Solar Energy Production Forecasting

Matthew Echols, Chris Olen,
Abhishek Chaturvedi, Targoon Siripanichpong

Background & Objectives

- Once cost prohibitive, solar power generation has seen explosive growth over the past few years and is quickly becoming one of the key energy producers in certain parts of the country

- Understanding the patterns and drivers of solar energy production will help illuminate opportunities in this growing industry and leverage this crucial tool in the fight against climate change

Dataset

Solar Energy Production

The monthly U.S. production of solar energy in trillions of British Thermal Units (Btu).

Data was gathered from the U.S. Energy Information Administration.

GDP

The monthly U.S. GDP in billions of dollars.

This is a measure of macroeconomic growth, which may be correlated to overall energy production and demand.

CPI of Gas

The Consumer Price Index for gasoline used by households in all forms (e.g heating).

This is a measure of how costly key alternative sources of energy to solar are for households.

Crude Oil Price

The weighted price of crude oil per barrel (across Dubai, Brent, and WTI producers).

This is a measure of how costly key alternative sources of energy to solar are for companies.

Silicon Production and Price

The production of silicon in tons and the price of silicon in dollars per ton.

This is a measure of how costly a key input of solar panels is.

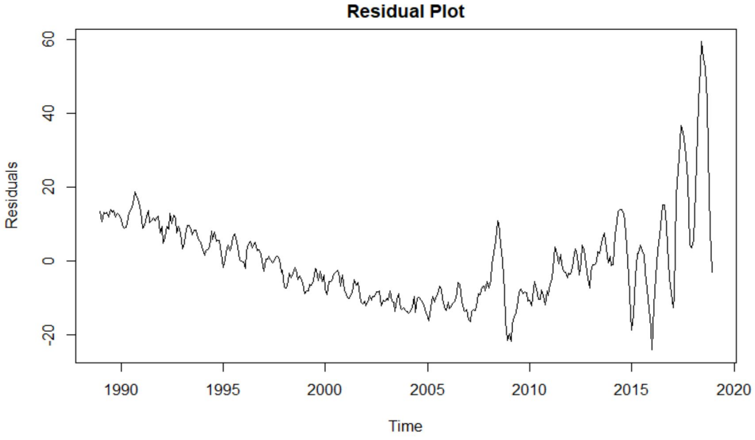
Cloud Cover

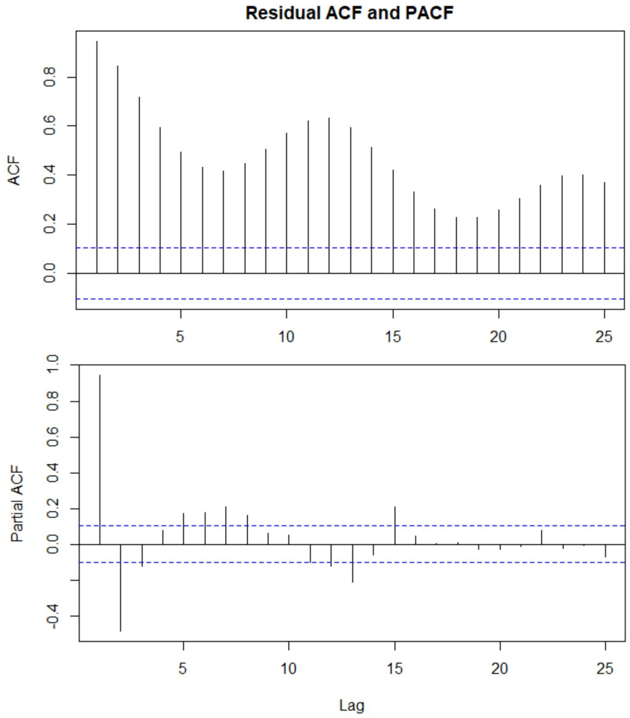
The cloud cover over key locations for solar energy production measured in %..

This is a measure of environmental effects on solar energy production.

Naive Cross-Sectional Approach

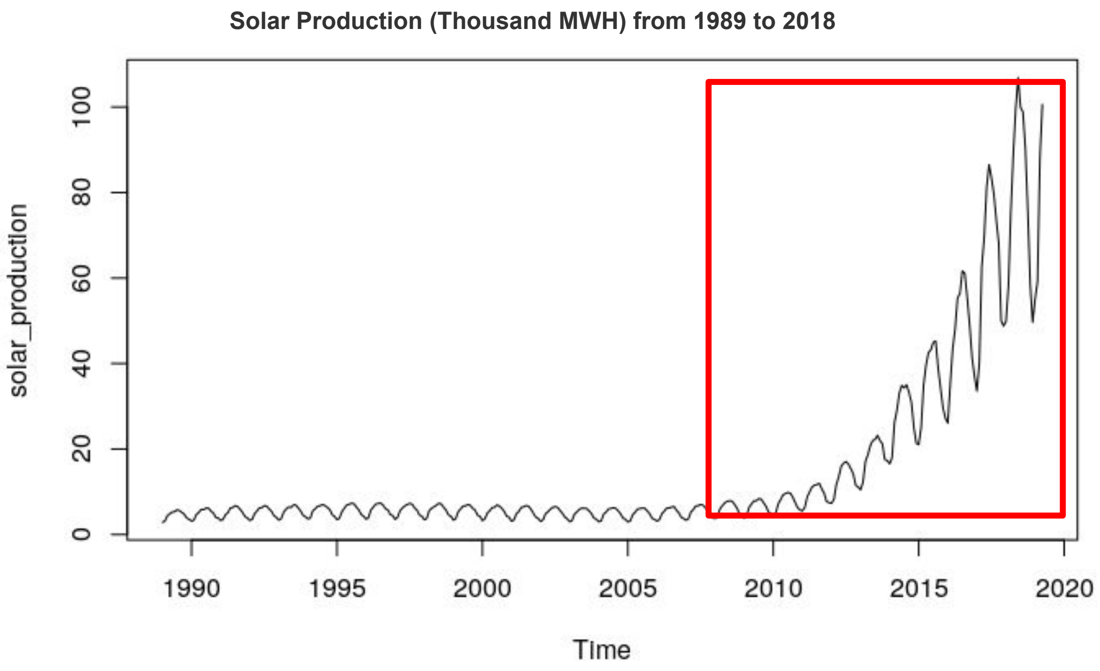
Residuals clearly indicate **autocorrelation** and **seasonality** - not something that can be removed using other cross-sectional modeling techniques.

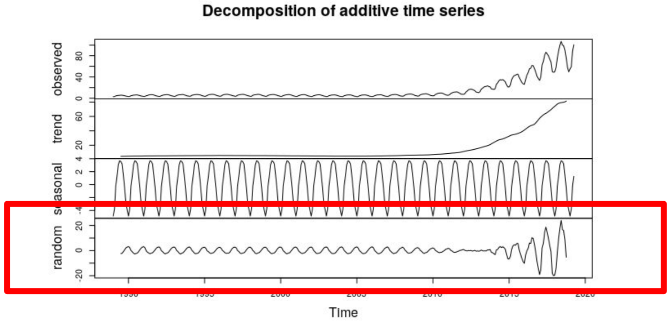


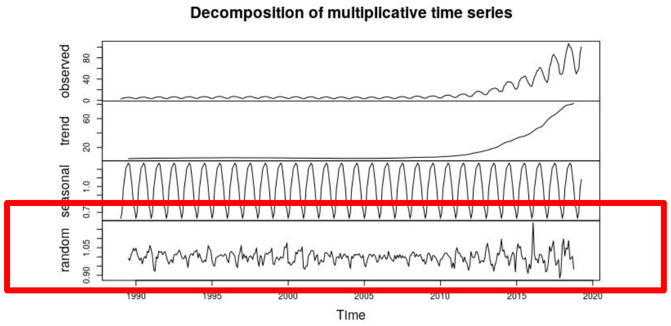


For reference, the model used for this was:
Solar_Production ~ GDP + Weighted_Crude_Oil_Price + Cloud_Cover (Adj R^2 = 0.6075)

