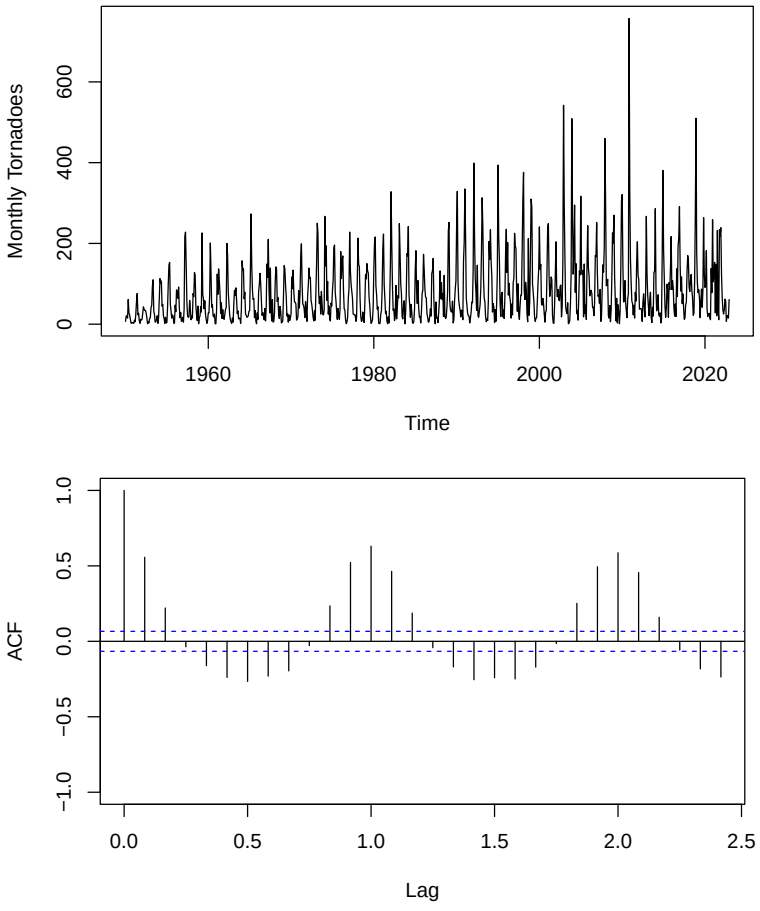
STAT 5550: Project Part 2

Nathan Honecker

Exploratory Data Analysis

Below, in `Figure 1`, we have a non-stationary time series, x_t, displaying counts of tornadoes that occurred each month from January of 1950 to December of 2022 in the United States. We can see a clear seasonal trend in the plot of the data and the month plot. There appears to be a global increasing trend over time with a non-constant mean, and the ACF slowly approaches zero, indicating non-stationarity. Also notice the non-constant variance in the data, indicating we should perform a log transformation.



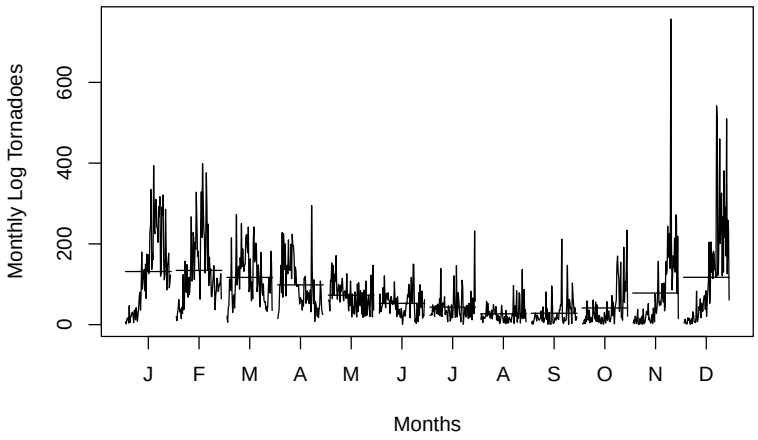


Figure 1: Plotted data x_t, ACF, and month plot.

