Time Series Forecast Model for Hourly Solar Power Near Duke University

Shufan Xia

1 Summary

Forecasting solar radiation is critical for incorporating green energy in energy grids reliably. In this analysis, our goal is to propose an Autoregressive Integrated Moving Average (ARIMA) model to predict the hourly solar irradiance measured by GHI for the area around Duke University. This analysis reviews one year length of data from 2020 Physical Solar Model (PSM) v3 Typical Meteorological Year (TMY) dataset. On comparing AIC and Mean Absolute Scaled Error (MASE), the result shows a Seasonal and Trend decomposition using Loess (STL) model using seasonal naive model for the seasonal component and ARIMA(5,1,2) for the remaining component captures the diurnal cycle in the data and provides better forecasting accuracy for one-month and one-week ahead forecasts.

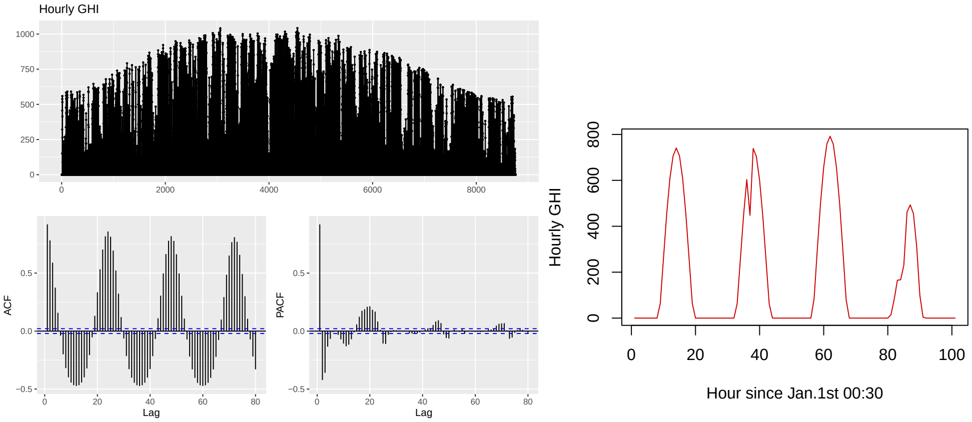
2 Introduction

Energy is operated with power grids that connect power plants of various sources to households and facilities. Accurate wind and solar forecasting enhance the value of renewable energy by improving the reliability and economic feasibility of these resources. Predicting solar energy has two components: having a reliable weather model and simulating photo-voltaic power based on a set of given weather conditions. Among all weather variables, solar irradiance directly affects the amount of solar energy available for generation. Therefore, predicting solar irradiance over time is a critical step in energy forecasting. Solar irradiance is often measured in direct normal(DNI), diffuse horizontal irradiance (DHI), and global horizontal irradiance (GHI). Most Photovoltaic systems are non-concentrating, meaning they utilize both direct and diffuse sunlight to generate power. Thus, predicting GHI is required.

Time series models are often used in solar power forecast in research. This analysis reviews the historical solar irradiance and other meteorological measurements in the area around Duke University and proposes a forecasting time series model. A time series ARMA model for hourly solar irradiance forecasting is proposed, and its performance is evaluated. The fluctuation of solar irradiance follows the diurnal cycle and the change in the relative position of the Sun over a year. Therefore, seasonality and trend in hourly GHI over time are expected, and a Seasonal-ARIMA and Seasonal and Trend decomposition using Loess(STL) model are also considered. The overall goal of this analysis is to investigate the seasonal trend of solar irradiance and propose a forecasting model for hourly GHI.