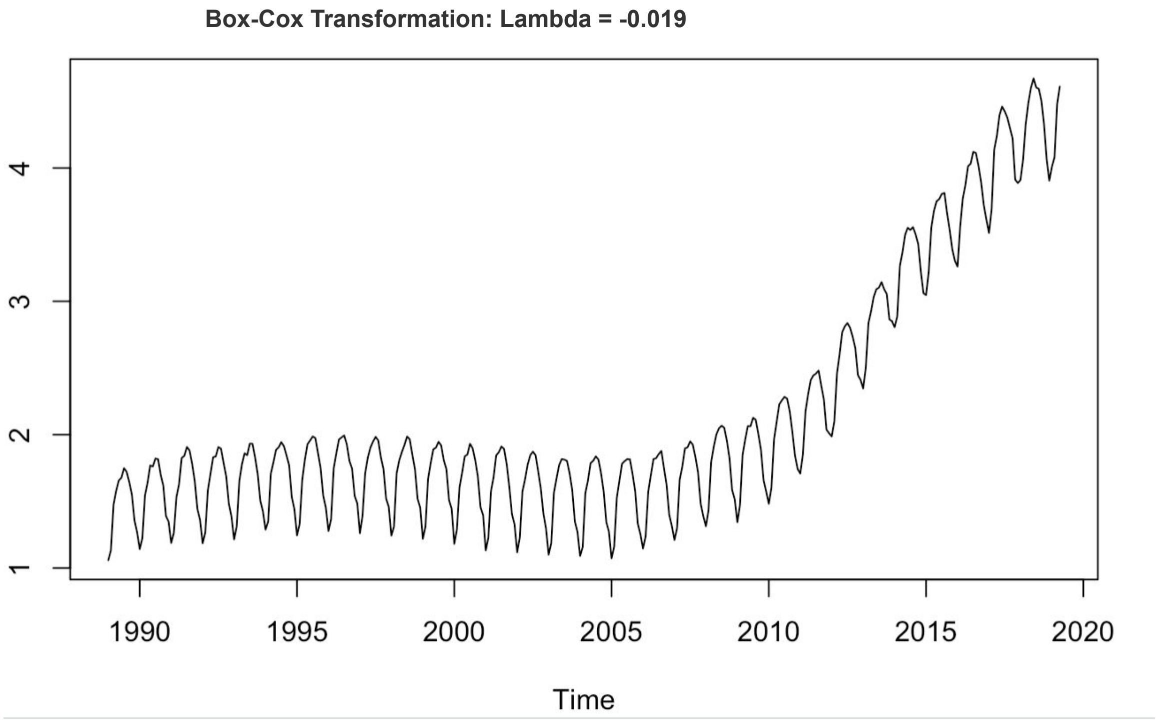
Series Decomposition



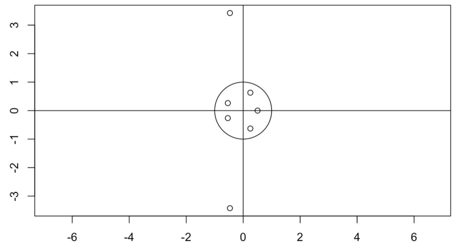




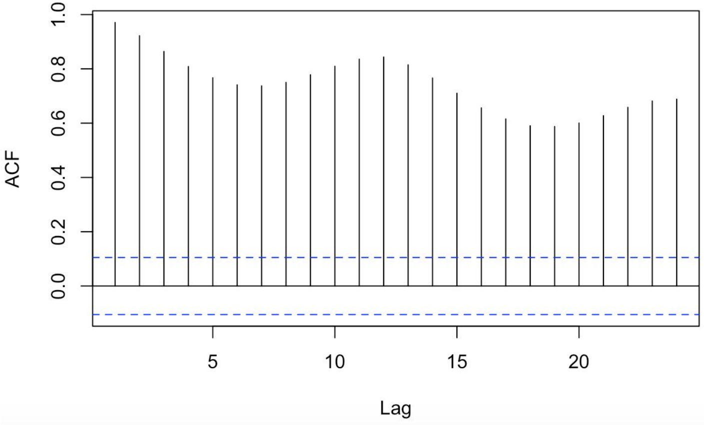
Box-Cox Transformation



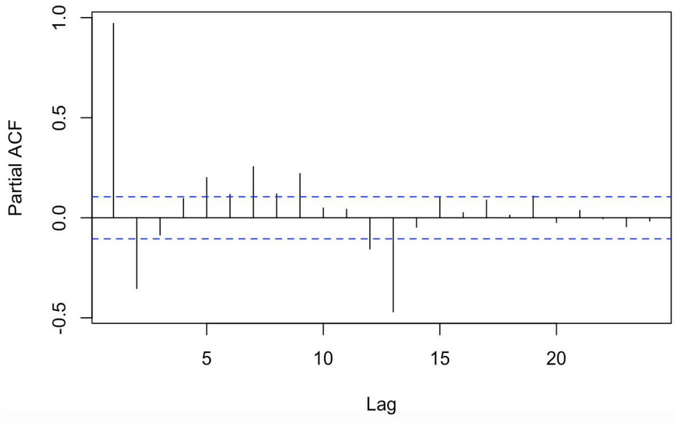
Autocorrelation and Non-Stationarity



ACF - No Differencing

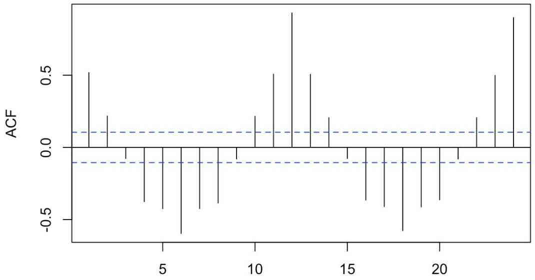


PACF - No Differencing

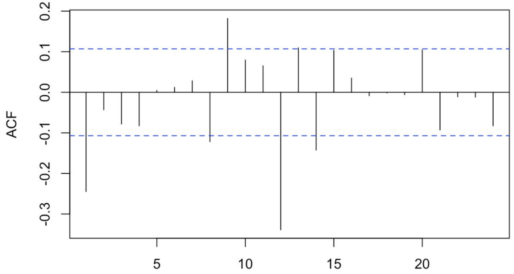


Differencing and Seasonality

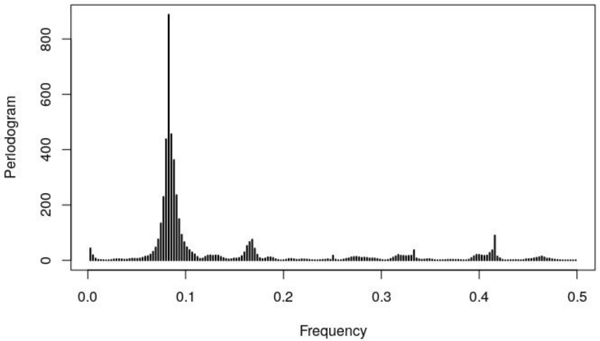
ACF - First Order Differencing