Trend and Seasonality

In Figure 2, we have performed a log transformation and are left with $y_t = log(x_t)$, and we can see the results of fitting the data to a regression model with a quadratric term, as a simple linear regression did not capture the trend very well. That is, $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + w_t$ where $w_t \sim wn(0, \sigma^2)$. The red line is the trend while the blue is the detrended data with $\mu = 0$, or $y_t - \hat{\beta}_0 - \hat{\beta}_1 t - \beta_2 t^2$. Notice that the positive trend is slight, but present. Notice the variance looks to be relatively constant, and thus the log transformation served its purpose.

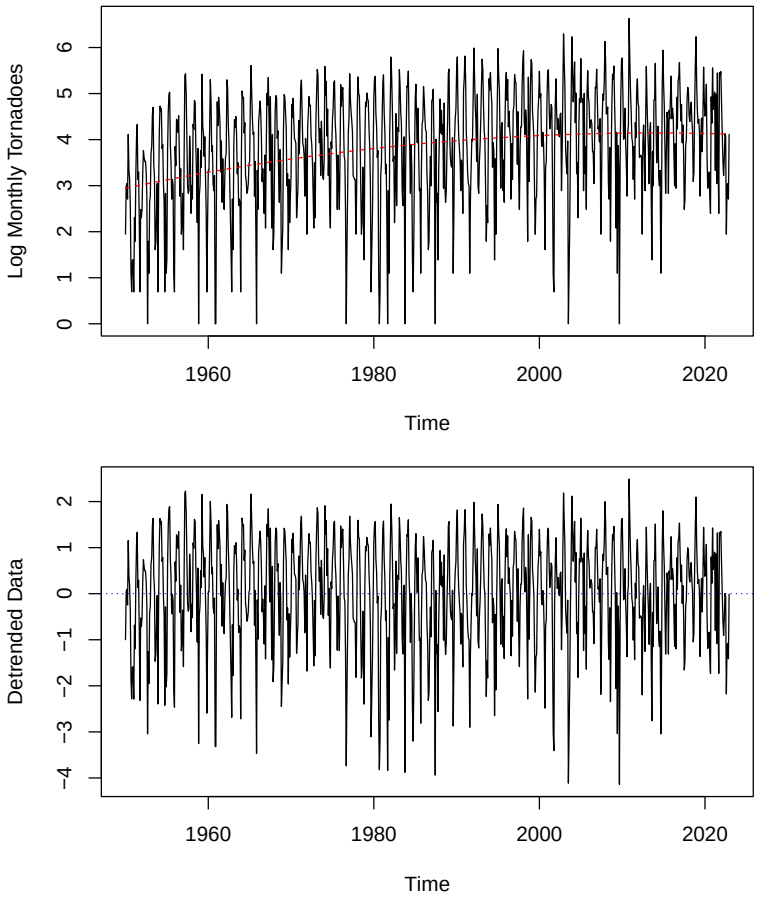


Figure 2: Log transformed data with trend fit and detrended data.

The seasonality proved difficult to capture. At first, I assumed a harmonic regression would work best, however it turns out that neither a harmonic regression nor a seasonal means captured the seasonality. We can see in Figure 3 below that the ACF of the seasonal means model displays seasonality, which means there is seasonality not captured. Note the seasonal means model captured the seasonal components the best out of all models that were fit. The seasonal means model:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \alpha_{m_t} + w_t$$

where $m_t \in \{1,\ldots,12\}$ represents the month. Then we have the "detrended" (but with still uncaptured seasonality) data modeled as:

