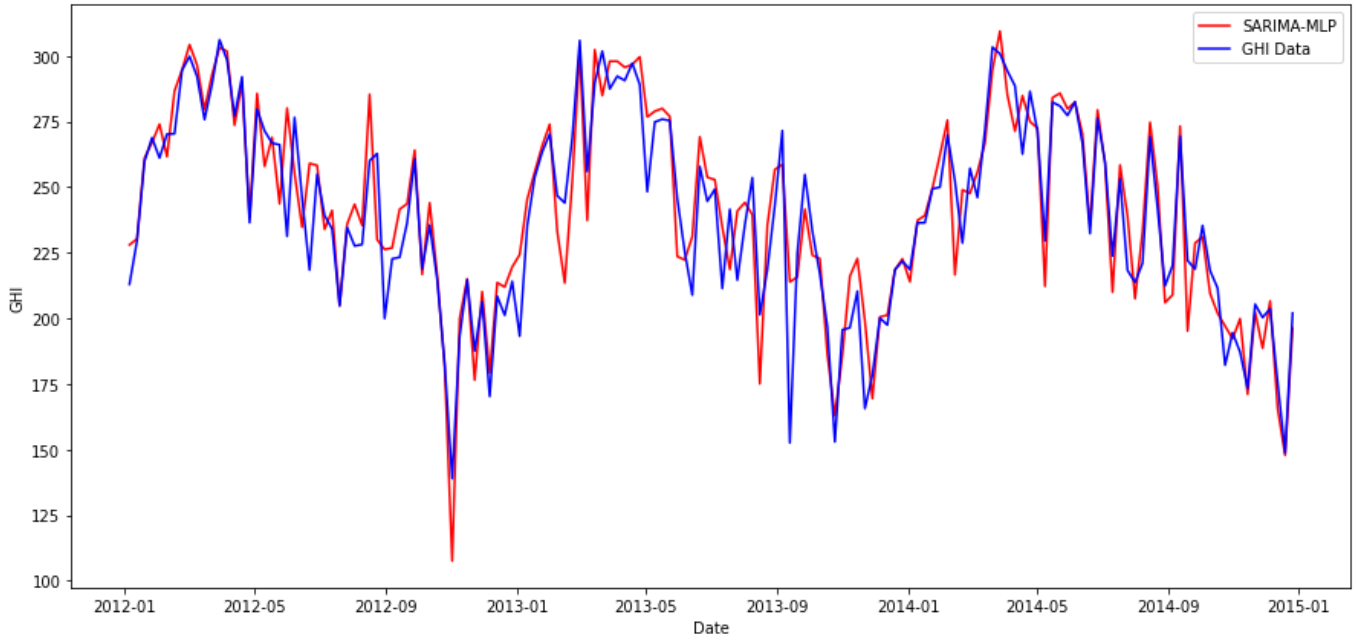


Figure 42 Predictions from SARIMA-MLP model

The SARIMA-MLP model for Karnataka dataset has MAPE value of 4.65 per cent and RMSE value of 14.44 which performs better than the LSTM model.

