Group #5
Professor Mohsen Pourahmadi
Statistics 626
4 August 2022

Analysis of Solar and Renewable Energy Generation

Introduction:

Our group is composed of Edmund Do, a PhD in Computer Science student on campus, Max Greenlee, an MS in Statistics distance student, Philip Kramer, an MS in Economics student on campus, and Casey Poulson, an MS in Statistics distance student. Philip Kramer is the Group Leader and will assist with programming, along with Casey Poulson. Edmund Do and Max Greenlee will focus on writing.

We retrieved our data from the U.S. Energy Information Administration (EIA). The agency "collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment," according to their website (About EIA). The dataset we selected for analysis reports the net electricity generation in Thousand MWh from utility-scale solar in the U.S. on a monthly basis. The span of the data is from January 2001 to March 2022. It is published in EIA's *Electric Power Monthly* and available online (Electricity Data Browser). We seek to identify trends and patterns in order to forecast net generation and capacity for solar and renewable energy sources. Generation of electricity at the utility level corresponds to the wholesale market for electricity. The market must clear through long-term bilateral contracts, day-ahead, and real-time markets due to our present inability to store electricity in a cost-effective manner. The goal of our analysis is motivated by this requirement as we seek to forecast and understand the relationships present in the data. Doing so will potentially assist grid planners in their ability to meet the needs of their region.

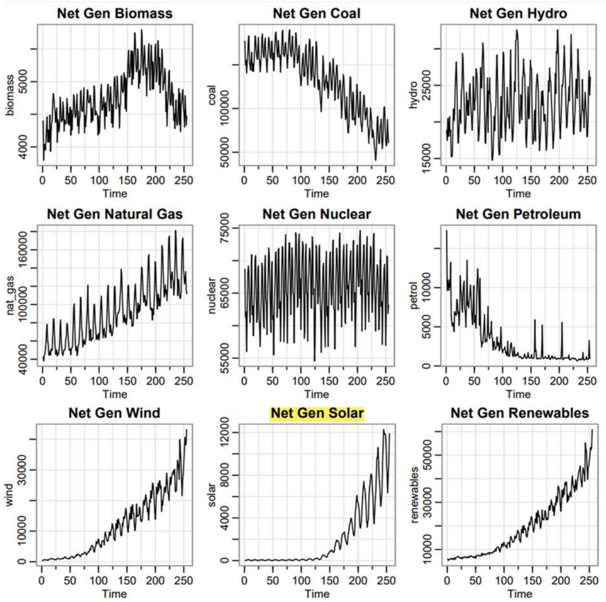
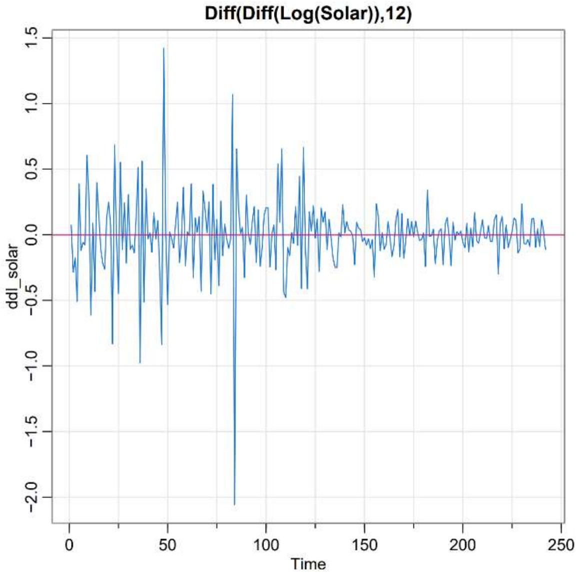


Figure 1: Highlighted Net Generation of Utility-Scale Solar Across All Sectors Monthly

Solar Data Modeling:

Noted above, there is a clear positive trend in the solar generation data, as well as clear cyclic seasonal patterns on a yearly 12-month basis. Prior to formulating a model, we must coerce the data to at least approximate stationarity. By first taking the log and first-order difference, the trend in the data is removed, but we still have non-constant variance, which we alleviate with seasonal differencing with a seasonal order of 12. As you can see in Figure 2 below, the data is now centered close to zero with relatively constant variance save for a few spikes.



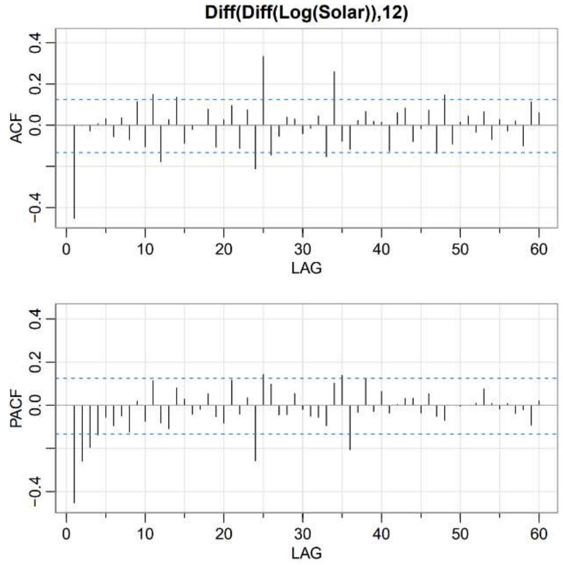
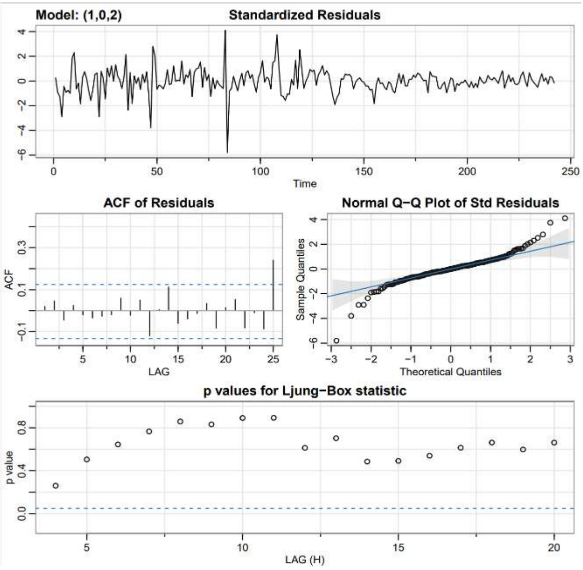