$$z_t = y_t - \hat{\beta}_1 t - \hat{\beta}_2 t^2 - \hat{\alpha}_{m_t}$$

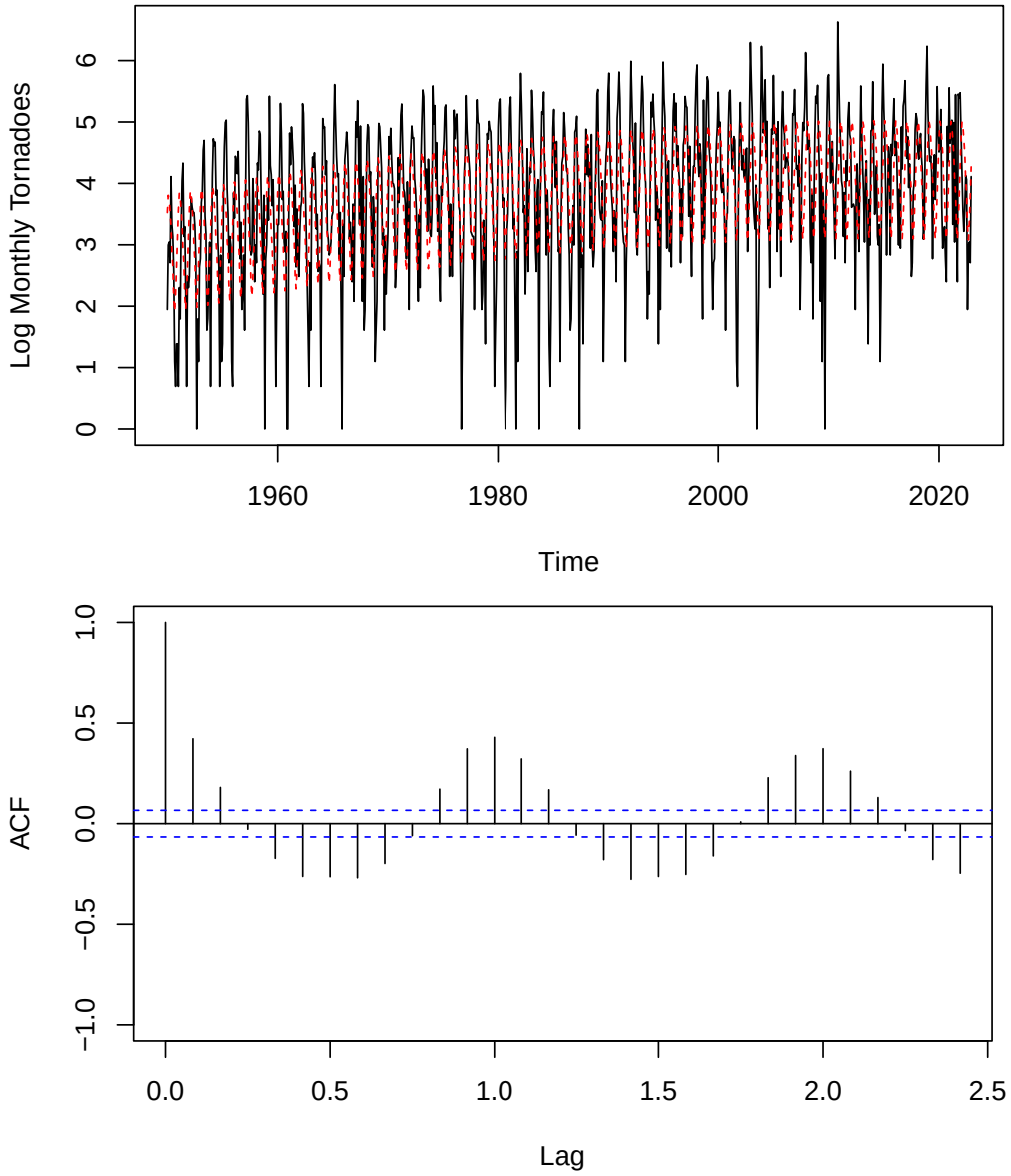


Figure 3: Seasonal means model and AFC.