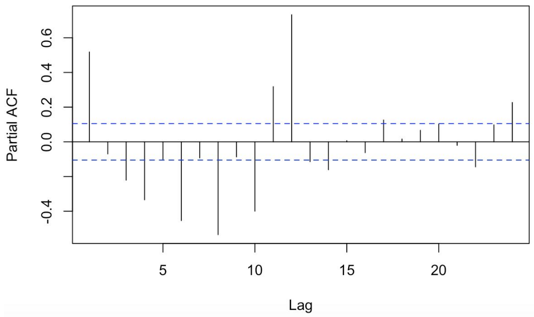
ACF - First Order + Seasonal Diff



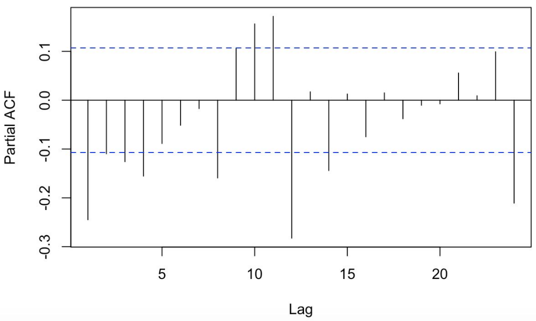
Periodogram



PACF - First Order Differencing

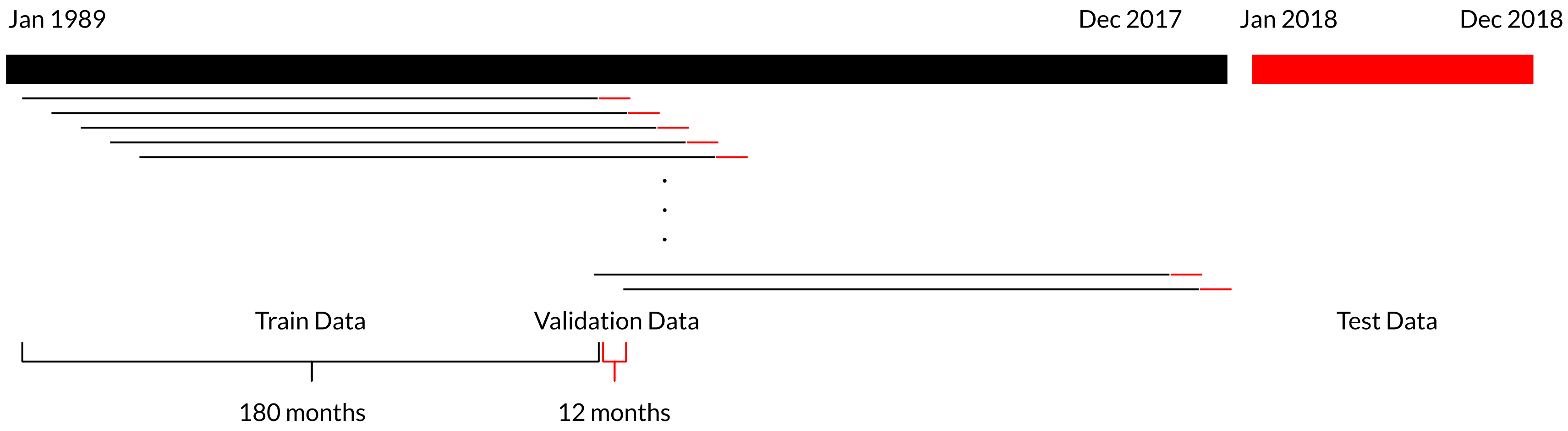


PACF - First Order + Seasonal Diff

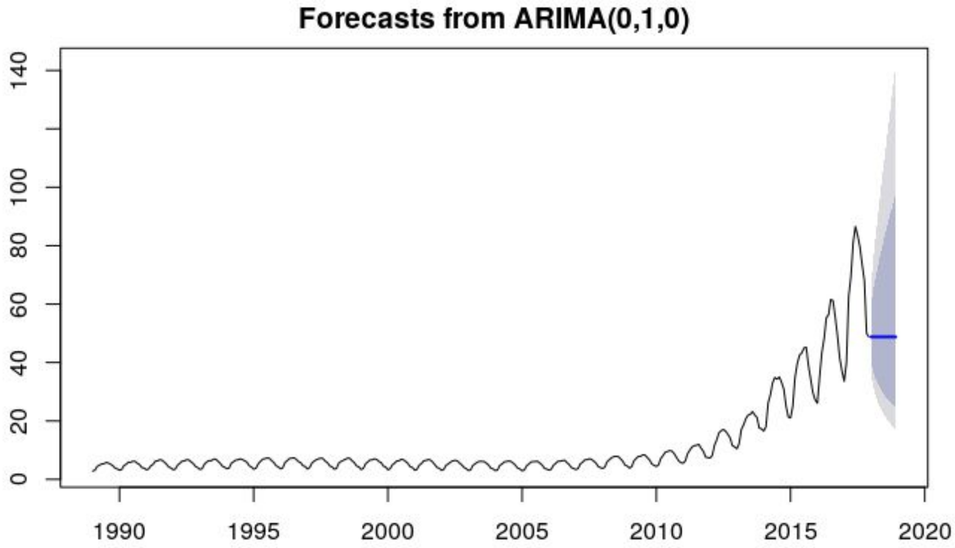


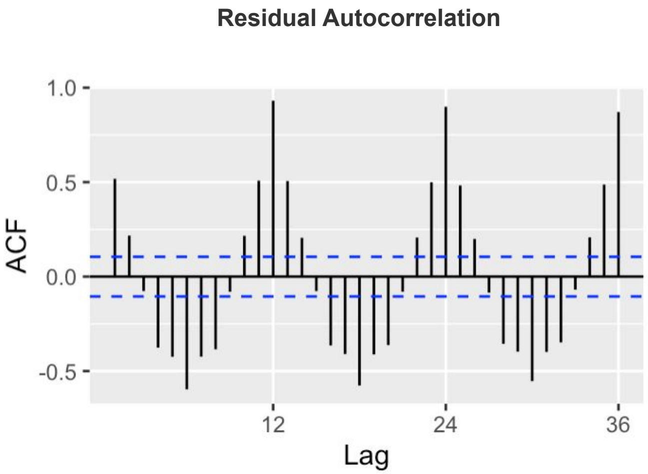
Model Validation Process

Evaluation Metrics: MSE & MAE

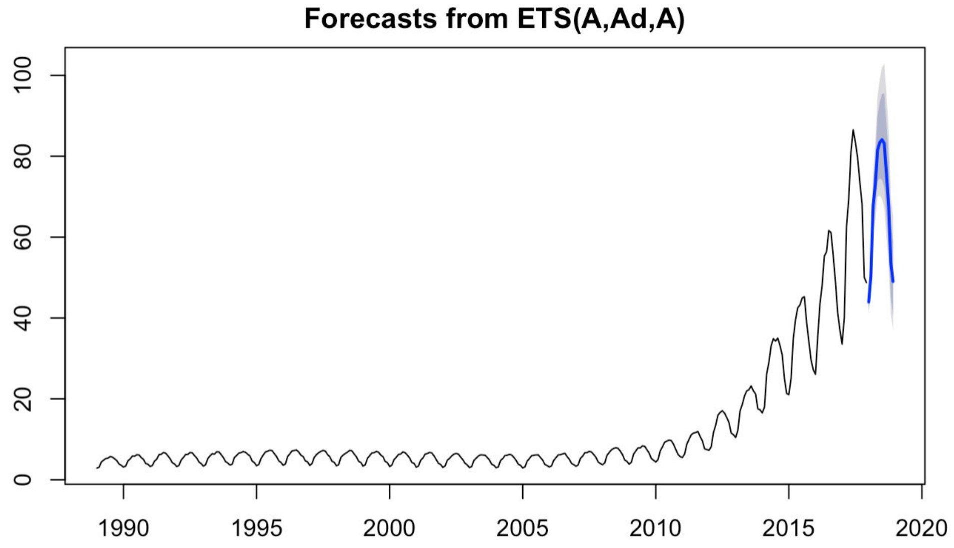


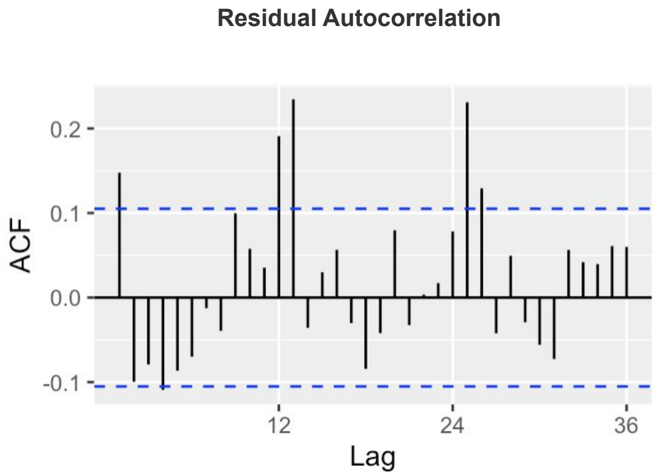
Baseline Model: ARIMA(0,1,0)