3 Data

This analysis reviews 2020 Physical Solar Model (PSM) v3 TMY data from the National Solar Radiation Database (NSRDB). NSRDB is managed and updated by a specialized team of forecasters at the National Renewable Energy Laboratory (NREL) supported by the U.S. Department of Energy's SunShot Initiative. PSM v3 TMY consists of hourly data of three measurements of solar irradiaance: global horizontal (GHI), direct normal(DNI), and diffuse horizontal irradiance (DHI), as well as 6 common meteorological measurements: dew point, surface albedo, wind direction, wind speed, temperature, and air pressure. Typical Meteorological Year (TMY) data are selected to best represent the average weather conditions of a specific location over multiple years in the past. Because weather conditions vary from year to year, to represent any long-term trend consistently, TMY data is used instead of data from any specific year. Any conclusion based on the TMY data is more transferable for forecasting any year in the future. The 2020 PSM3 v3 TMY data includes years from 1999 to 2020. It has 8730 entries, one entry for each hour of each day over a year. Each month in the year corresponds to a specific month from one of the years between 1999 and 2020. For example, January from 2002, February from 1999, and etc.. Table 1 in Appendix lists all the variables and their units contained in the data.



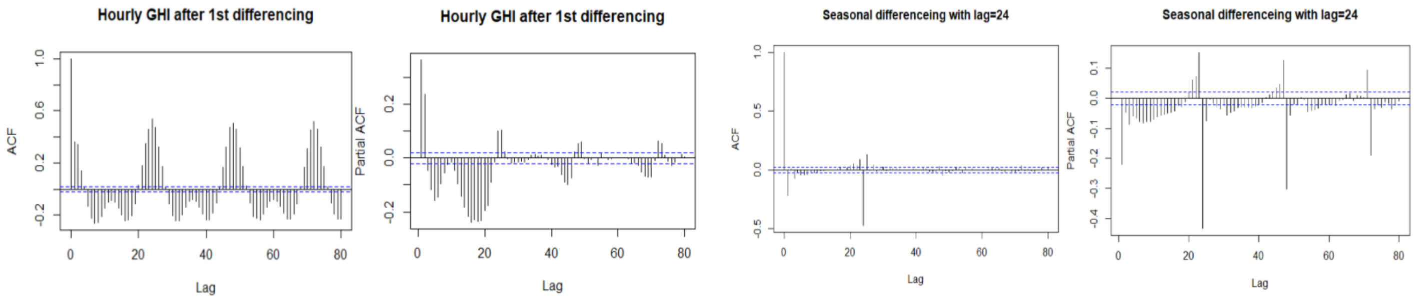
(a) (b)

Figure 1: Hourly GHI plotted over a year from PSM vs TMY data, ACF and PACF plots

Hourly GHI near Duke university over a year ranges from 0 to 1042.0 with a median of 0 and a mean of 189.4. This distribution is highly skewed right because 50% of the hourly GHI data observed is during the night, thus 0. To utilize a series of R functions built-in for time series analysis, the hourly GHI data is turned into a time series object with the ts() command. Fig 2.(a) shows the seasonal pattern of GHI over a year. Overall, the hourly GHI exhibits a parabolic trend. This is expected since sunshine is stronger during the summer in the Northern hemisphere. Zooming in on this time series (Fig 2.(b)) shows hourly GHI peaks every 24 hours, which mirrors the rise and fall of the Sun. In this report, the term "seasonality" follows the definition in Hyndman(2018). Seasonality occurs at a fixed frequency, while Hyndman refers to cyclicity as repeated patterns with non-fixed frequency. If the data were drawn from multiple years, cyclicity might be available. Therefore, the two observations above are considered as seasonality and trend, and suggest including "seasonality" is critical in building a forecasting model. The strong daily seasonality is also reflected in the ACF autocorrelation plot in Fig 2. The absolute values of ACF decrease slowly as the lags increases, which is due to the trend, while the sinusoidal oscillating shape is due to the daily seasonality. Since autocorrelation tapers off slowly, the hourly GHI data are non-stationary. Moreover, the p-value of the KPSS test is smaller than 0.05, indicating the data is not stationary. Therefore, differencing or transformation on hourly GHI data is required.

Both the Augmented Dickey-Fuller (ADF) and KPSS test suggest taking first-order differing makes hourly GHI stationery. This says if an ARIMA model on the original hourly GHI were used, the value of d in the model parameter would be 1. On the partial autocorrelation plot for the 1st order differencing data, partial autocorrelation decreases to below the threshold after 26 lags, suggesting only using an AR model would require lags about 26 (p=26). However, the partial correlation never dies off to zero, which means that the forecasting needs a combination of AR and MA models.