SARIMA Modeling

Because we were unable to capture the seasonality through regression or smoothing, ARMA modeling was not successful. Looking at the ACF of our attempted detrended data, we cannot narrow down specific ARMA models to test, and every one I did test did not look adequate. Therefore we will consider the possibility that our data is stochastic and therefore must rely on differencing to appropriately capture the seasonality.

Below, in `Figure 4`, we see the data after taking one seasonal difference. This model is $\nabla_{12} y_t$, as we are differencing the log transformed data, y_t. We see a spike in the ACF at lag 12 and a spike in the PACF at lags 12, 24, and 36. This indicates that a SARIMA model is appropriate.

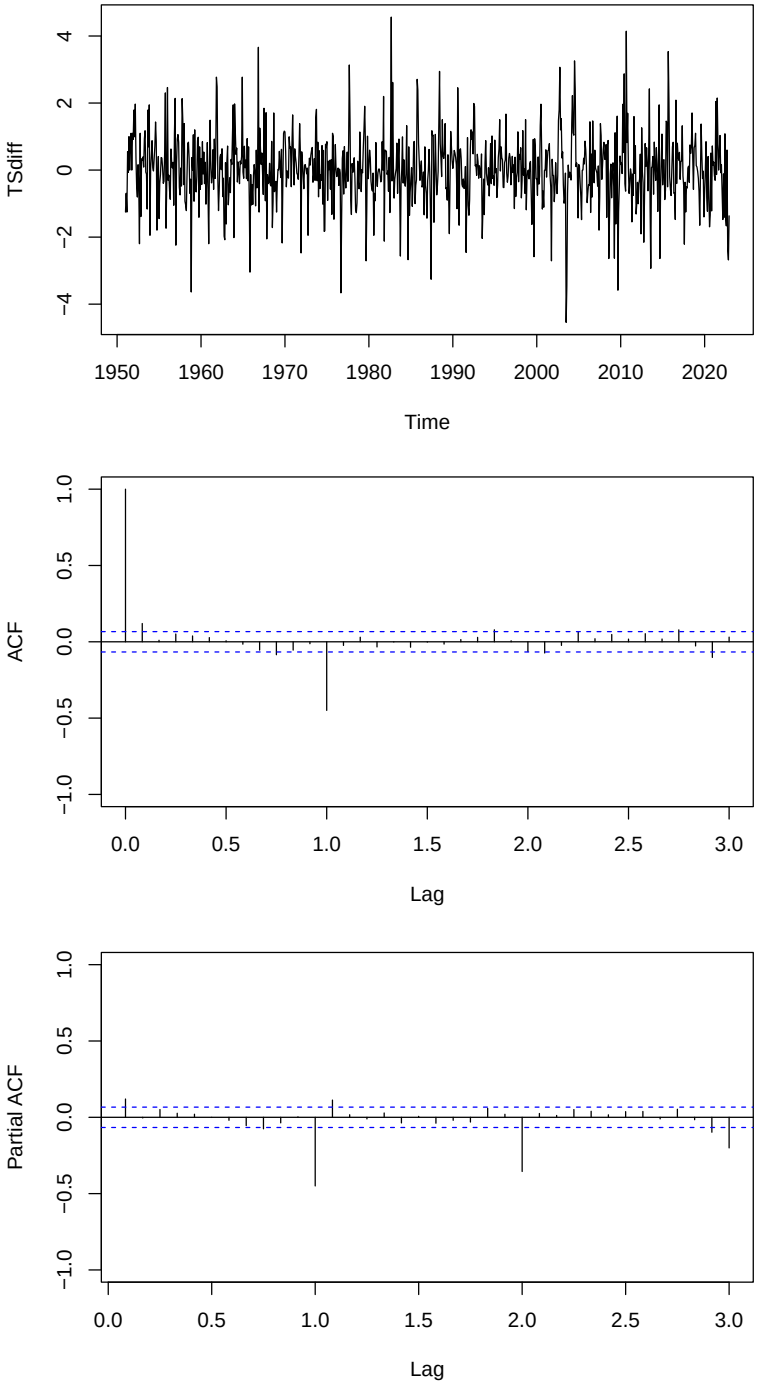


Figure 4: Differenced data, ACF, PACF.