6.7.2 Andhra Pradesh

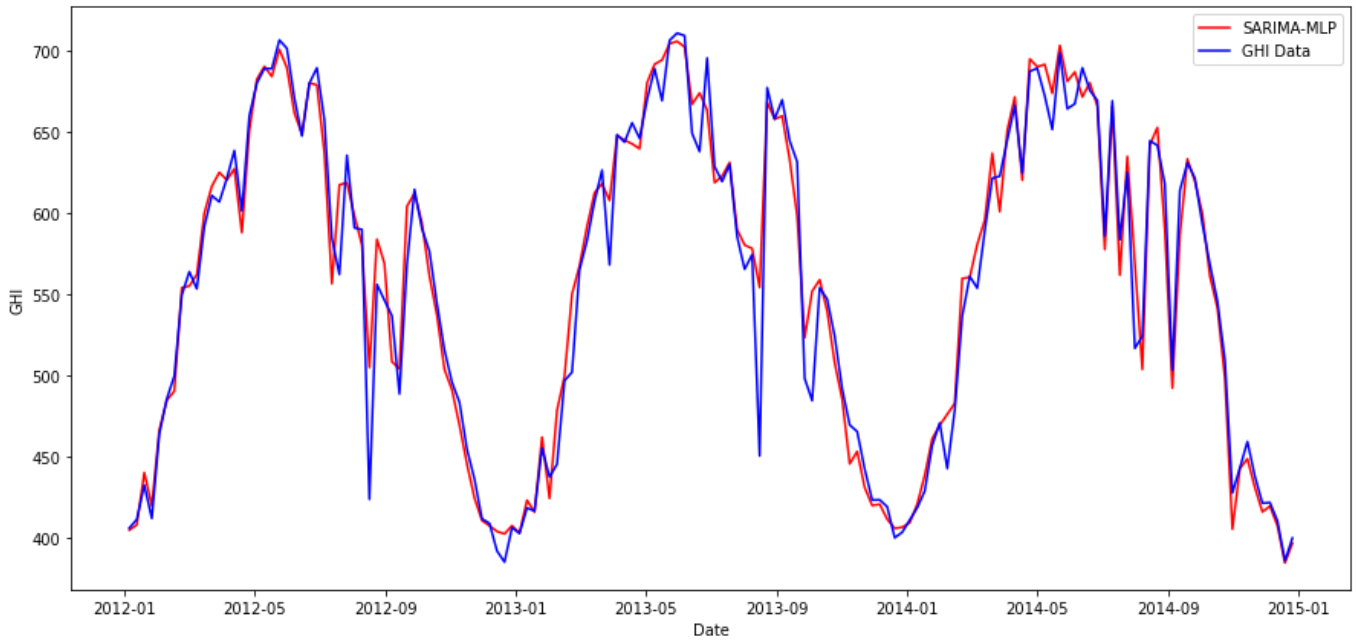


Figure 43 Predictions from SARIMA-MLP model

The SARIMA-MLP model for Andhra Pradesh dataset has MAPE value of 2.31 per cent and RMSE value of 19.25 which performs better than the MLP model.

