Time Series Analysis: Dissolved Oxygen Levels in the Peconic River in Riverhead, Long Island: 2013-2016

Samantha Benedict
Professor Dana Sylvan
STAT 715
City University of New York, Hunter College

Abstract

The Peconic River has seen a fluctuation in water quality measures, specifically dissolved oxygen levels within the years of 2013 to 2015. In this project we study a time series analysis of the dissolved oxygen levels in the Peconic River and compare it to water temperature within the three year span. We propose a square root ARIMA(3,1,3) model and predict dissolved oxygen on the year 2016.

1. Introduction

1.1 Data

The data anlayzed in this project were collected by the United States Geological Survey. All water data were gathered at the County Road 105 bridge, in Riverhead, Long Island between 2012 and present. A water quality measurement device took recordings every six minutes and daily maximums, minimums, and averages were reported on a number of different measurements.[1] These include:

Dissolved Oxygen (mg/L)
Water Temperature (°C)
Estuary Elevation (Ft.)
Specific Conductance (uS/m)
Salinity (PSU)
Nitrate (mg/L)
Turbidity (FNU)
Chlorophyll (ug/L)
pH level (std. unit)

In this project, average daily dissolved oxygen and average daily water temperature were compared between the years of 2013 and 2015 simply because these years had the most complete data. There were still many missing data values, so the imputeTS package was employed in R, which utilizes time correlation to impute missing values based on the observations around it.

Prediction will later be discussed on dissolved oxygen levels for 2016 and compared to a test set of data already recorded for that year. All of the time series analysis performed in this project was done using R 3.6.1.

1.2 Background

The Peconic Bay Estuary System is located on the eastern end of Long Island between North and South Forks. **Figure 1** illustrates a map for geographical reference.[2] The fish and wildlife habitat is the freshwater portion of the river, which extends approximately 15 miles, from County Route 63 in the center of Riverhead, to the headwaters in Peconic River County Park. The entire length of the Peconic River is a productive habitat for warmwater fisheries.[3]

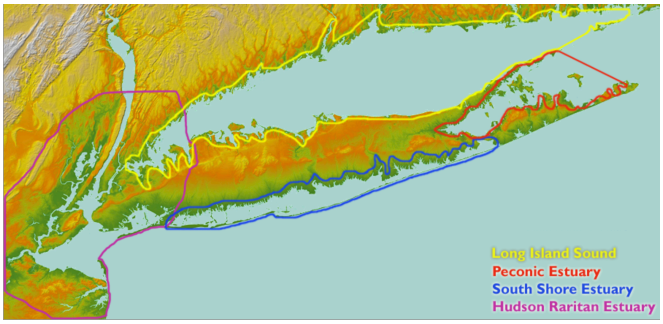


Figure 1 Map of Peconic Bay Estuary System

In order to recognize a time series analysis of dissolved oxygen levels, a definition of dissolved oxygen must be understood. Dissolved oxygen is the amount of gaseous oxygen (O_2) dissolved in the water. Oxygen enters the water by direct absorption from the atmosphere, by rapid movement, or as a waste product of plant photosynthesis. Adequate dissolved oxygen is important for good water quality and necessary to all forms of life. Dissolved oxygen levels that drop below 5.0 mg/L cause stress to aquatic life. Oxygen levels that fall below 1-2 mg/L may result in large fish kills.[4]

1.3 Exploratory Analysis

Presented in **Figures 2 and 3** are additive time series plots of average daily dissolved oxygen and average daily water temperature between January 1, 2012 and December 31, 2015 as well as their individual components removed for analysis.

We can see from the trend components in **Figures 2 and 3** that neither of these time series plots are stationary, both with time dependent means. As time increases, average daily dissolved oxygen decreases. The opposite is true for average daily water temperature.

This is confirmed by an Augmented Dickey Fuller test, which suggests a null hypothesis of non-stationarity. Performing this test on both of the aforementioned time series produced p-values greater than alpha equal to 0.01, so we fail to reject the null hypothesis and conclude non-stationarity.