The model that appears to fit best is ARMA$(1,1,1)$x$(0,1,1)_{12}$=

$$z_t = (1-\phi_1 B)\nabla_{12}\nabla_1 y_t = \delta + (1+\Theta_1 B^{12})(1+\theta_1 B)w_t$$

When capturing the linear trend of the data, we concluded a quadratic term was necessary to capture the overall trend. Considering this, it is no surprise that the model that fits best differences twice, one seasonal difference to capture the seasonality and another first order difference to capture the quadratic nature of the trend. Thus `D=1` and `d=1`. Based on the PACF, `p` is either 1 or 0 and `P` is most likely 0 since the lags at 12, 24, and 36 appear to be steadily decreasing. Based on the ACF, `q` is either 1 or 0 and `Q` appears to equal 1 since the ACF cuts off after `lag=12`.

```
Call:
arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
    include.mean = !no.constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = tr
        REPORT = 1, reltol = tol))

Coefficients:
         ar1      ma1     sma1
      0.1805  -0.9833  -0.8146
s.e.  0.0353   0.0097   0.0206

sigma^2 estimated as 0.6541:  log likelihood = -1050.51,  aic = 2109.03
```

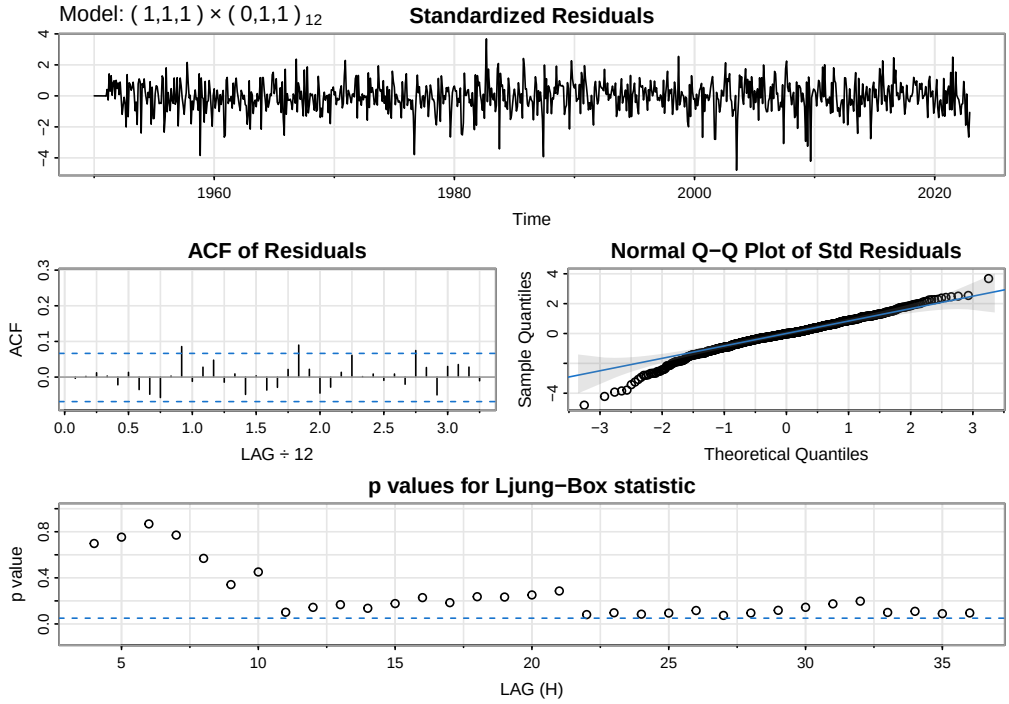


Figure 5: ARIMA(1,1,1)x(0,1,1)12 diagnostics.