6.7.3 Rajasthan

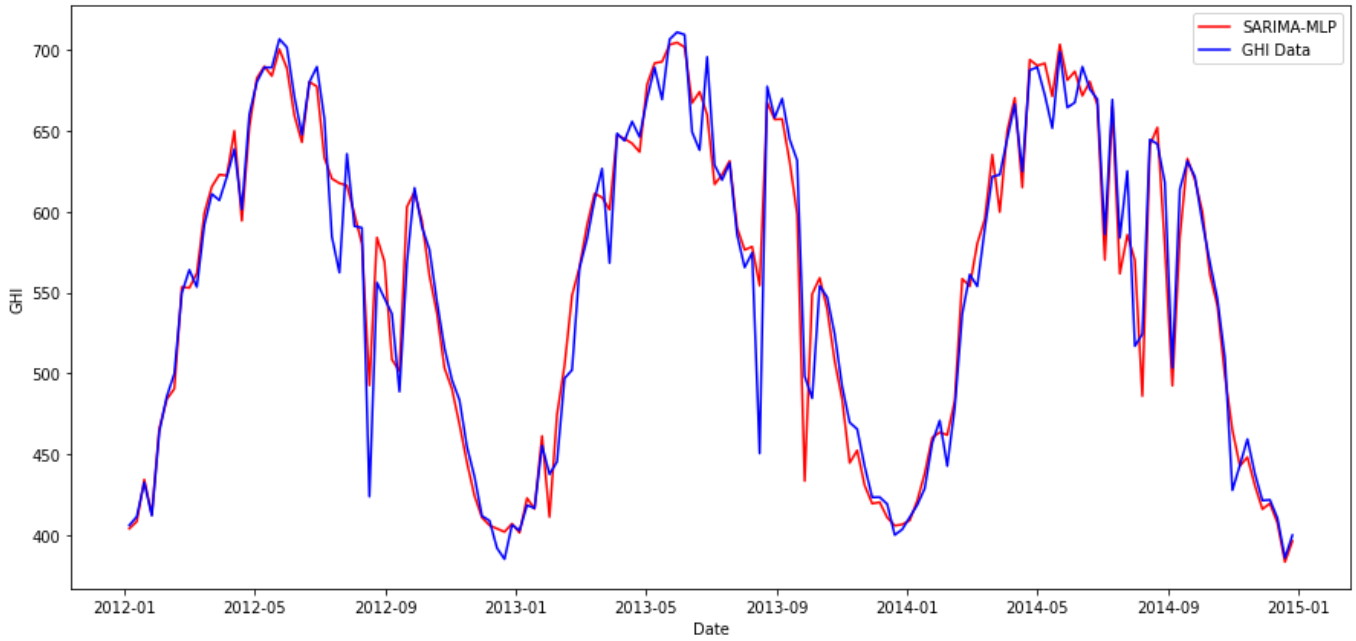


Figure 43 Predictions from SARIMA-MLP model

The SARIMA-MLP model for Rajasthan dataset has MAPE value of 2.47 per cent and RMSE value of 20.21 which performs better than the LSTM model.

6.7.4 Gujarat

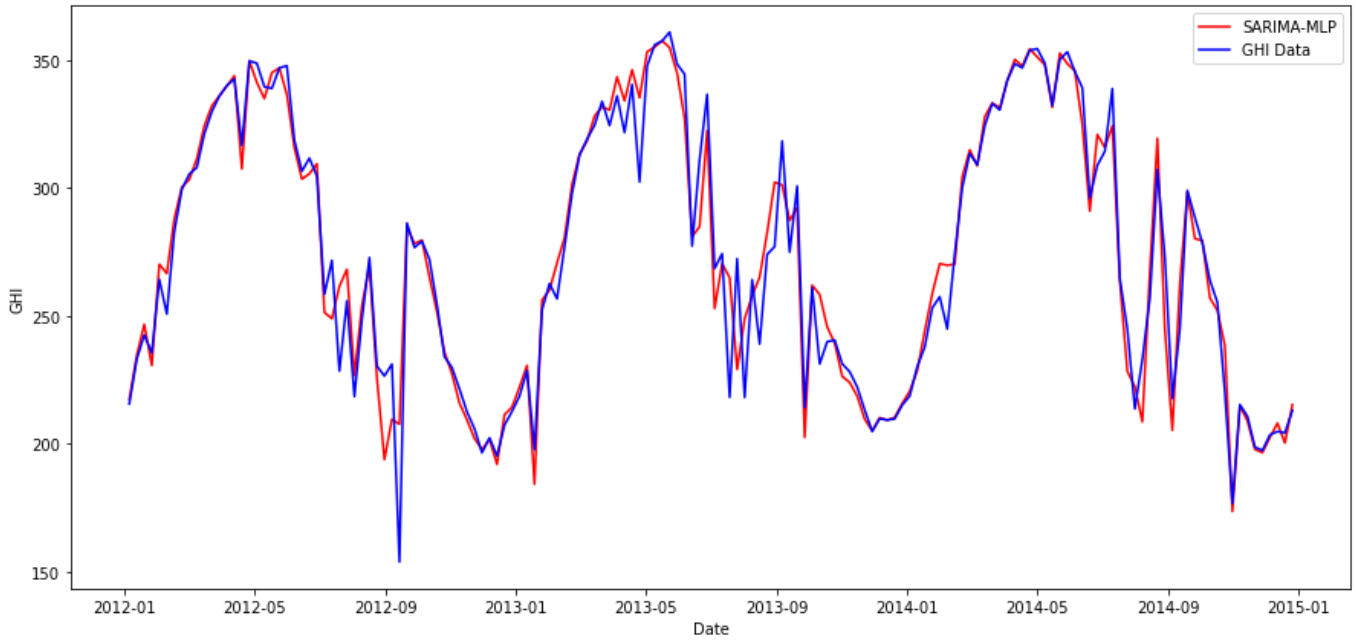


Figure 43 Predictions from SARIMA-MLP model

The SARIMA-MLP model for Gujarat dataset has MAPE value of 2.99 per cent and RMSE value of 12.19 which performs better than the LSTM model.