Figure 2: Seasonally Differenced Solar Data Series (left)
Figure 3: Seasonally Differenced Solar Data ACF & PACF (right)

Our first modeling attempt involved fitting an ARMA(1,2) component as well as an GARCH(1,1) component for the residuals to accommodate the heteroskedasticity we have grappled with before. We first examine the sample ACF and PACF of the series to come up with the ARMA component, see Figure 3. We see the ACF cuts off after lag 1, suggestive of MA(1), and the PACF cuts off around lags 1 or 2, suggestive of adding an AR term. After fitting many different order models, we found that an ARMA(1,2) model looked the best in terms of diagnostics, see Figure 4 below. The ACF of the residuals looks decent, with one spike at lag 25, and the Q-test fails to reject the null that the residuals are white for all lags displayed. However, the standardized residuals clearly have less variability on the right side of the plot, and the normal QQ plot tells us the residuals are heavier-tailed than normal. To model the variances more appropriately, we look at the ACF and PACF of the squared residuals from this model to formulate a GARCH component, see Figure 5 below. Note the ACF and PACF cuts off after lag 1, suggestive of ARMA(1,1) residuals. For our first model, we used an ARMA(1,2) + GARCH(1,1) model, see Solar Model 1 below. Estimated coefficients are given with standard error in parentheses.



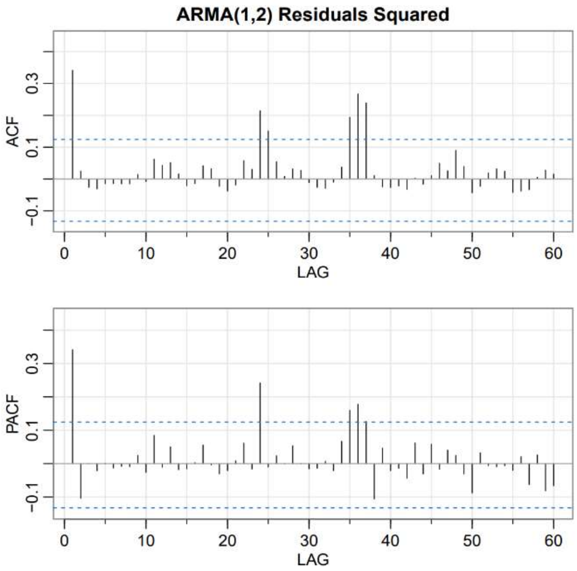


Figure 4: ARMA(1,2) Residual Diagnostics (left)
Figure 5: ARMA(1,2) Residuals^2 ACF & PACF (right)

Solar Model 1: $\hat{x}_t = \hat{\mu} + \hat{\varphi}_1 x_{t-1} + \hat{\theta}_1 w_{t-1} + \hat{\theta}_2 w_{t-2} + \hat{\sigma}_t^2$ *with* $\hat{\sigma}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 r^2_{t-1} + \hat{\beta}_1 \hat{\sigma}^2_{t-1}$

$$\hat{x}_t = -0.001_{(0.007)} - 0.862_{(0.083)} x_{t-1} + 0.289_{(0.096)} w_{t-1} - 0.594_{(0.067)} w_{t-2} + 0.0001_{(0.0002)} + 0.037_{(0.0196)} r^2_{t-1} + 0.951_{(0.019)} \hat{\sigma}^2_{t-1}$$

We note here that all coefficients are significant at level 0.05 except for the two constant terms. Next, we will look at residual diagnostics from Solar Model 1, see Figure 7 below. The standardized residual plot looks similar to white noise. The ACF plot looks good with only one value at lag 12 outside the 95% confidence bands, and the normal QQ plot suggests the residuals fit a normal distribution quite well. The Q-test of whiteness fails to reject the null that the residuals are white for lags 1-24. Figure 6 plotted to the right is the seasonally differenced solar data in blue with red-dashed fitted values from Solar Model 1 overlaid. This model does a fair job at capturing the heteroskedastic nature of the data, but is still rather conservative at the extreme values. We are not dissatisfied with these results, but we hope to coerce the model residuals closer to white noise with different modelling techniques. Next, we attempt to fit a seasonal model to the solar data, hoping to accommodate the cyclic patterns this time with seasonal AR and/or MA terms.