The estimated model is:

$$z_t = (1 - 0.1805B)\nabla_{12}\nabla_1 y_t = \delta + (1 - 0.8146B^{12})(1 - 0.9833B)w_t$$

where $w_t \sim$ iid N(0,0.6541). In `Figure 6`, we see the diagnostic plots for the model. It appears to fit relatively well, but there are some concerns. The ACF of the residuals shows three autocovariance values outside of the acceptable range, which is more than we would like. The QQ plot isn't terrible but strays from the normal line at the lower tail of the data which could be cause for concern regarding the normality assumption. The Ljung-Box plot does indeed have all p-values above 0.05, but barely. We can conclude white noise.

Below is a table outlining all of the SARIMA models that were considered. We can see the ARIMA$(1,1,1)$x$(0,1,1)_{12}$ model is the best fit even though it does not have the lowest AIC. This is because all of its parameters are statistically significant.

Model	$\hat{\phi_1}(se)$	$\hat{\theta_1}(se)$	$\hat{\theta_2}(se)$	$\hat{\Theta_1}(se)$	AIC
ARIMA(1,1,1)x$(0,1,1)_{12}$	0.1805 (0.0353)	-0.9833 (0.0097)		-0.8146 (0.0206)	2109.03
ARIMA(1,0,1)x$(0,1,1)_{12}$	0.3315 (0.1866)	-0.1398 (0.1962)		-0.8092 (0.0213)	2107.22
ARIMA(1,0,2)x$(0,1,1)_{12}$	0.5743 (0.3307)	-0.3815 (0.3342)	-0.0554 (0.0801)	-0.8077 (0.0214)	2108.77
ARIMA(0,1,2)x$(0,1,1)_{12}$		-0.8091 (0.0331)	-0.1685 (0.0333)	-0.8128 (0.0205)	2110.06
ARIMA(1,1,2)x$(0,1,1)_{12}$	0.2243 (0.2051)	-1.0284 (0.2102)	0.0438 (0.2043)	-0.8148 (0.0206)	2110.98
ARIMA(0,1,1)x$(0,1,1)_{12}$		-0.9665 (0.0137)		-0.8022 (0.0199)	2132.7

Commentary

-ARIMA(1,1,1)x$(0,1,1)_{12}$: This model was chosen because its parameters are all statistically significant, its diagnostic plots look just as good as any of the other models, and its AIC is lowest among other models with statistically significant parameters and adequate diagnostics.

-ARIMA(1,0,1)x$(0,1,1)_{12}$: Neither $\hat{\phi}_1$ nor $\hat{\theta}_1$ are significant at the 95% level.

-ARIMA(1,0,2)x$(0,1,1)_{12}$: Neither $\hat{\phi}_1$ nor $\hat{\theta}_1$ are significant at the 95% level.

-ARIMA(0,1,2)x$(0,1,1)_{12}$: This is a decent model, as all of the parameters are statistically significant and the (not included) diagnostic plots look good. The only reason this model was not chosen is because its AIC is slightly higher than the other model that was chosen.

-ARIMA(1,1,2)x$(0,1,1)_{12}$: Neither $\hat{\phi}_1$ nor $\hat{\theta}_2$ are significant at the 95% level. This model had good diagnostics, it is clear the second differencing (d=1) is important in capturing the trend.

-ARIMA(0,1,1)x$(0,1,1)_{12}$: While all the parameters are statistically significant, the AIC is large and the Ljung-Box has nearly all p-values below 0.05.