Exponential Smoothing (A,Ad,A)





$$y_t = (l_{t-1} + 0.85b_{t-1} + s_{t-m})(1 + \epsilon_t)$$

$$l_t = (l_{t-1} + 0.85b_{t-1} + 0.612(l_{t-1} + 0.85b_{t-1} + s_{t-m})(\epsilon_t)$$

$$b_t = (0.85b_{t-1} + 0.155(l_{t-1} + 0.85b_{t-1} + s_{t-m})(\epsilon_t)$$

$$s_t = s_{t-m} + 0.191l_{t-1} + 0.85b_{t-1} + s_{t-m})(\epsilon_t)$$

sARIMA: Parameter Selection

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12
0  x o o o o o o x x o o  x  o
1  x o o o o o o o x o o  x  x
2  x o o o o o o o x o o  x  o
3  x o x x o o o o o o o  x  o
4  x o o x o o o o o o o  x  o
5  x o o x o o o o o o o  x  x
6  x o o x x o o o o o o  x  o
7  o x x x x o o o o o o  x  o
8  x x x o x o o x o o o  x  o
9  x x x o x o o o x o o  x  o
10 x o x x x x o o x o o  x  o
11 x x x x x o o x x x x  x  x
12 o x o x o o o o o o o  x  x
```

```
        [,1]    [,2]   [,3]
[1,]  0.1342 -0.0639 -0.035
[2,]  0.2645 -0.0472 -0.103
[3,]  0.3808 -0.0877 -0.079
```

- ACF and PACF suggest ARIMA(1,1,1)(1,1,1)[12]
- Insignificant coefficients on fit models suggest *seasonal* ARMA(0,1) instead of (1,1)
- EACF on differenced series additionally suggests ARMA(0,1) and ARMA(2,1)
- Finally, significant PACF spikes at lag 4 led us to try ARIMA(4,1,1)(0,1,1)[12] with AR order 2 and 3 coefficients dropped

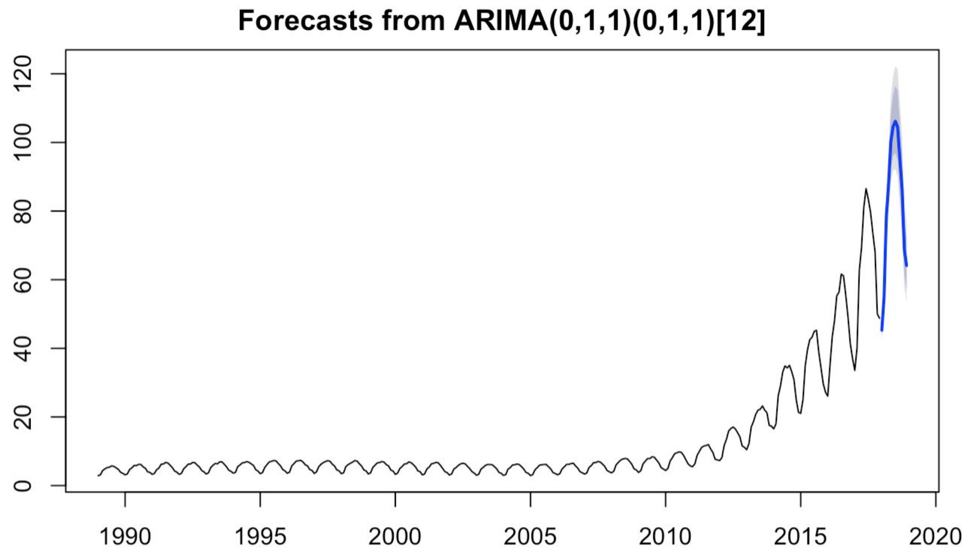
```
Coefficients:
         ar1  ar2  ar3      ar4      ma1     sma1
      0.2854    0    0  -0.0760  -0.5218  -0.5971
s.e.  0.1522    0    0   0.0592   0.1368   0.0490