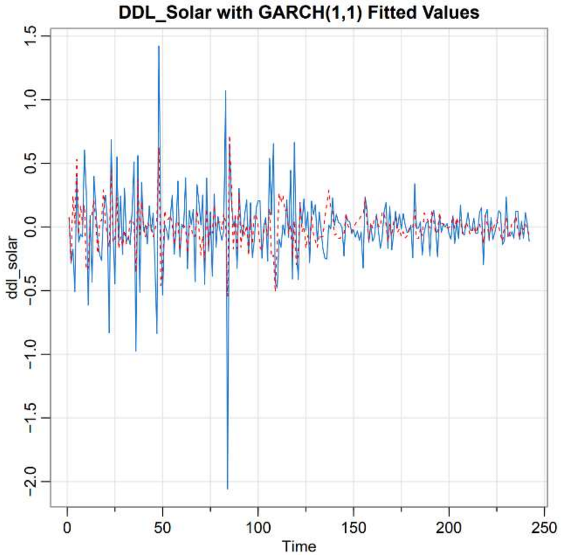


Figure 6: Solar Model 1 Fitted Values

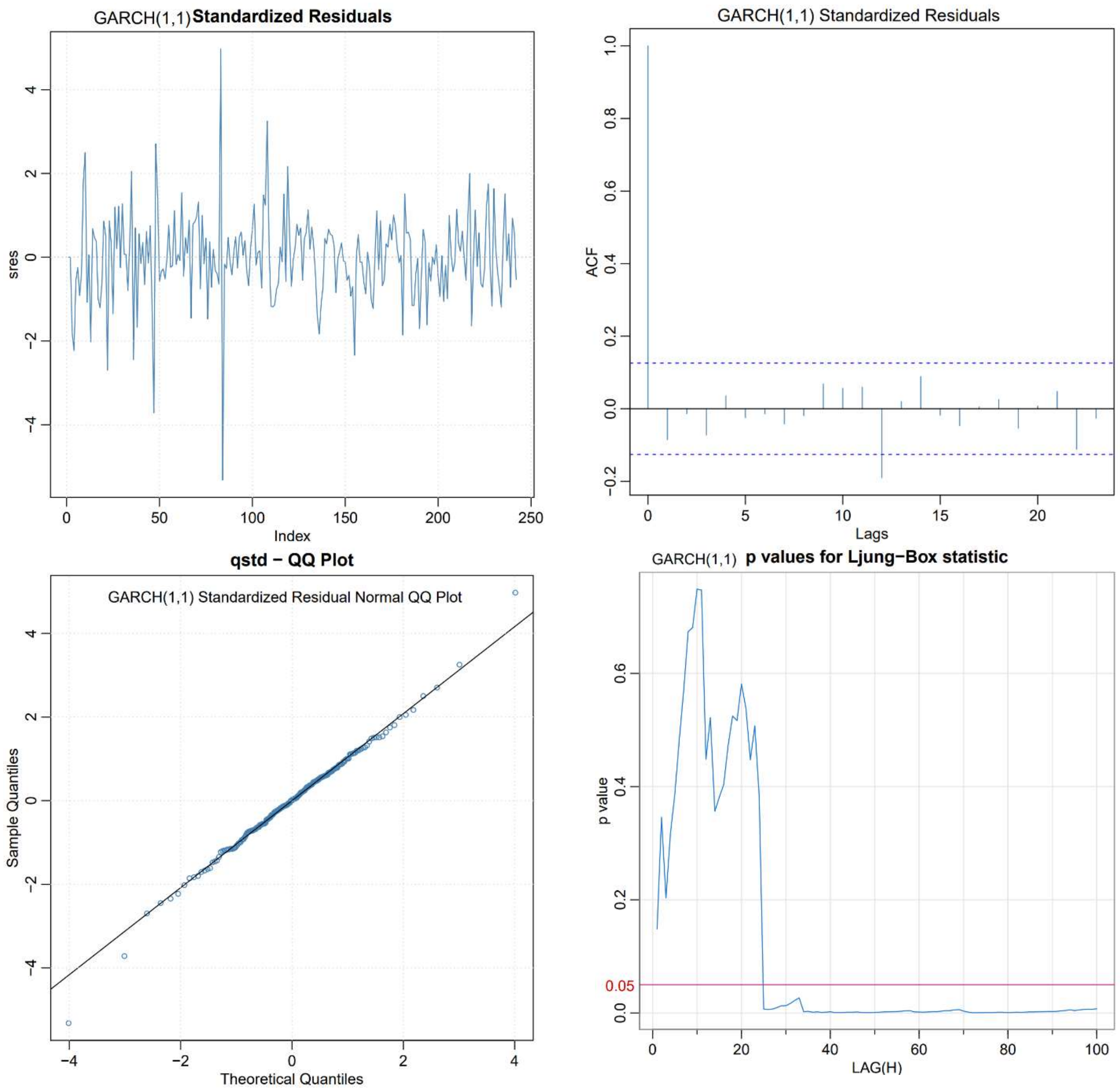


Figure 7: Solar Model 1 Residual Diagnostic Plots

In the previous model, we dealt with the seasonality of the data by taking seasonal difference. In the next model, we try to do so with a SARIMA fit. Figure 9 plotted below shows the first-differenced and log of the solar data, along with its corresponding ACF and PACF plots. It is evident in Figure 9 that the data is centered at zero but with non-constant variance. It is our hope to bring those values down with seasonal terms, and the ACF and PACF is plotted mainly to show that the underlying seasonality is still very much present in the data. Through trial-and-error we arrived at a SARIMA model for the data-Solar Model 2-expressed below.