6.7.5 Tamil Nadu

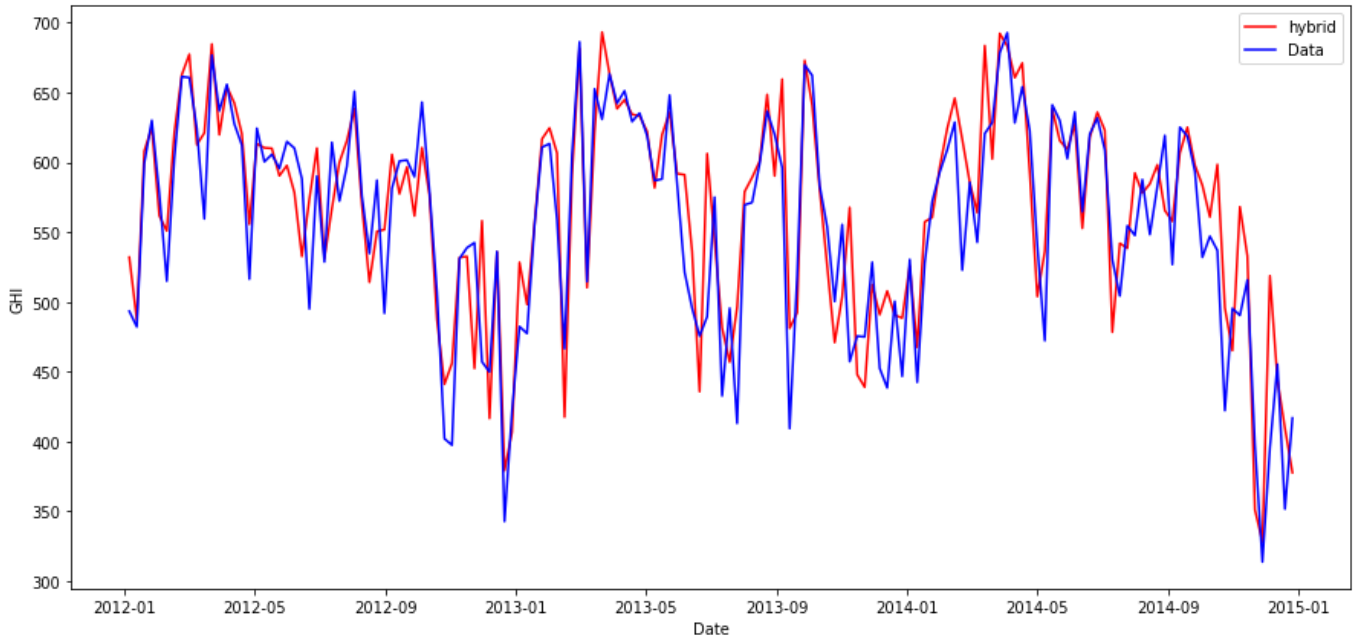


Figure 43 Predictions from SARIMA-MLP model

The SARIMA-MLP model for Tamil Nadu dataset has MAPE value of 5.53 per cent and RMSE value of 31.92 which performs better than the LSTM model.