sigma^2 estimated as 0.0009593:  log likelihood=689.01
AIC=-1368.02   AICc=-1367.83   BIC=-1348.95
```

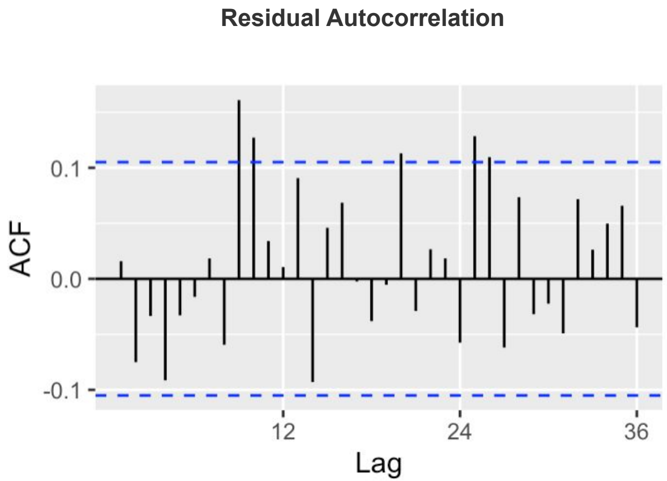
sARIMA: Parameter Selection (Continued)

- The two models with the lowest AICc scores AND Box-Ljung test p-values that did not reject the null hypothesis at the .01 level are:

1. ARIMA(0,1,1)(0,1,1)[12]
2. ARIMA(0,1,2)(0,1,1)[12] *(provided by Auto Arima)*

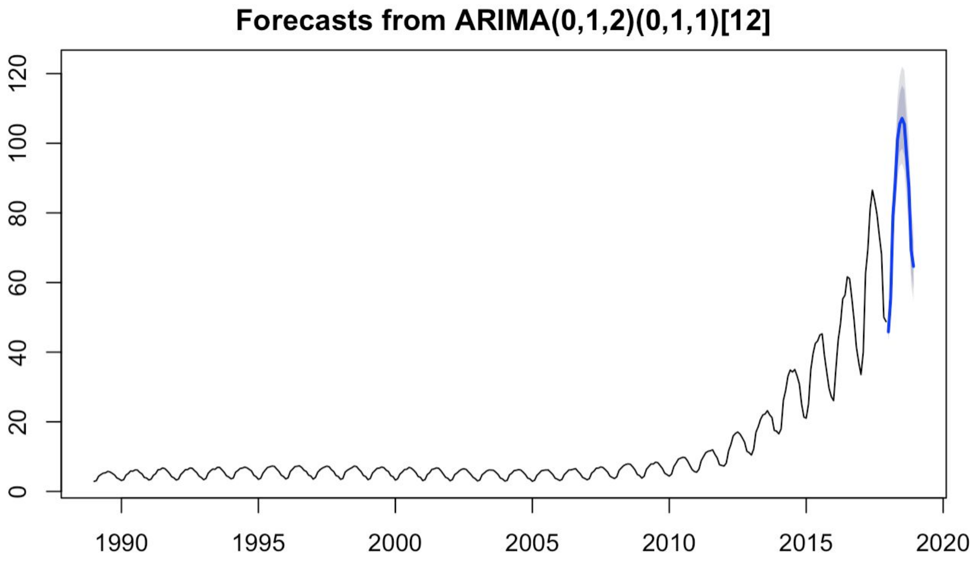
ARIMA (0,1,1)(0,1,1)[12]

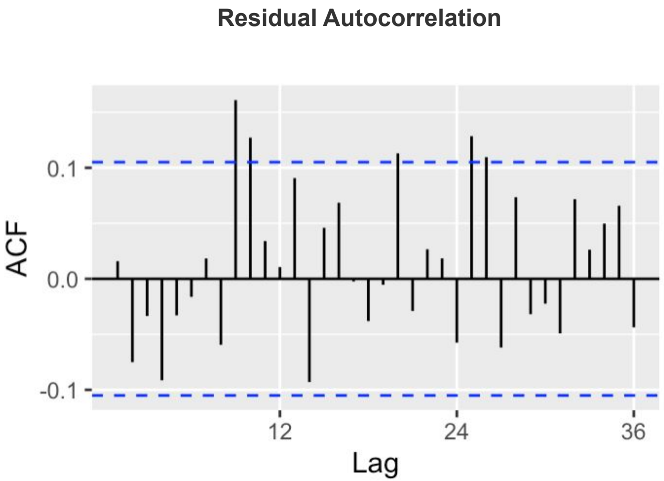




$$y_t = (1 + -0.2384e_{t-1})(1 + -0.6214e_{t-12})$$

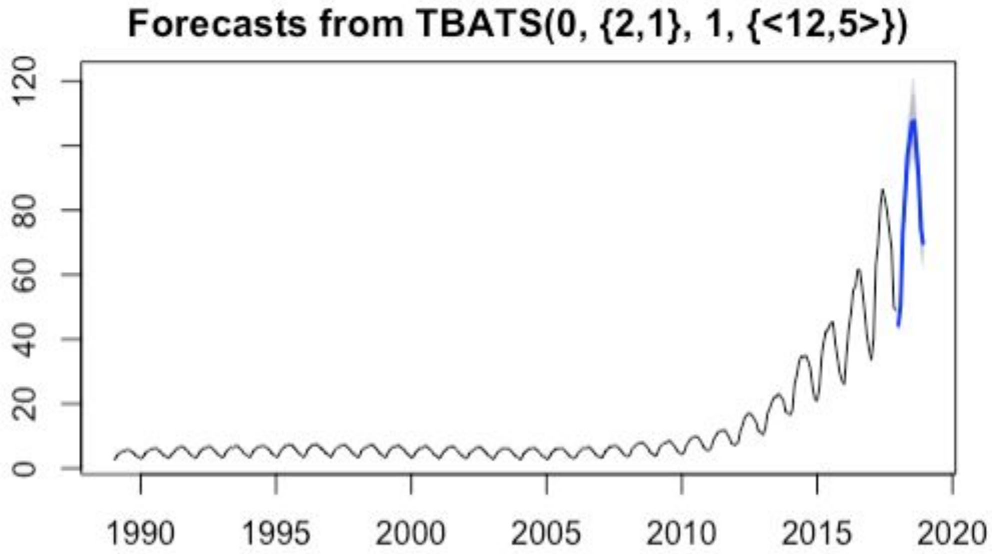
ARIMA (0,1,2)(0,1,1)[12] - Auto Arima

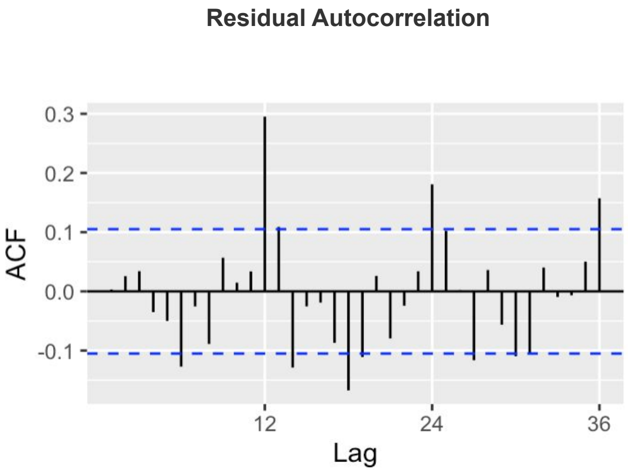




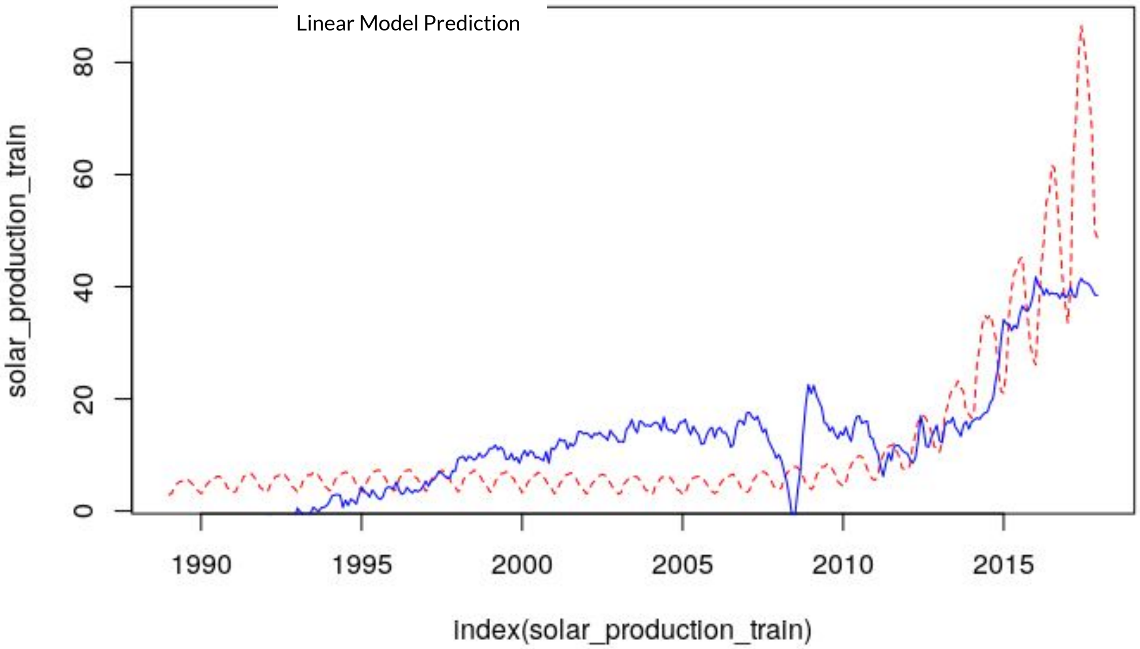
$$y_t = (1 + -0.2262e_{t-1} + -0.1053e_{t-2})(1 + -0.6104e_{t-12})$$

TBATS (0, {2,1}, 1, {<12,5>})





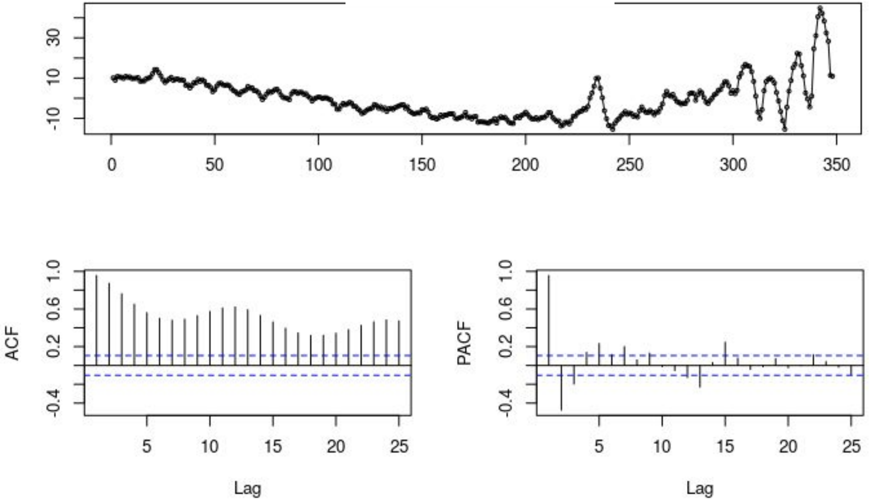
Regression with ARIMA Residuals



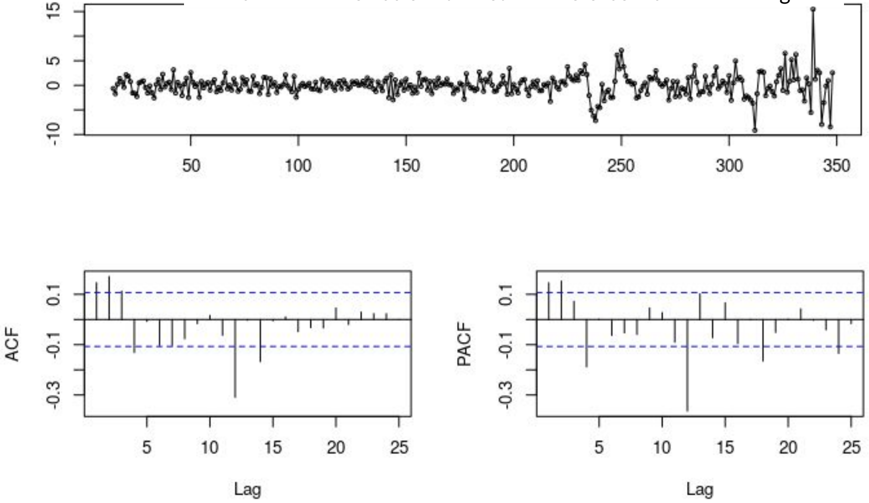