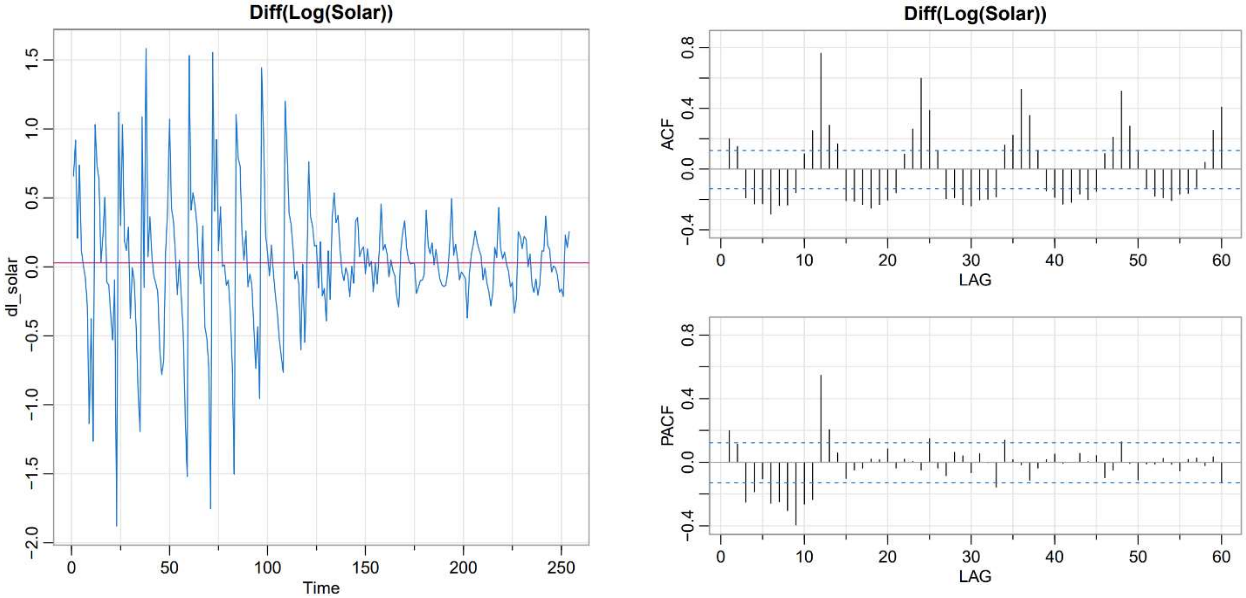


Figure 9: Diff(Log(Solar)) and its ACF and PACF Plots

Solar Model 2: $\hat{x}_t = \hat{\mu} + w_t + \hat{\theta}_1 w_{t-1} + \hat{\theta}_2 w_{t-2} + \hat{\Phi}_1 x_{t-12} + \hat{\Phi}_2 x_{t-24} + \hat{\Phi}_3 x_{t-36}$

$$\hat{x}_t = 0.021_{(0.039)} + w_t - 0.599_{(0.064)} w_{t-1} - 0.145_{(0.064)} w_{t-2} + 0.8_{(0.064)} x_{t-12} + 0.0004_{(0.087)} x_{t-24} + 0.144_{(0.069)} x_{t-36}$$

We note that all coefficients are significant at level 0.05 except the constant and the SAR(2) term. To judge the performance and usability of this model, we look next at the residual diagnostics as well as forecasted values, plotted below in Figure 10.

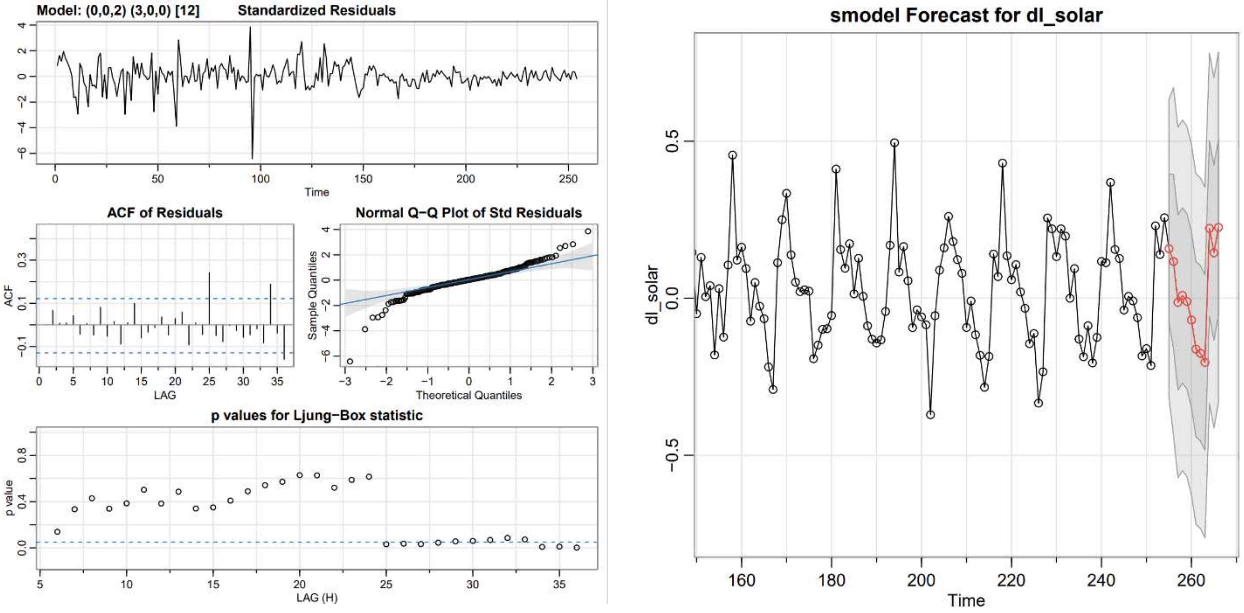


Figure 10: Solar Model 2 Residual Diagnostics and 12-month Forecast

Upon examination of the diagnostic plots above, we immediately see that the non-constant variance problem persists in the standardized residuals, and the QQ plot suggests an approximate normal distribution to the residuals although with heavy-tails. The ACF looks alright with only two values outside the confidence bands at lags 25 and 34, and the Q-test fails to reject the null hypothesis that the residuals are white for about the first 24 lags. The forecast plot suggests that the predicted values do closely follow the pattern of the transformed solar data, but the heteroskedasticity remains an issue, which we hope to address in the final solar model that follows.