6.7.6 Telangana

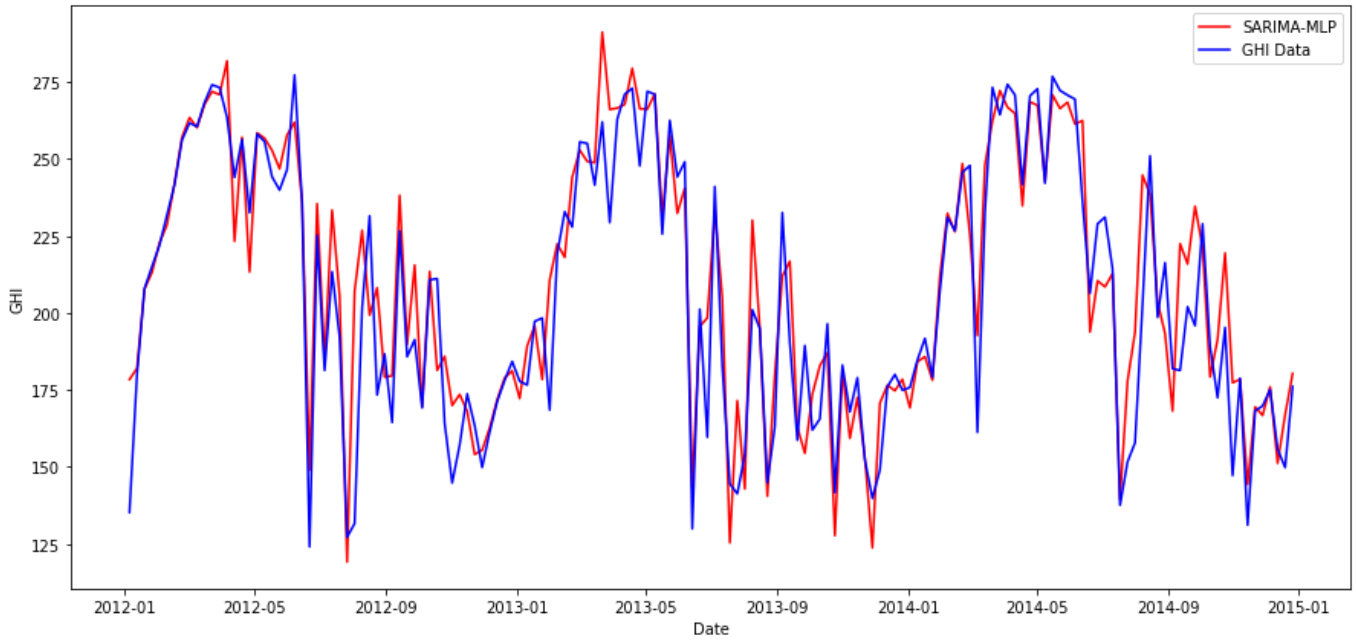


Figure 43 Predictions from SARIMA-MLP model

The SARIMA-MLP model for Telangana dataset has MAPE value of 6.59 per cent and RMSE value of 17.19 which performs better than the LSTM model.

7 Observations and Conclusions

1. GHI index is highly correlated to DHI, DNI, Clearsky GHI, Clearsky DNI, Clearsky DHI and Temperature.
2. From the time series decomposition, we can clearly see there is a clear seasonal component in GHI.
3. The solar sites situated in Rajasthan, Andhra Pradesh and Tamil Nadu had higher median GHI values than solar sites in Karnataka, Gujarat and Telangana.
4. Solar site situated in Andhra Pradesh had least coefficient of variance followed by Rajasthan, but the number of outliers in Andhra Pradesh dataset is greater than Rajasthan. Gujarat and Karnataka had almost identical coefficient of variance.Telangana had the greatest coefficient of variance.
5. The SARIMA model performs best for the monthly dataset for all states except Gujarat which MLP model gave best predictions. The most accurate results were obtained for solar site located in Rajasthan.
6. For the weekly dataset, LSTM models outperform all the other models. Again the most accurate predictions were obtained for Rajasthan.
7. For the daily dataset, the predictions of LSTM and MLP models give similar results. Solar sites in Andhra Pradesh and Rajasthan had most accurate results from MLP model, while Tamil Nadu and Telangana favoured LSTM model. Karnataka and Gujarat had similar results from LSTM and MLP models. The most accurate results were obtained from the solar site in Andhra Pradesh.
8. The SARIMA-MLP hybrid models outperformed LSTM model for weekly dataset of all the states.
9. Conducting visual analysis on the SARIMA-MLP model of Rajasthan shows that majority of the occurrences where SARIMA predictions are preferred over MLP predictions occurred during the Monsoon season of Rajasthan. (Scatter plot in Appendix).
10. The MLP models had the least training times in comparison to the other models while the LSTM model had the greatest training time for the above datasets.