```
Call:
lm(formula = EnergyProduction ~ GDP + weighted_crude_oil_price +
    cloudCover, data = solar[1:348, ])

Residuals:
    Min      1Q  Median      3Q     Max
-18.034  -7.299  -0.943   5.831  45.061

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.456e+01  2.486e+00  -9.878   <2e-16 ***
GDP                       3.962e-03  1.958e-04  20.238   <2e-16 ***
weighted_crude_oil_price -2.724e-01  2.463e-02 -11.064   <2e-16 ***
cloudCover                4.184e+00  3.316e+00   1.262    0.208
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.651 on 344 degrees of freedom
Multiple R-squared:  0.5847,    Adjusted R-squared:  0.5811
F-statistic: 161.4 on 3 and 344 DF,  p-value: < 2.2e-16
```

Regression with ARIMA Residuals

Linear Model Residuals

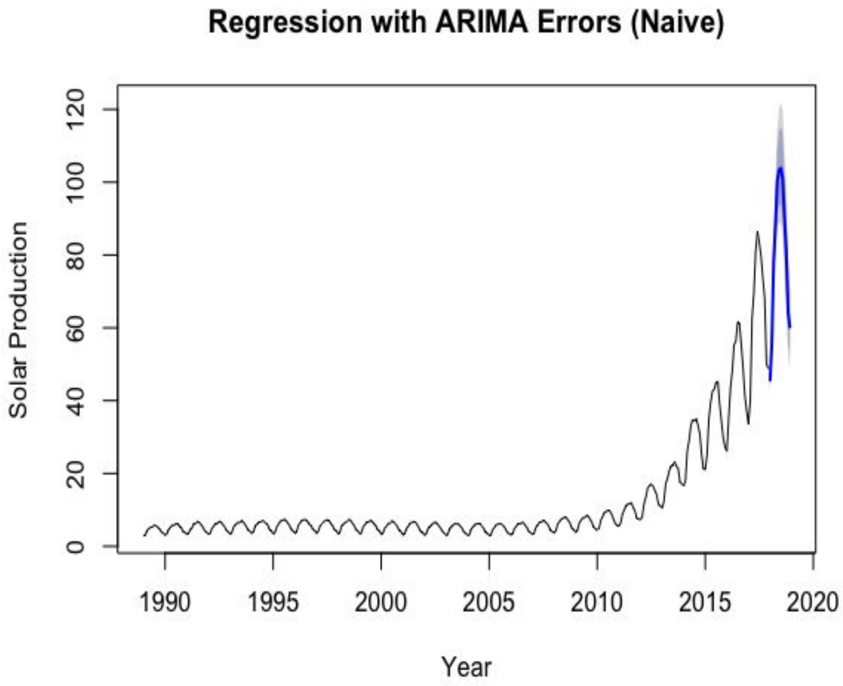


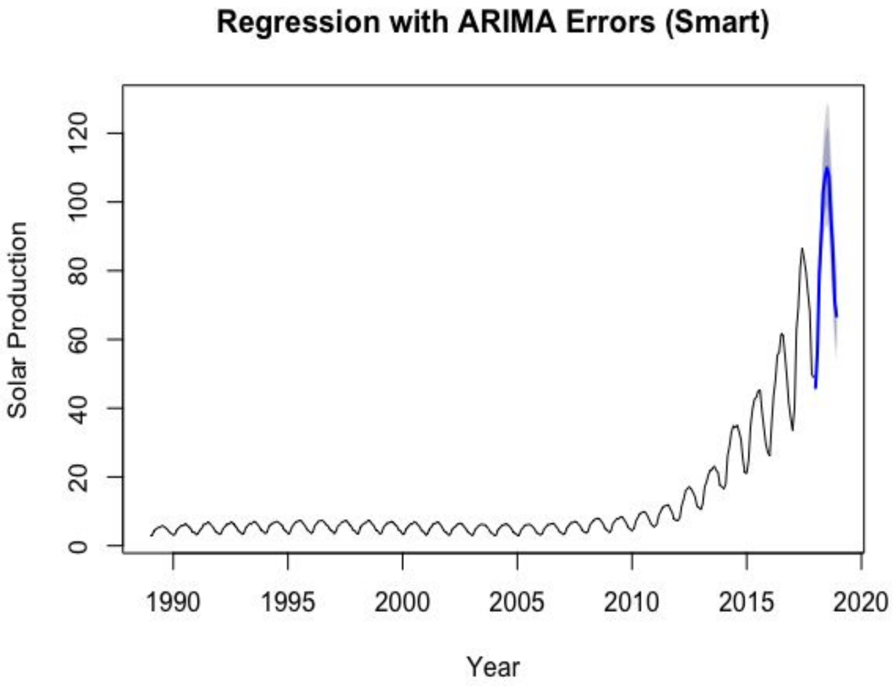
Linear Model Residuals with first-order & seasonal differencing



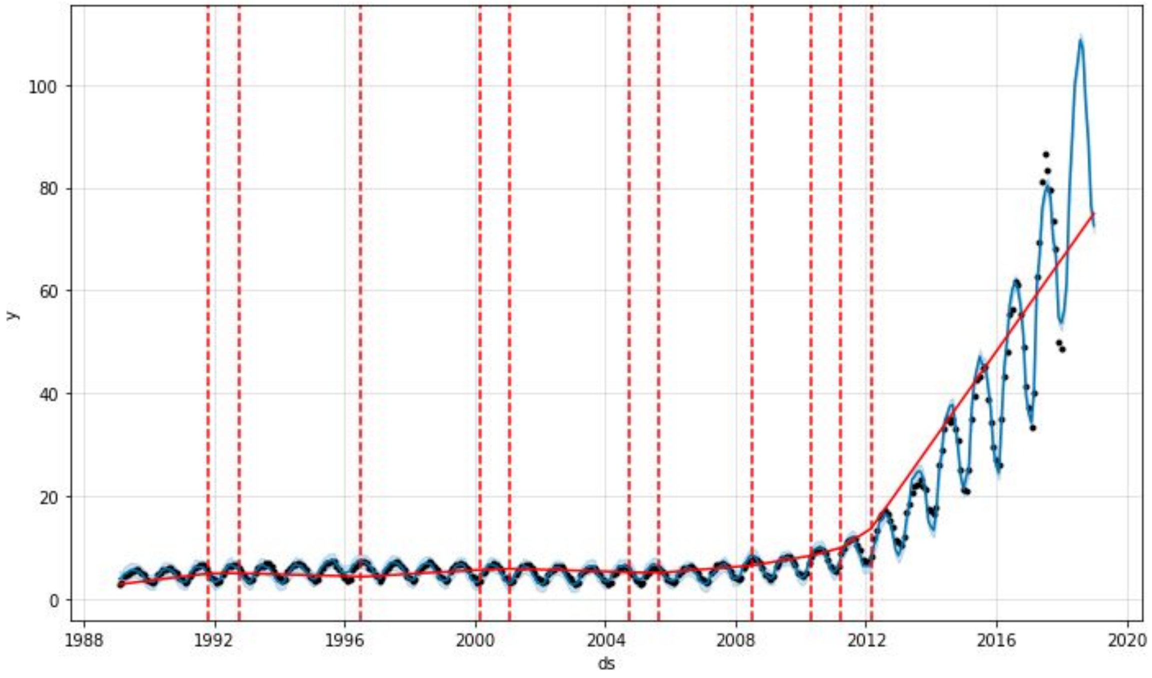
Regression with ARIMA Residuals

ARIMA(2,1,2)(1,1,1)[12]



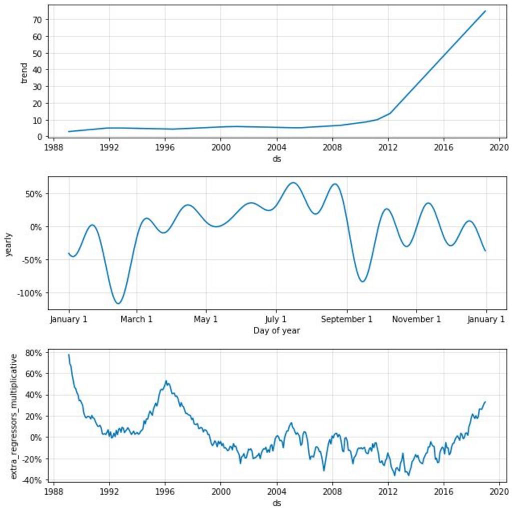


Prophet



Model Development

- Seasonality Type
  - Additive
  - Multiplicative
- Univariate vs Multivariate Models



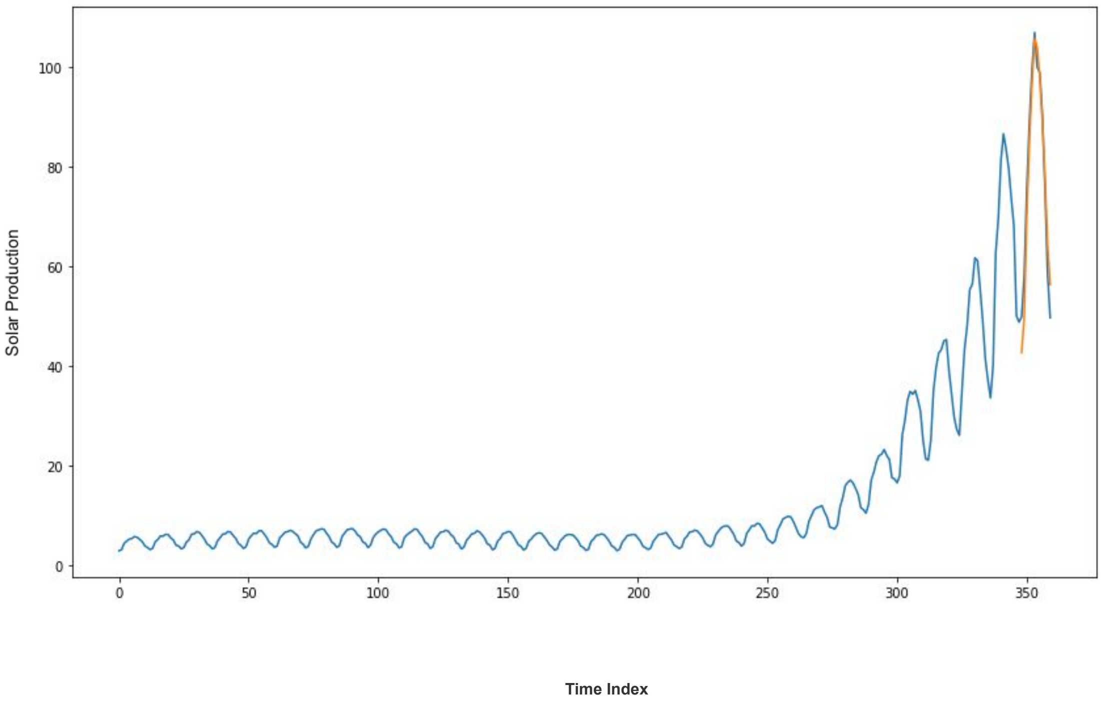
Recurrent Neural Network

Model Development

- Model Hyperparameters
  - Number of lags
  - Number of layers
  - Number of nodes/layers
- Compiler
  - Objective function: MSE
  - Optimizer: adam
- Regularizations
  - Dropouts

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_3 (LSTM)                (None, 36, 50)            10400
_________________________________________________________________
lstm_4 (LSTM)                (None, 50)                20200
_________________________________________________________________
dense_4 (Dense)              (None, 32)                1632
_________________________________________________________________
dropout_3 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_5 (Dense)              (None, 12)                396
=================================================================
Total params: 32,628
Trainable params: 32,628
Non-trainable params: 0
_________________________________________________________________
```