Solar Data-Lagged Regression Modeling

We also wanted to investigate fitting a model to the solar data based on information from another data set. The coal net generation data, from the same EIA source, has the scatterplot shown. There appears to be a quadratic relationship between the variables. However, this type of plot does not tell us anything about the relationship over time. For that, we turn to a Cross Correlation Function (CCF) plot (Figure 11).

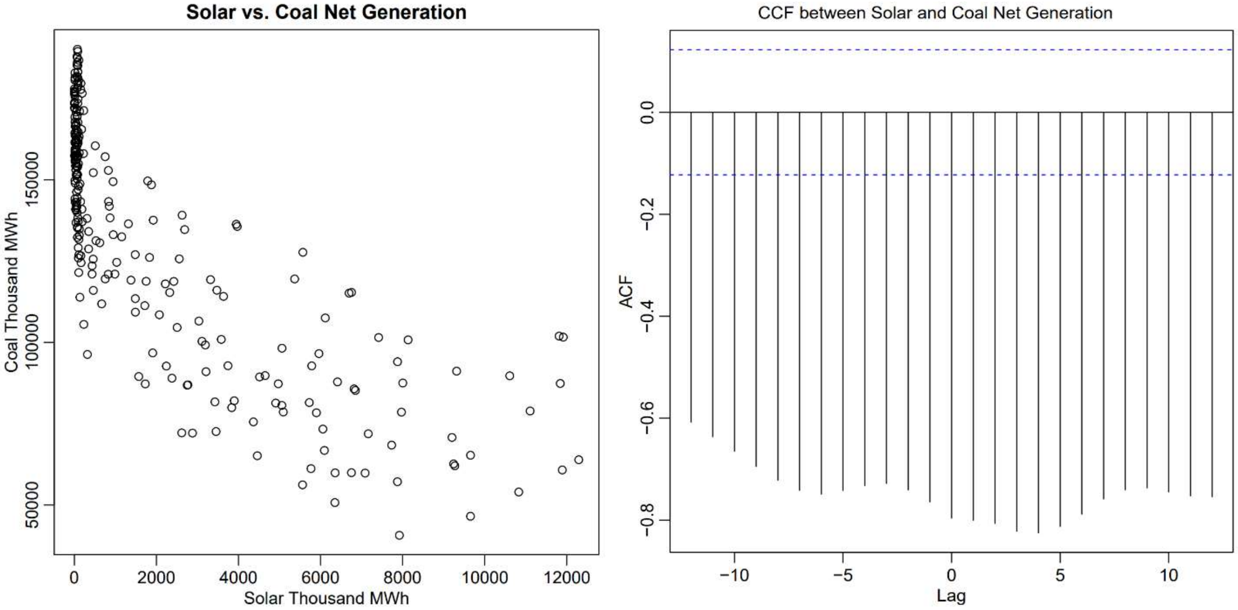
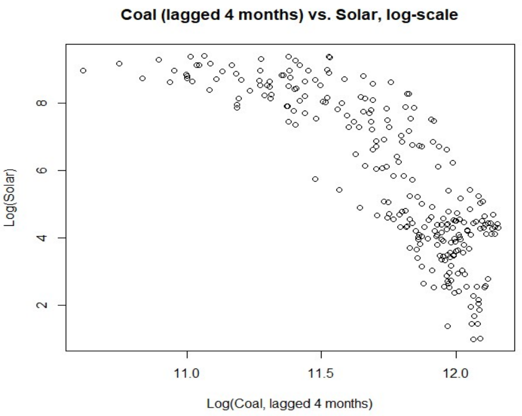


Figure 11: Solar vs. Coal with CCF Values

This plot shows relatively strong negative correlations across a large range of values. The largest correlation is at lag 4, which because of how the plot was constructed means that the coal data lags the solar data. Using that lag, we can create a new series that aligns the coal and solar data. We also apply a log transformation to attempt to deal with some of the heteroskedasticity issues. A scatterplot of the intersected data series, log transformed, is given in Figure 12. The quadratic relationship noted in the initial scatterplot remains and is even clearer in this plot. Therefore, we can attempt to remove some of the trends in the solar data using regression with this lagged and logged coal data. We fit the line using R and find the following coefficients:

$$y_t = -524.5261 + 97.6260x_{t-4} - 4.4644x_{t-4}^2$$



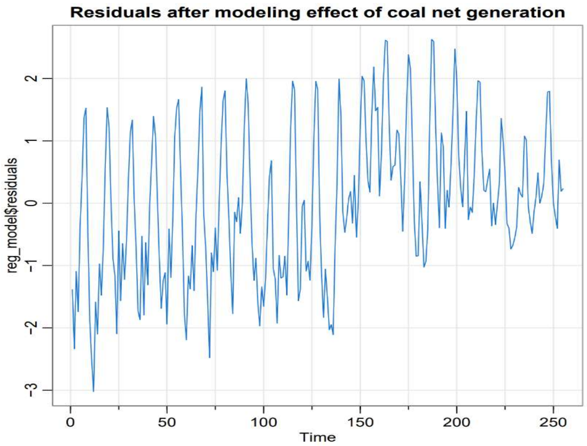


Figure 12: Scatterplot of lagged and logged coal data vs. log solar data (left)
Figure 13: Model residuals after lagged regression (right)