Forecasting

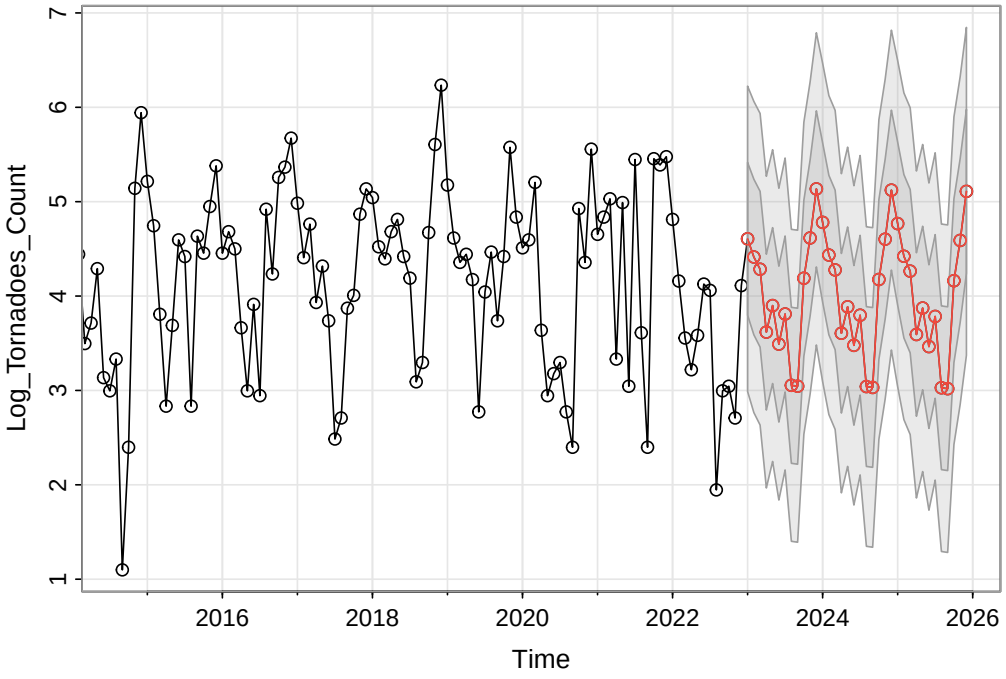


Figure 6: Forecasted z_t.

In Figure 6, we see a plot forecasting log tornado count using our SARIMA model, z_t. Note that the dark shaded region is a 95% prediction interval.

Model Comparison

While a quadratic trend seems to capture the global trend quite well, I was unable to reasonably estimate the seasonal components of the data using smoothing and regression methods. I tried fitting various moving average filters, a seasonal means model, and a harmonic regression model, but all of those left us with seasonal components unaccounted for and non-stationary ACFs. Because of this, no ARMA models were included in this report. While I did fit several ARMA models, their diagnostics were frankly unimpressive and were unable to achieve stationarity.

So with that, we have a SARIMA model, z_t, that appears to fit the data well enough. Considering the diagnostics, I would like the p-values in the Ljung-Box plot to be higher, the QQ plot appears acceptable but does give me some pause regarding the normality assumption, and the ACF has a few too many auto-covariances outside of the bounds, but overall it appears to capture trend and seasonality to the best of my ability given the tools at my disposal.

Conclusion

The tornado data is represented by a non-stationary time series, x_t, and in attempting to model that data, I performed a log transformation to achieve a more constant variance and meet normality assumptions, leaving us with y_t. The trend was modeled with a linear and quadratic term. I then attempted to estimate the seasonality of y_t using regression and smoothing techniques, but was unsuccessful. Because of this, I failed to find an appropriate ARMA model. From there, I was able to estimate the seasonality and trend by differencing, achieving a stationary time series, and thus search for the best SARIMA model. It became

clear that differencing twice more accurately captured the trend and seasonality, and I concluded that an ARIMA(1,1,1)x$(0, 1, 1)_{12}$ model, outlined above and referred to as z_t, is the best SARIMA model for this data.