Solar Production Projection (Thousand MWH) from 1989 to 2018



Model Comparisons

Metrics	Random Walk - ARIMA (0,1,0)	ARIMA (0,1,1)(0,1,1)[12]	Auto - ARIMA (0,1,2)(0,1,1)[12]	ETS	TBATS	Regression with ARIMA Residuals (Smart)	Regression with ARIMA Residuals (Naive)	Prophet (Univ.)	Prophet (Mult.)	RNN
Avg. Cross Validation MSE	92.345	6.504	6.378	17.962	11.15	-	-	2.731	**2.528**	-
Avg. Cross Validation MAE	5.593	1.404	1.402	2.166	1.765	-	-	2.428	2.168	-
Forecast MSE	1335.065	48.324	54.273	172.31	101.132	58.300	30.336	12.279	**10.296**	23.239
Forecast MAE	30.5103	5.629	5.908	11.488	8.509	6.0672	4.607	11.032	7.805	**4.188**

Model Selection and Observations

- We select Multivariate Prophet as the best performing model.
    - Multivariate Prophet has the best ability to account for outliers in the data as evidenced by the minimal distance between its MSE and MAE scores
    - Multivariate Prophet’s forecast errors are the lowest, on average
- State Space models struggle in particular with modeling outliers (see: MSE scores)
- There is not enough complex seasonality in the models to required TBATS