8 Future Work

1. The models for monthly, weekly and hourly datasets have been evaluated only till December 2014 due to the non-avalability of dataset for 2015-2021. So the future work will invlove evaluating the performance during recent years.
2. The SARIMA-MLP model discussed in this project is for weekly dataset. Similar concept can be used to construct hybrid models for Monthly dataset to improve performance in some states.
3. The visual analysis of hybrid models shows that cloud cover during monsoon season affects the choice of the model, thus a more statistical approach based on Cloud Cover Index can be used to decide the choice for the model. The Cloud Cover Index can be calculated using GHI index and Clearsky GHI index.
4. The LSTM model discussed above is the Vanilla LSTM model, there are several other LSTM variants which can be used to improve performance and to decrease the training time.

9 References

1 National electricity plan (2016), Volume 1, Generation, Central Electricity Authority (CEA),Ministry of Power, GOI .Available at http://www.cea.nic.in/reports/committee/nep/nep_dec.pdf .

2 Power Sector at a Glance All India. (2020, 02 02). Retrieved from Ministry of Power GoI: https://powermin.nic.in/en/content/power-sector-glance-all-india

3 *Brown to Green: The G20 Transition Towards A Net-Zero Emissions Economy 2019 report.* (2020, 02 02). Retrieved from Enerdata 2019: https://www.climate-transparency.org/wp-content/uploads/2019/11/Brown-to-Green-Report-2019.pdf

4 Each Country's Share of CO2 Emissions. (2020, 02 02). Retrieved from Union of Concerned Scientists: https://www.ucsusa.org/resources/each-countrys-share-co2-emissions.

5 World Energy Scenarios Composing energy futures to 2050 (2013), World energy Council. https://www.worldenergy.org/wp-content/uploads/2013/09/World-Energy-Scenarios_Composing-energy-futures-to-2050_Full-report.pdf .

6 India’s potential for integrating solar and on- and offshore wind power into its energy system. (2020, 02 02). Retrieved from Nature Communications: https://www.nature.com/articles/s41467-020-18318-7

7 Annual Report Ministry of New and Renewable Energy 2019. (2020, 02 05). Retrieved from Ministry of New and Renewable Energy, GoI: https://mnre.gov.in/img/documents/uploads/file_f-1597797108502.pdf

8 Physical Progress. (2020, 02 05). Retrieved from Ministry of New and Renewable Energy: https://mnre.gov.in/the-ministry/physical-progress

9 Vashishtha, D. S. (2021, 02 18). Differentiate between the DNI, DHI and GHI? Retrieved from First Green Consulting: https://firstgreenconsulting.wordpress.com/2012/04/26/differentiate-between-the-dni-dhi-and-ghi/

10 Brockwell, P. J., Davis, R. A., & Fienberg, S. E. (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

11 H.M. Diagne, P. Lauret, M. David, Solar irradiation forecasting: state-of-the-art and proposition for future developments for small-scale insular grids, in: n.d. https://hal.archives-ouvertes.fr/hal-00918150/document (Accessed 15 February 2021)

12 Wu Ji, Keong Chan Chee, Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN, Solar Energy, Volume 85, Issue 5, May 2011, Pages 808-817.

13 Al-Sadah, F. H., Ragab, F. M., & Arshad, M. K. (1990). Hourly solar radiation over Bahrain. *Energy*, *15*(5), 395-402.

14 Reikard, G. (2009). Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, *83*(3), 342-349.

15 Hugo T. C. Pedro and Carlos F. M. Coimbra, “Assessment of forecasting techniques for solar power production with no exogenous inputs,” Solar Energy,vol. 86, no. 7, pp. 2017–2028, 2012.

16 Dazhi Yang, Panida Jirutitijaroen, and Wilfred M. Walsh, “Hourly solar irradiance time series forecasting using cloud cover index,” Solar Energy, vol. 86, no. 12, pp. 3531–3543, 2012.