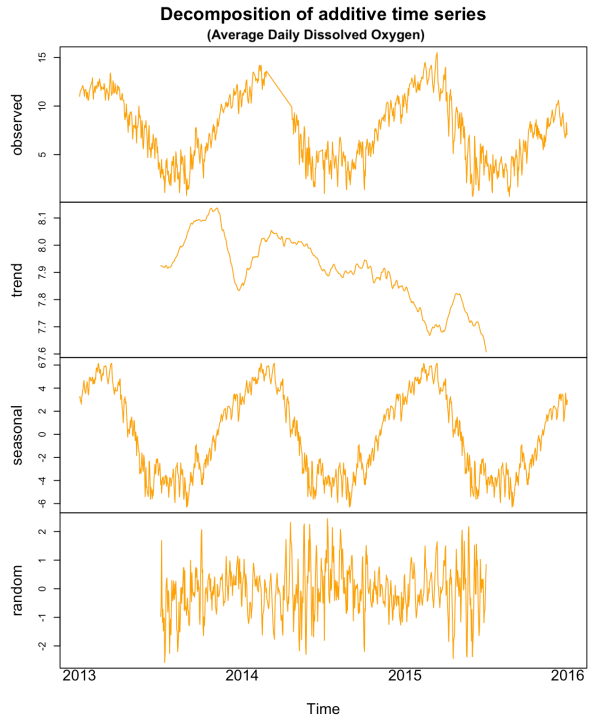


Figure 2 Time Series Plot of Average Daily Dissolved Oxygen

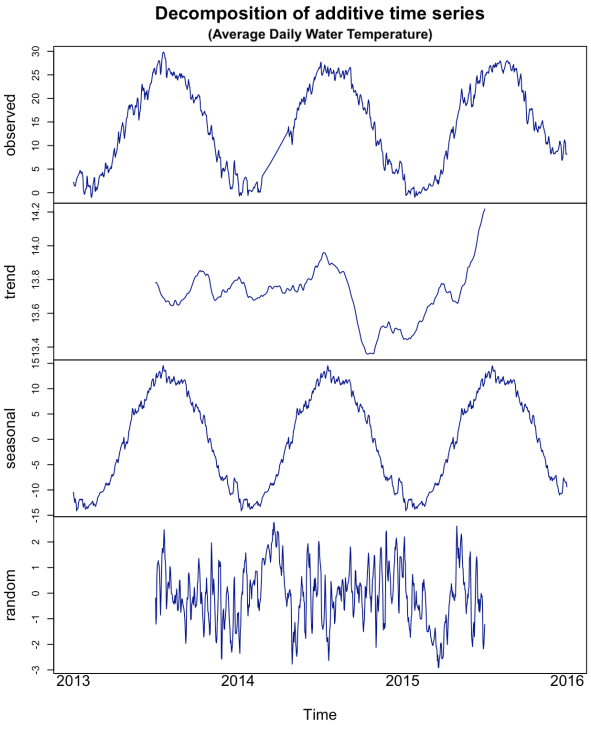


Figure 3 Time Series Plot of Average Daily Water Temperature

Since these time plots indicate a clear inverse relationship, both time series were plotted on the same plot for visual comparison as shown in **Figure 4.** You can see that in the colder months when water temperature is low, dissolved oxygen levels are high. Additionally, when water temperature is high, the dissolved oxygen levels fall below 5 mg/L which is considered the danger zone. Since we do see a negative trend, this pattern could be cause for concern in future years which is why a prediction model could be beneficial.