Note that y_t is the log solar data series and x_{t-4} is the lagged and logged coal data series. The residuals of this model are shown in Figure 13. The lagged regression did not fully remove the trends in the data, and certainly did not remove the cyclical patterns, but the combination of this with the log transformation removed much of the heteroskedasticity. We can further iterate on this model to fully reduce the data to stationarity.

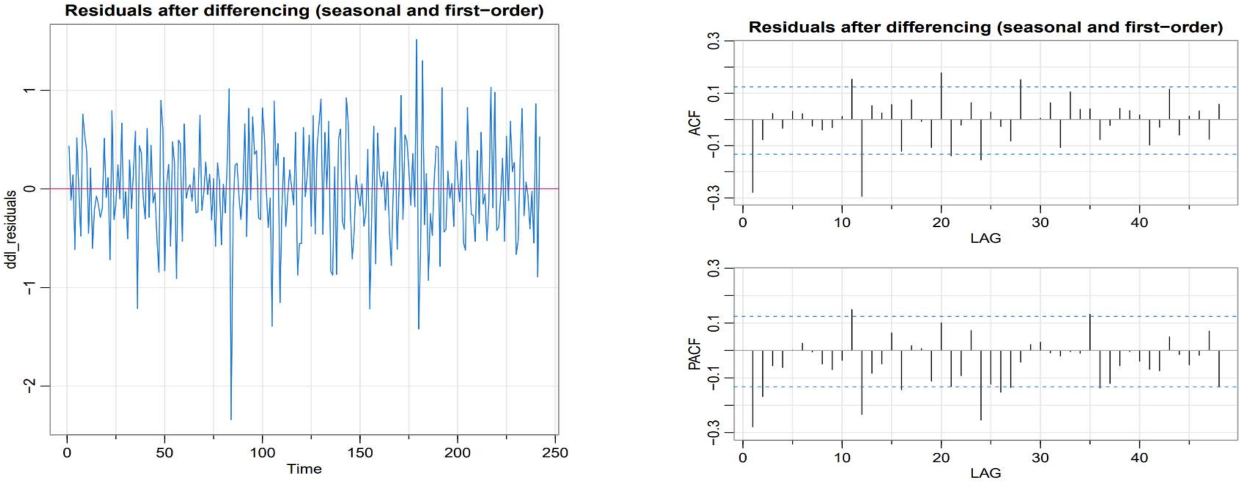


Figure 14: Residuals after differencing

We applied first order and seasonal differencing (with a period of 12). This process removed the trend present in the model residuals and most of the seasonality effects. A time series plot of the data after this process is shown in Figure 14. From the plot, it seems the data is close to stationary and thus the ACF and PACF have a valid interpretation. These plots are also given in Figure 14. We first look at the seasonal lags (i.e., 12, 24, 36, 48, etc. in this case). There is a gradual decline in the PACF and a sharp decline in the ACF after lag 12, which is suggestive of an MA(1) term for the seasonal part of the SARIMA model. For the non-seasonal part of the model, the interpretation is less clear. There is clearly correlation in both the ACF and PACF at lag 1. The

PACF appears to decline rather than drop off. We tried fitting several low order models to the nonseasonal portion, including an AR(1), MA(1), and ARMA(1,1). The best model for this portion was the ARMA(1,1), so the complete model ended up being a SARIMA(1,1,1,0,1,1), applied to the residuals of the data after a log transformation and lagged regression. The model diagnostic plot and the P/ACF of the residuals are in Figure 15. The final model equation is:

$$\nabla^{12}\nabla log\ (\widehat{x_t}) = -524.526 + 97.626\ log\ log\ (C_{t-4})\ - 4.464 log\ (C_{t-4})\ ^2 + 0.516x_{t-1}$$
$$- 0.821w_{t-1} - 0.655w_{t-12} + w_t$$