17 Yanting Li, Yan Su, and Lianjie Shu, “An ARMAX model for forecasting the power output of a grid connected photovoltaic system, “Renewable Energy, vol.66, pp. 78–89, 2014.

18 P. Lauret, C. Voyant, T. Soubdhan, M. David, P. Poggi, A benchmarking of machine learning techniques for solar radiation forecasting in an insular context, Sol. Energy 112 (2015) 446e457.

19 Wu Ji, Keong Chan Chee, Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN, Solar Energy, Volume 85, Issue 5, May 2011, Pages 808-817.

20 Reikard, G. Predicting solar radiation at high resolutions: A comparison of time series forecasts.Sol. Energy2009,83, 342–349

21 Farhath, Z.A.; Arputhamary, B.; Arockiam, D.L. A Survey on ARIMA Forecasting Using Time Series Model.Int. J. Comput. Sci. Mobile Comput.2016,5, 104–109.

22 M. Bouzerdoum, A. Mellit, A. Massi Pavan, A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant, Sol. Energy 98 (PC) (2013) 226e235

23 Y.-K. Wu, C.-R. Chen, H. Abdul Rahman, A novel hybrid model for short-term forecasting in PV power generation, Int. J. Photoenergy (2014)

24 A. Gensler, J. Henze, B. Sick and N. Raabe, "Deep Learning for solar power forecasting — An approach using Auto Encoder and LSTM Neural Networks," *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary, 2016, pp. 002858-002865, doi: 10.1109/SMC.2016.7844673.

25 Jie Shi, Wei-Jen Lee, Yongqian Liu, Yongping Yang, and Peng Wang, "Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines,"IEEE Transactions on Industry Applications,vol. 48, no. 3, pp. 1064–1069, 2012.

26 Mashud Rana, Irena Koprinska, and Vassilios G. Agelidis, "2D-interval fore-casts for solar power production," Solar Energy, vol. 122, pp. 191–203, 2015.

27 Adel Mellit, Alessandro M. Pavan, and Mohamed Benghanem, "Least squares support vector machine for short-term prediction of meteorological time series," Theoretical and Applied Climatology, vol. 111, no. 1, pp. 297–307, 2013.

28 Makbul A. M. Ramli, Ssennoga Twaha, and Yusuf A. Al-Turki, "Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study," Energy Conversion and Management, vol. 105, pp. 442–452, 2015.

29 Betul B. Ekici, "A least squares support vector machine model for prediction of the next day solar insolation for effective use of PV systems," Measurement, vol. 50, no. 1, pp. 255–262, 2014.

30 Kasra Mohammadi, Shahaboddin Shamshirband, Chong W. Tong, Muhammad Arif, Dalibor Petkovi, and Sudheer Ch, "A new hybrid support vector machine wavelet transform approach for estimation of horizontal global solar radiation," Energy Conversion and Management, vol. 92, pp. 162–171, 2015.

31 Lanre Olatomiwa, Saad Mekhilef, Shahaboddin Shamshirband, Kasra Moham-madi, Dalibor Petkovi, and Ch Sudheer, "A support vector machine firefly algorithm-based model for global solar radiation prediction," Solar Energy, vol. 115, pp. 632–644, 2015.