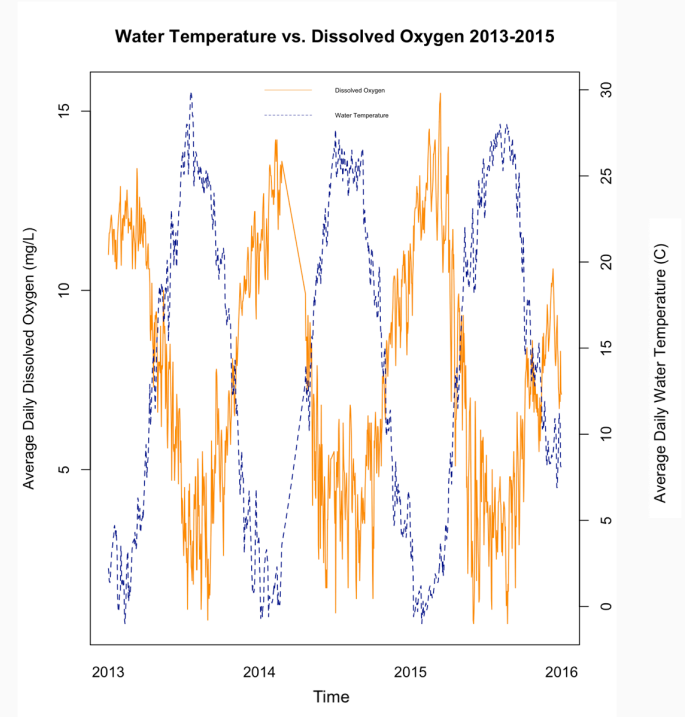
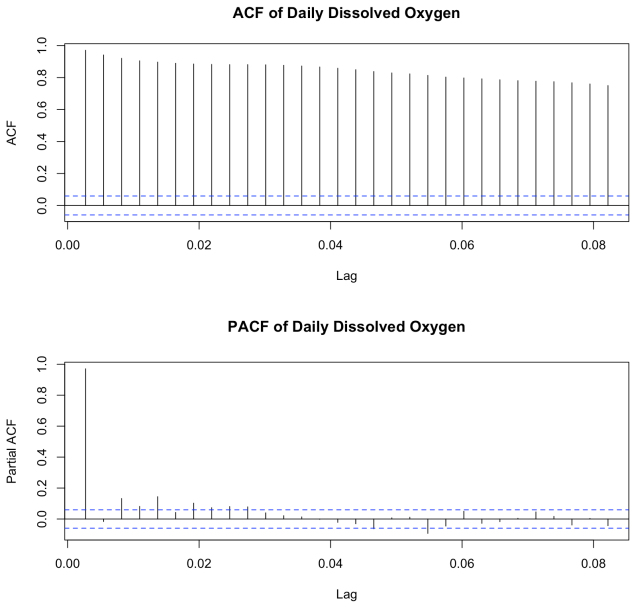


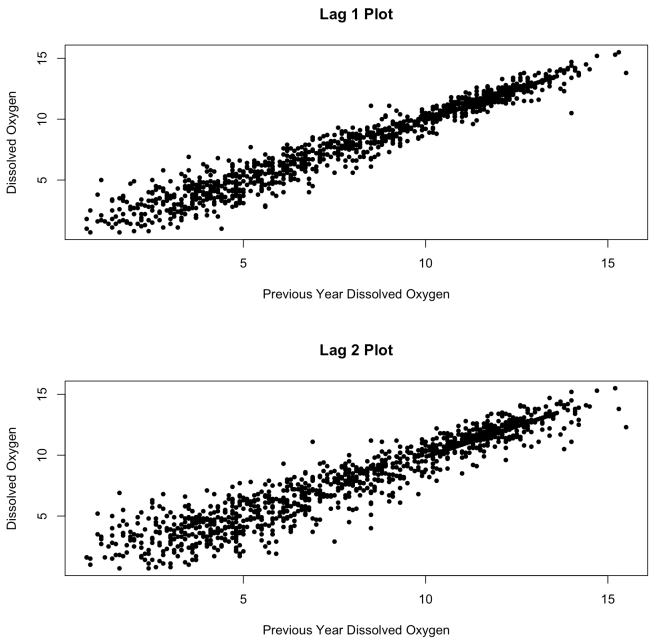
Figure 4 Time Series Dissolved Oxygen vs. Water Temperature

The ACF and PACF plots of average daily dissolved oxygen are shown in **Figures 5(a) and 5(b).** Both of these plots portray autocorrelation values that lie well beyond the 95% white noise confidence bounds.

Plots of dissolved oxygen versus dissolved oxygen at lags one and two can be seen in **Figures 6(a) and 6(b)**. These plots indicate that the correlation is still very strong at later time lags. Although not shown here, this correlation hardly weakened when plotted against lags three and beyond.