Despite both stationary tests confirming that taking the first-order differencing makes the data stationary, 1st order differing still shows seasonality because autocorrelation peaks at a seasonal lag period of 24. The autocorrelation and partial autocorrelation plot for seasonal difference with lag=24 on top of first-order differencing shown in Fig 2b. ADF and KPSS tests both indicate the data with



(a)

(b)

Figure 2: a) ACF and PACF plot on the data after 1-st order differencing b) after 1-st order differencing and lag=24 seasonal differencing.

first order and lag 24 seasonal differencing is stationary. On the ACF plot, the significant spike until lag 3 suggests a non-seasonal MA(3) component, and the significant spike at lag 24 suggests a seasonal MA(1) component, i.e. ARIMA(0,1,3)(0,1,1)24. Meanwhile, the partial autocorrelations plot shows a spike at every 24 lags. This together indicates adding AR to the non-seasonal and seasonal part of ARIMA is potentially necessary. Another method of fitting seasonal timer series is to use the STL decomposition method, which is dicussed in the next section.

As to the meteorological measurements included in the data, exploratory data analysis follows the same approach of multiple linear regression but does not consider the response variable, i.e. hourly GHI. Examining the distribution of surface albedo suggests this variable is discrete with 6 levels. Thus, surface albedo is redefined as a categorical variable. Temperature has a strong linear association with GHI compared to other variables since the temperature is correlated with the season, thus the overall available sunlight received. A multiple linear regression model including all the six main effects yields a significant p-value on the coefficients for all. This means that a forecasting model that combines multiple linear regression and an ARIMA model should include all six meteorological measurements..

4 Model

This section evaluates and compares several forecasting models for hourly GHI data. Since the first-order differencing data is stationary, the model building process starts with using a basic ARIMA model found by the auto.Arima() command. The model is modified sequentially to generate better forecasting accuracy. The subsequent models include a dynamic regression model combining multiple linear regression and an ARIMA model, a seasonal ARIMA model, and finally a STL decomposition model. Model evaluation is based on residual diagnostic tests, and model accuracy is assessed with AIC score. To evaluate the forecasting accuracy of a model, a one-month-ahead forecast is calculated. The training data uses the data from January to November. For each model proposed, the order of parameters and their values are estimated from the training data, then used to run forecasting. Evaluating forecasting accuracy uses marginal absolute scaled error (MASE) to measure percentage errors. MASE greater than one indicate that the forecast result is worse than using naïve method. Since the goal of fitting a series model is to forecast, the model fitting process is guided by forecasting preformance.

The auto.Arima() command in R suggests using an AR(4) and MA(2) with 1st order differencing (p,d,q = 4,1,2). However, this model fails the residual diagnostic test. The distribution of the residuals has mostly constant variance but includes many outliers. The distribution of residuals is skewed right distribution. The autocorrelations of the residuals are not within the threshold limits and still show the sinusoidal-like seasonality for the diurnal cycle. Furthermore, the Ljung-Box test returns a small p-value, suggesting that the residuals are not white noise. The AIC of this model on the training dataset is 95150, and the MASE for one-month ahead forecast is 2.18, larger than one.

As discussed in the EDA, partial autocorrelation on the first-order differing data dies off after 26 lags. Thus, the first revised model uses ARIMA(26,1,2). Adding more lags to the AR or the MA components is equivalent to including more predictors in the model, thus fitting an ARIMA (26,1,2) is computationally expensive. This revised ARIMA model fails the residual diagnostic test. Despite the AIC score reducing to 93015, MASE for one month ahead forecast is still larger than 1. Thus, it is hard to say this model performs better.