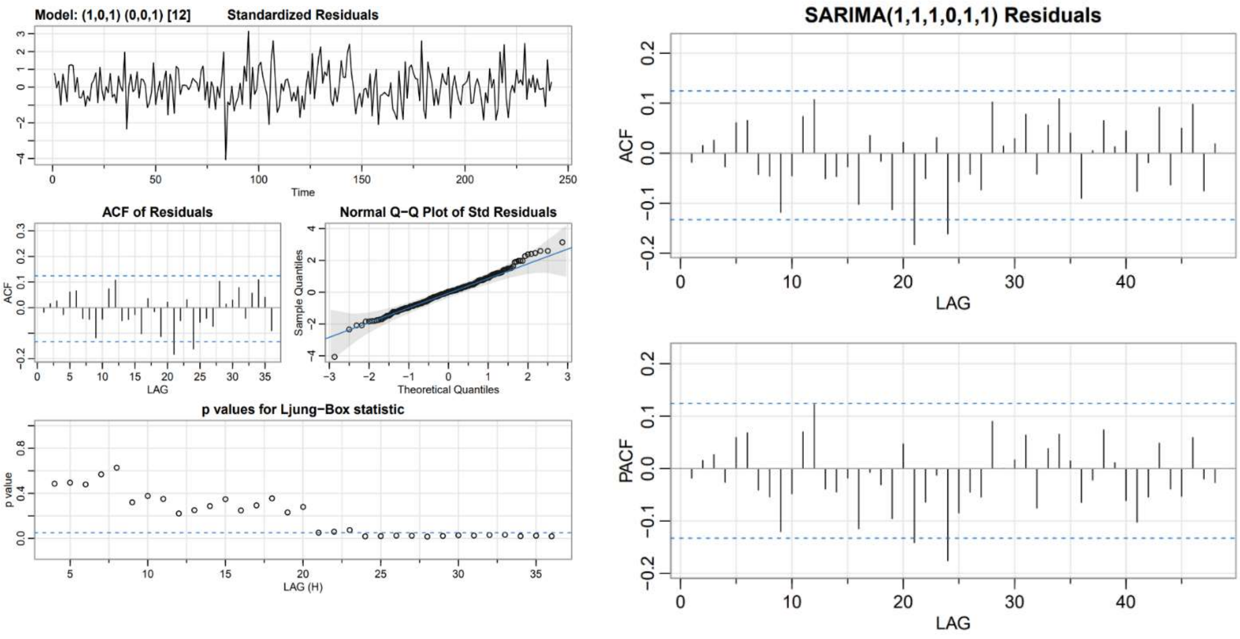


Figure 15: Final lagged regression residual diagnostics (above)

The diagnostics for the model all look reasonably like white noise. The time series plot has no clear trend and looks random. The ACF and PACF do not have any lags well outside the bounds of white noise. The Q-test statistics are all non-significant before lag 20. Finally, the normal reference distribution plot shows the data is well-represented by a normal distribution. As a last step, we can look at the forecasts from the model, which are given in Figure 16. The forecast looks like it captures both the trend and the pattern of seasonality well. This model performs quite well, even in comparison to the more advanced GARCH techniques.

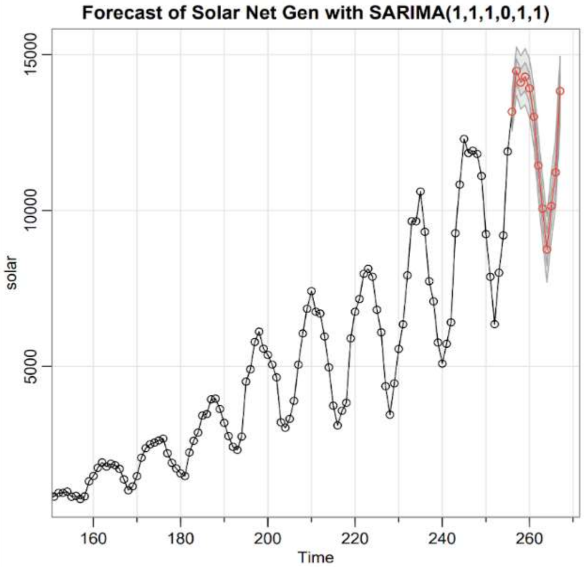


Figure 16: Lagged regression model predictions

Renewable Data Modeling:

In our second presentation we attempted to model the aggregated renewable energy generation data, and so we do so again with more tools in our belt. The series is clearly non-stationary as seen in Figure 1, and so we coerce it to approximate stationarity with first log transformation, first-order, and seasonal differencing. This series, plotted in Figure 17 below, is centered at zero, and though we see smaller variability in the beginning of the series, overall, it appears well-enough to continue. The ACF of this series cuts off after lag 12, suggestive of adding a SMA(1) term, and the PACF tails off around lag 3 or 4, suggestive of adding a MA(3) term to the model. We fit just such a model in Renewable Model 1 expressed below Figure 18.