Figures 5(a) ACF (top) and (b)PACF (bottom) of Dissolved Oxygen

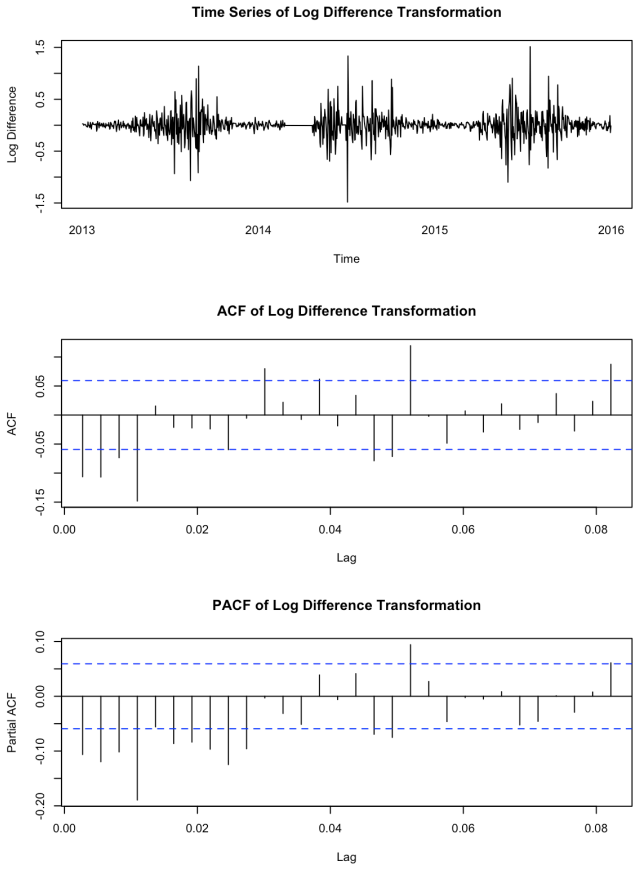


Figures 6(a) Lag 1 and (b) Lag 2 Plots of Dissolved Oxygen

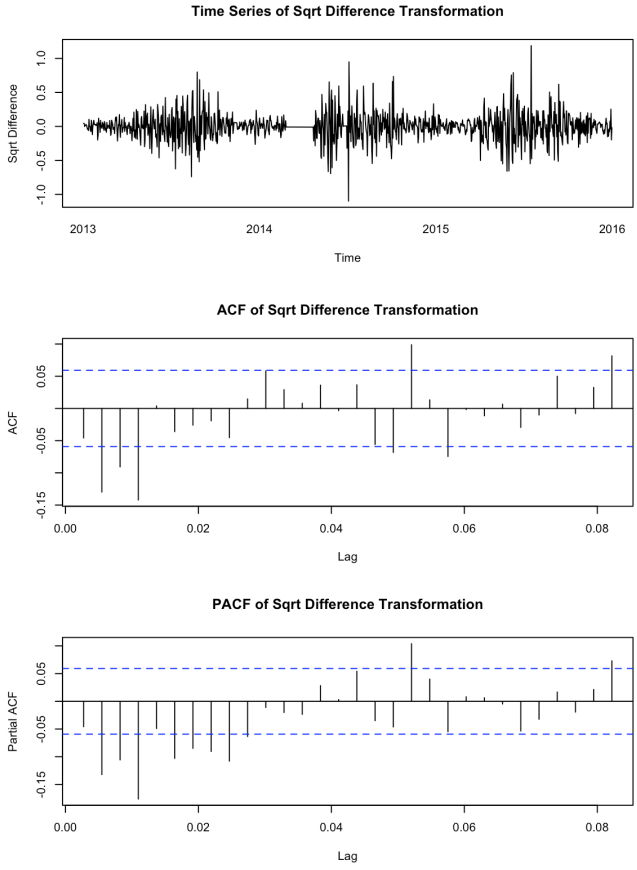
1.4 Transformations

In order to fit an appropriate time series model and be left with white noise residuals, we have attempted to first remove stationarity by way of differencing in combination with other possible transformations. After differencing just once, stationarity was successfully removed. Again, an Augmented Dickey Fuller test was performed to check this fact and a p-value of less than .01 confirmed that differencing did in fact remove time dependence in the mean function.