The preliminary multiple linear regression in the EDA suggests all the six meteorological variables are significant predictors; thus a dynamic model using both the multiple linear regression and an ARIMA model is considered to see if the model performance is improved. Similarly, the main concern from the residual diagnostic test is the seanoality pattern on the autocorrelation plots of the residual. While the AIC score on the training data increase slightly to 93698, MASE for one-week ahead forecasting is 1.683, which is still not ideal.

Since the original time series model shows strong seasonality, and all the models above exhibit seasonality in the autocorrelation plot of residuals, a seasonal-ARIMA model is fitted. The baseline seasonal ARIMA model is ARIMA(0,1,3)(0,1,1)24 based on the discussion in last section. As the auto. ARIMA on the original hourly GHI suggests using AR(4), ARIMA(4,1,3)(1,1,1)24 is chosen. The residual diagnostic plot shown in Appendix 5. The residuals have mostly constant variance but include many outliers, and the distribution of residuals is approximately normal. The autocorrelations of the residuals are mostly within the threshold value and do not exhibit any seasonality. The AIC of this model on the training data set for one-month-ahead forecasting is 92000. And the MASE score on the one-month-ahead forecasting is 0.88. This suggests this seasonal-ARIMA model has better forecasting performance than the naive method. Different numbers of lags for AR ranging from 1 to 15 on the non-seasonal part have been experimented with, but neither the residual plots have significant differences nor the MASE score decreases significantly. Still, the forecasting accuracy is low.

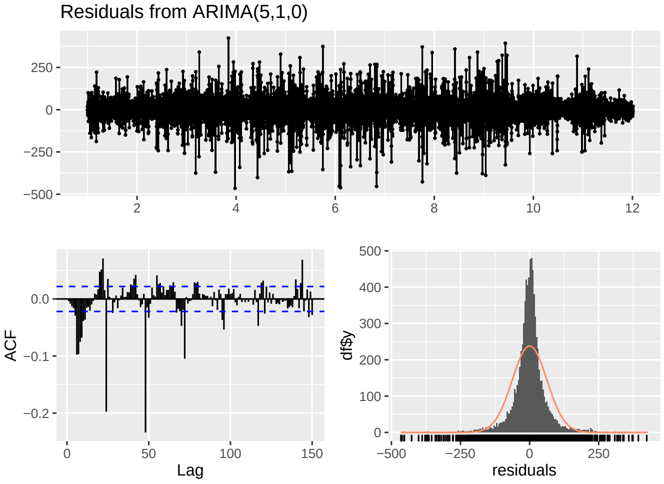
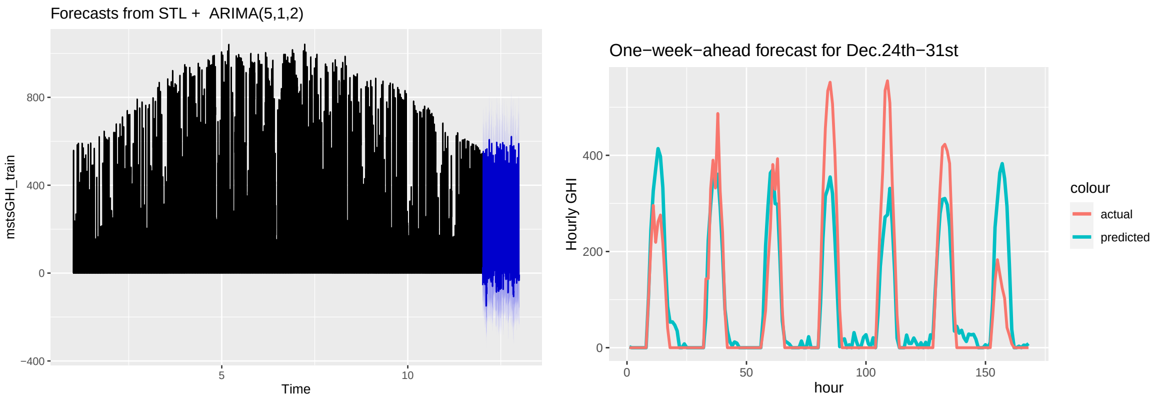