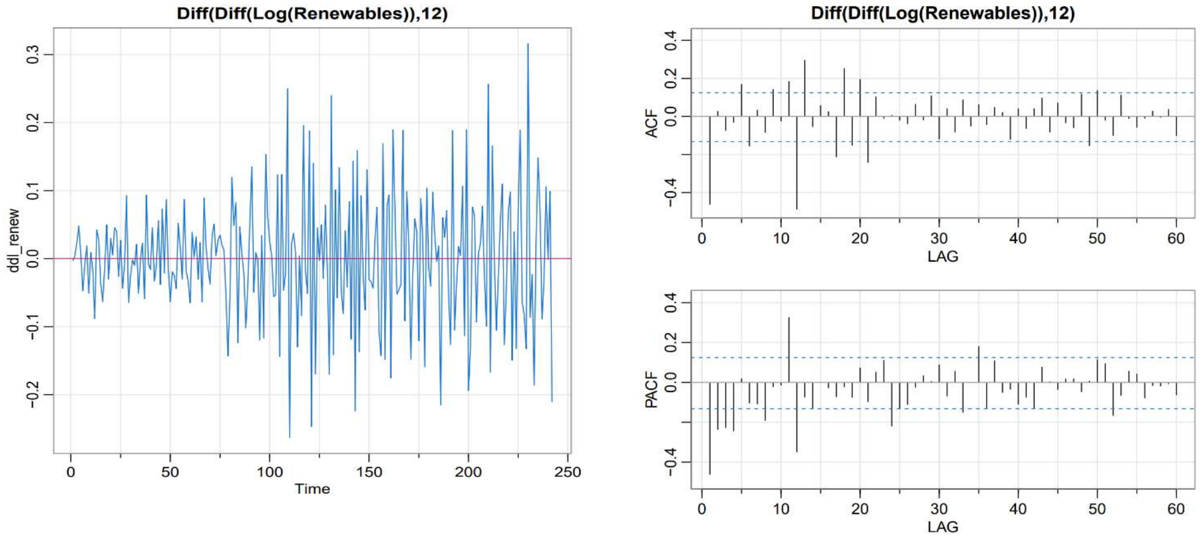


Figure 17: Seasonally Differenced Renewable Data with ACF & PACF

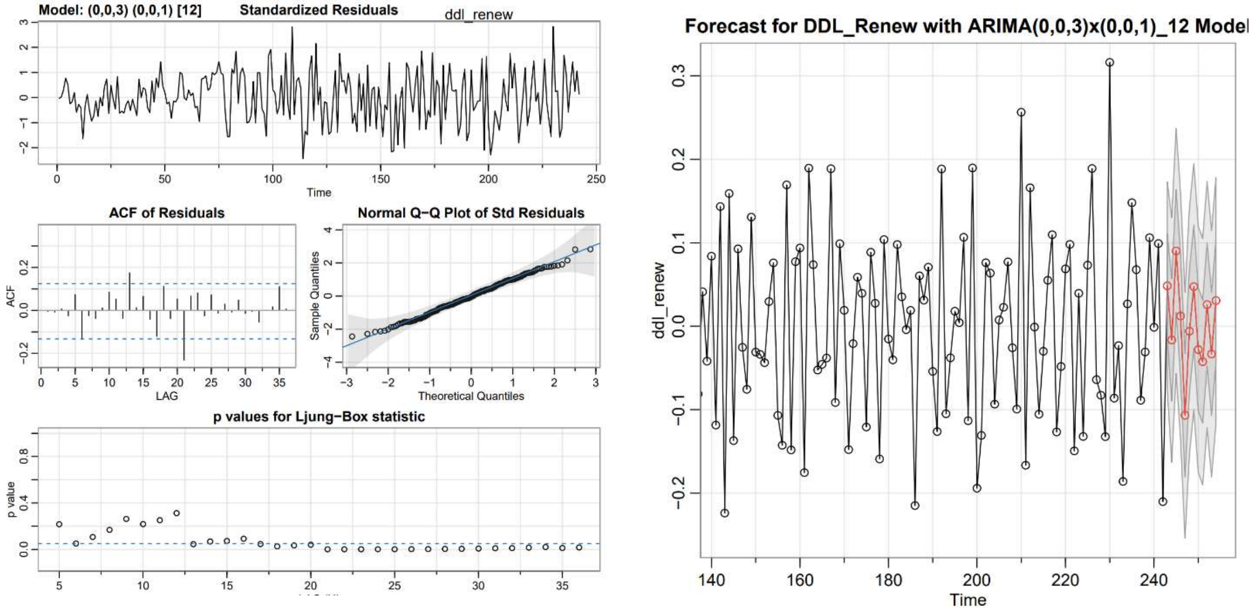


Figure 18: Renewable Model Residual Diagnostics and 12-Month Forecast

$$\textbf{\textit{Renewable Model 1:}}\ \hat{x}_t = \hat{\mu} + \hat{\theta}_1 w_{t-1} + \hat{\theta}_2 w_{t-2} + \hat{\theta}_3 w_{t-3} + \hat{\Theta} w_{t-12} + w_t$$

$$\hat{x}_t = 0.0003_{(0.0003)} - 0.622_{(0.067)} w_{t-1} - 0.043_{(0.076)} w_{t-2} - 0.123_{(0.067)} w_{t-3} - 0.692_{(0.054)} w_{t-12} + w_t$$

Diagnostics and 12-month forecasted values for Renewable Model 1 are displayed above in Figure 18. We note the standardized residuals appear approximately white with slightly less variability on the left. The ACF has two spikes outside of the confidence bands at lags 13 and 21, but mostly fall inside the bands, and the QQ plot suggest a good approximate normal fit. It is hard to see but the Q-test fails to reject the null that the residuals are white noise for lags 1-20. The 12-month forecast looks good and seems to follow the bumping-around pattern of the seasonally differenced renewable data quite well. Overall, we were satisfied with this model.

We did attempt fitting several GARCH models to the residuals from ordinal ARMA fits to the renewable data, but decided to omit the output again for the sake of brevity, and because Renewable Model 1 above outperformed any other fits. We also found the AIC and BIC values for Renewable Model 1 above were smaller than those for our GARCH fits, so we decided to stick with Renewable Model 1 for this data.

Conclusion

Working with heteroskedastic data proved to be an exciting challenge throughout the project. Through the implementation of SARIMA, GARCH, and regression models, we feel confident in our ability to model and forecast potential growth in solar and aggregate renewable energy generation. Based on our model diagnostics, we feel most confident recommending our lagged regression solar model and the ARIMA(0,0,3)x$(0,0,1)_{12}$ renewable generation model. We propose that the analysis summarized in this report is applicable for grid planners, policy makers, and investors, as such knowledge is crucial to the complex yet essential intersection between energy generation and consumer demand due to our present inability to store energy in a large scale and cost-effective manner. Future analysis of the data might involve multivariate analysis of market trends through PCA and FAVAR due to the endogenous relationship between all types of energy generation.

References

US Energy Information Administration. *About EIA*. n.d. 12 June 2022.

—. *Electricity Data Browser*. March 2022. CSV. 12 June 2022.