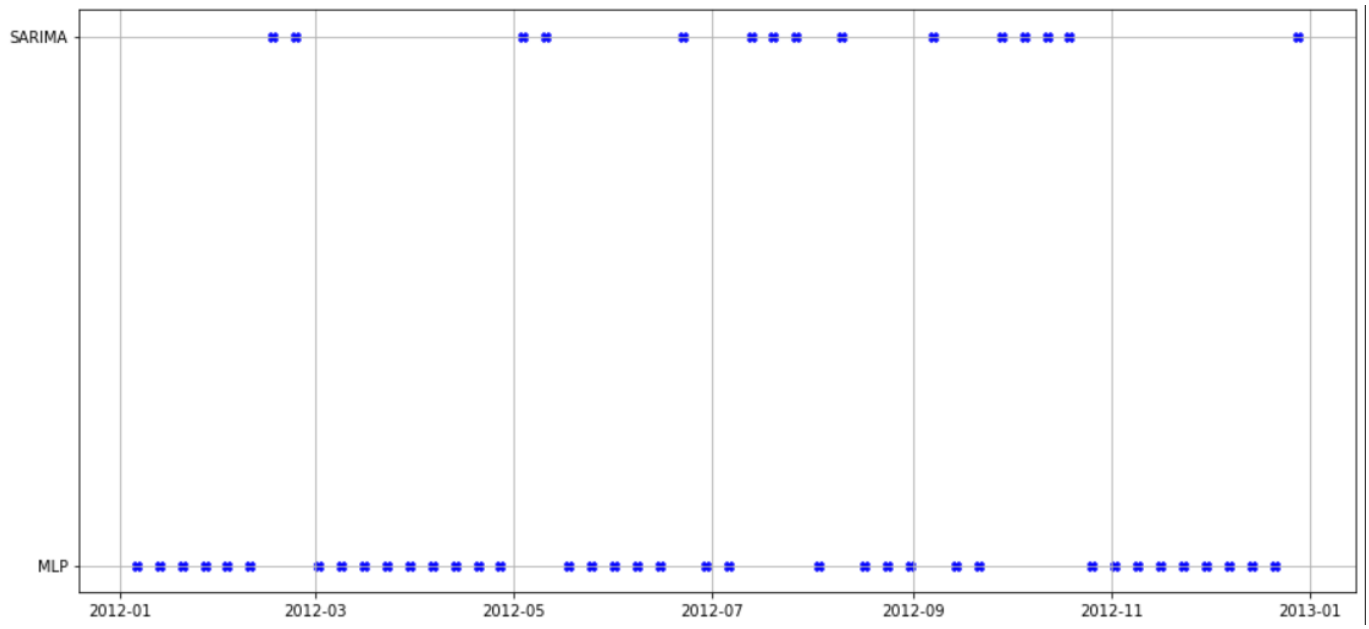
32 Cao, J. and Cao, S. Study of forecasting solar irradiance using neural networks with pre-processing sample data by wavelet analysis. Energy, 31(15):3435 – 3445, 2006.ISSN 0360-5442. doi: http://dx.doi.org/10.1016/j.energy.2006.04.001. URL http://www.sciencedirect.com/science/article/pii/S0360544206001009. {ECOS} 2004 - 17th International Conference on Efficiency, Costs, Optimization, Simulation, and Environmental Impact of Energy on Process Systems 17th International Conference on Efficiency, Costs, Optimization, Simulation, and Environmental Impact of Energy on Process Systems.

33 Dong, Z., Yang, D., Reindl, T., and Walsh, W. M. Satellite image analysis and a hybrid ESSS/ANN model to forecast solar irradiance in the tropics. Energy Conversion and Management, 79(0):66 – 73, 2014. ISSN 0196-8904. doi: http://dx.doi.org/10.1016/j.enconman.2013.11.04.

34 Yona, A., Senjyu, T., Saber, A., Funabashi, T., Sekine, H., and Kim, C.-H. Application of neural network to 24-hour-ahead generating power forecasting for pv system. In Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE, pages 1 – 6, July 2008. doi: 10.1109/PES.2008.4596295.

35 Wu, Y.-K., Chen, C.-R., and Abdul-Rahman, H. A novel hybrid model for short-term forecasting in PV power generation. International Journal of Photoenergy, 2014(0):1 –9, 2014. doi: http://dx.doi.org/10.1155/2014/569249.

10 Appendix



Scatter Plot depicting the distribution of predictions of MLP and SARIMA models in the hybrid model of Rajasthan

The above scatter plot shows the distribution of predictions from MLP and SARIMA models in the hybrid model . From the plot we can observe than 2/3rd of the occurences of SARIMA model happen during the monsoon season of the State during mid June and late September and retreating monsoons during early October.