Figure 3: Residual diagnostic test for the STL model

Finally, the STL decomposition method is examined. The original data is decomposed into daily, monthly seasonal components, and the remaining seasonally-adjusted component using mstl(). The seasonal naïve method is used for the seasonal component, while ARIMA method is used for the seasonally-adjusted component. Autocorrelation of the seasonally-adjusted component after 1st order



(a) (b)

Figure 4: (a) One-month ahead (b) One-week ahead forecast using seasonal naive + ARIMA(5,1,2) STL model

differencing dies off after 2 lags (Fig 6 in Appendix), suggesting a MA(2) component. auto.Arima() on the seasonally adjusted component suggests ARIMA(5,1,2), which agree with the observation from the autocorrelation plot. Therefore ARIMA(5,1,2) is applied to the seasonally adjusted component. Fig 3 shows the residual diagnostic on the STL model. Residuals have mostly constant variance and are approximately normally distributed. The autocorrelation residuals mostly stays within the threshold values. Although seasonal autocorrelation spikes are visible but die off after 3 periods. The AIC score on the training data reduces significantly to 88847. The MASE score on the one-month ahead forecasting decreases to 0.17. Appendix 2 lists the estimated value for each coefficient in the model. Fig 4.(a) shows the forecasted value for hourly GHI in December. All predicted negative value should be treated as 0 since GHI must be zero or positive. Between the last two model, seasonal-ARIMA and STL model, since the forecasting of the latter is significantly smaller, the STL model proposed above is the final model. This model is fitted on all the data from Jan.1st to Dec.24 th to predict the hourly GHI for the last week of December. Fig 4.(b) shows the predicted results captures the 24 hour diurnal cycles, and the MASE score is 0.14.

5 Conclusion

The hourly GHI data for the area around Duke has strong daily seasonality and parabolic trend over one year. It is found that using STL decomposition time series with seasonal naive method on the seasonal components and ARIMA(5,1,2) on the non-seaonsal components yields satisfying forecasting results. However, the model has several limitations. First, the autocorrelation plot on the residuals still exhibits weak seasonality. Using STL decomposition method does not fully adjust for seasonality. Second, since the time length of the data is one year, the forecast model is not applicable to predict hourly GHI in future years. This model is appropriate for short-term forecast. If a similar analysis is done on data over multiple years, it would be interesting to see any cyclicity. Still this analysis provides a step by step guideline on how to analyze hourly GHI time series data using ARIMA model. In future, Neural Networks and machine learning methods could be further explored to improve forecasting accuracy.

References

[1] Mohammed H. Alsharif, Mohammad K. Younes, and Jeong Kim. "Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea". In: *Symmetry* 11.2 (2019). ISSN: 2073-8994. DOI: `10.3390/sym11020240`. URL: `https://www.mdpi.com/2073-8994/11/2/240`.

[2] Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. English. 2nd. Australia: OTexts, 2018.

[3] Gordon Reikard. "Predicting solar radiation at high resolutions: A comparison of time series forecasts". In: *Solar Energy* 83.3 (2009), pp. 342–349. ISSN: 0038-092X. DOI: `https://doi.org/10.1016/j.solener.2008.08.007`. URL: `https://www.sciencedirect.com/science/article/pii/S0038092X08002107`.

[4] Bismark Singh and David Pozo. "A Guide to Solar Power Forecasting using ARMA Models". In: *arXiv e-prints*, arXiv:1809.03574 (Sept. 2018), arXiv:1809.03574. arXiv: `1809.03574 [stat.AP]`.