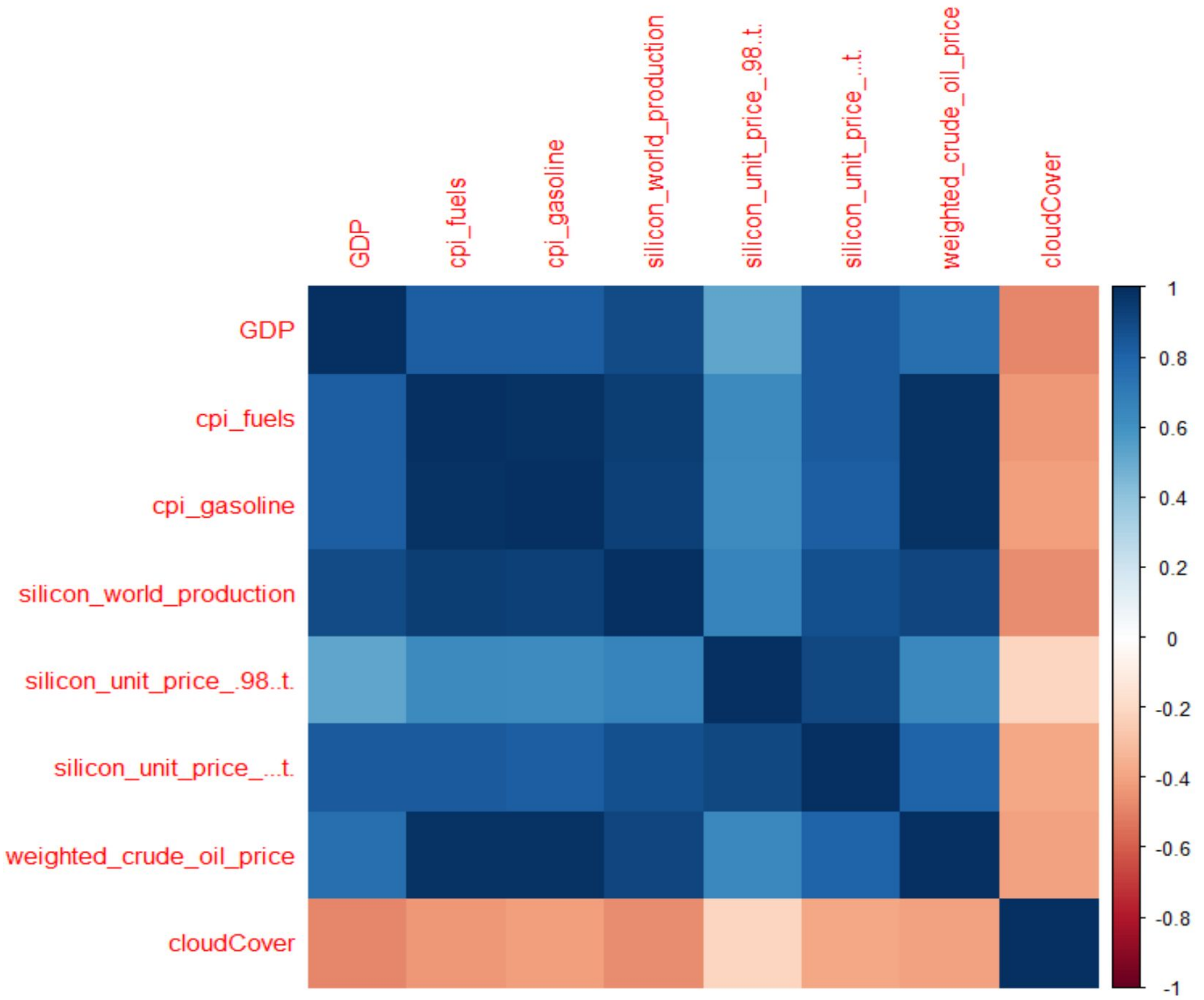
Future Work

- We note that in many of our models, we see a rise in the residuals during the last three to four years; as a result, we propose exploration of an intervention model and/or modeling the past ten years separately in order to more closely evaluate recent signals
- We also propose the addition of predictors to capture the energy output of competitor power sources (e.g. wind, hydroelectric, coal, natural gas)
- We propose the application of Bayesian Structural Time Series as an additional approach, given the success of Prophet (a Bayesian-based model)

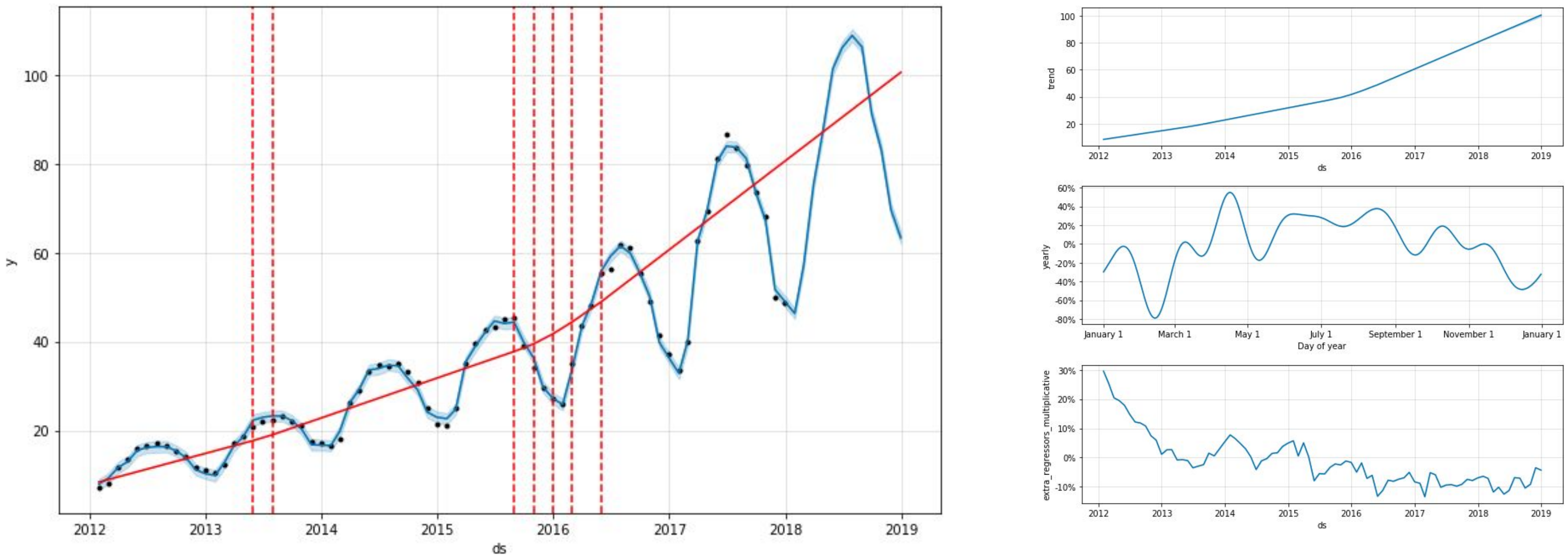
Appendix

Predictor Collinearity

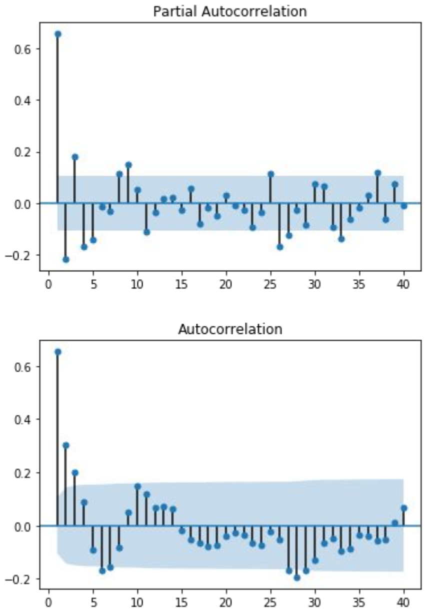
- Most predictors are clearly highly correlated with one another
- For this reason, we only selected a subset of predictors when developing any cross-sectional model we developed in our project.



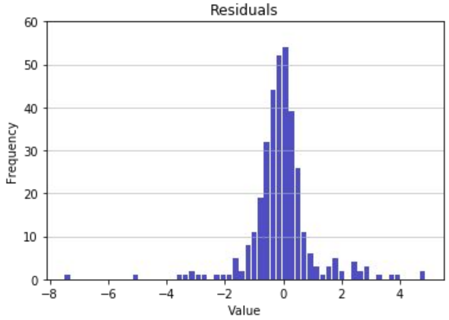
Prophet (2012-)



Prophet



Partial Autocorrelation
Autocorrelation



Residuals
Frequency
Value