Data Source: NSRDB: National Solar Radiation Database https://nsrdb.nrel.gov/about/tmy.html

6 Appendix

All the results in this analysis are reproducible using the R script file 'SolarForecast.R' in `https://github.com/shufan1/HourlyGHITimeSeries.git`.

variable	GHI	Dew point	Surface albdeo	Wind direction	Wind speed	Temperature	Pressure
unit	W/m^2	° C	Unitless	degree	m/s^2	° C	mbar

Table 1: variables and their unit used in the data set. Source: https://nsrdb.nrel.gov/about/tmy.html

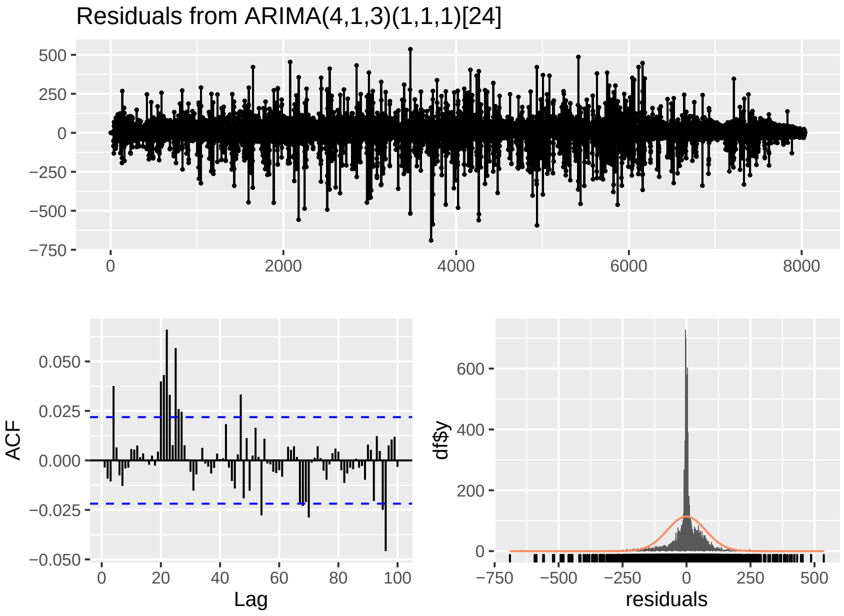


Figure 5: Residual diagnostic for the seasonal-ARIMA ARIMA(4,1,3)(1,1,1)24 model

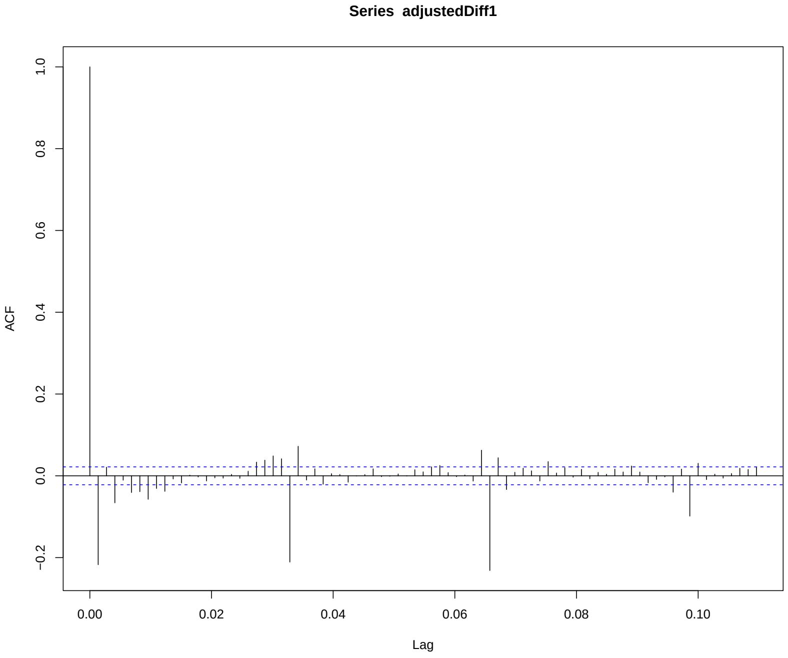


Figure 6: Autocorrelation plot of the seasonally-adjusted component after 1st order differencing.

	ar1	ar2	ar3	ar4	ar5	ma1	ma2
estimated value	0.163	0.520	0.042	-0.020	-0.026	-0.462	-0.528

Table 2: Estimated value for each coefficients in the STL naive+ARIMA(5,1,2) model. Standard error is not available. NaNs produced for standard error