Additionally, some transformations were performed in effort to improve the ACF and PACF plots. Both the log transformation and the square root transformations were tested on the differenced time series. **Figures 7(a)-(c) and 8(a)-(c)** demonstrate the effects of these transformations.



Figures 7(a)-(c) Time Series, ACF, and PACF of Log Difference (respectively)



Figures 8(a)-(c) Time Series, ACF, and PACF of Square Root Difference (respectively)

As you can see in **Figures 7(a) and 8(a)** both the square root difference and log difference transformations have removed non-stationarity but we can still see seasonal clusters. Additionally, Figures **7(b) and (c) and 8(b) and (c)** show very similar ACF and PACF plots for both of the transformations. While autocorrelation values still exist outside of the white noise confidence bounds, these transformations did improve the plots. After some trial and error during model selection, as well as a slightly superior time series plot, the square root transformation proved to be the most influential in fitting an appropriate model.

2. Methods

2.1 Model Specification

This time series proved to be particularly difficult when attempting to fit a proper model that would remove the autoregressive and moving-average components present in our series and leave behind only white noise residuals.

While the dissolved oxygen time series exhibits clear seasonality, when fitting the integrated autoregressive–moving-average (ARIMA) models with seasonality parameters included, they performed poorly in comparison to those without. This poor performance was seen in the ACF and PACF plots, when testing the model residuals for white noise, and overall, they produced inadequate forecasts regardless of their fairly competitive AIC scores.

After experimenting with many ARIMA models, the best fitting model was proven to be the ARIMA(3,1,3) on the square root transformation without seasonality parameters included. This model is presented in **Eq. 1.** Overall, it had superlative ACF and PACF plots and a competitively low AIC score. Some of the models that were explored can be seen in **Table 1**, which lists the type of transformation applied, the ARIMA model tested, and the corresponding AIC score.

Transformation	ARIMA Model (Seasonality Parameters)	AIC
None	3 1 3	2629.88
Square Root	3 1 3	-542.59
	2 1 2	-530.24
	3 1 2	-532.86
	0 1 3	-511.59
	2 1 3 (2 0 2)	-528.12
	2 1 2 (1 0 1)	-530.75
Log	1 1 2	-360.37
	3 1 3	-363.13
	2 1 2	-359.16
	0 1 2	-305.91
	3 1 2 (1 1 1)	-305.22

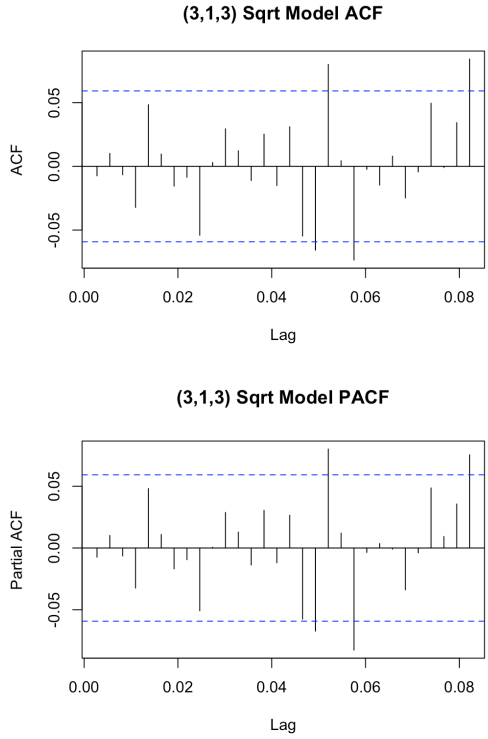
Table 1 Tested ARIMA models with Corresponding AIC Scores

$$\nabla \mathbf{Y} = \mathbf{0.7534}(\mathbf{Y_{t-1}} - \mathbf{Y_{t-2}}) + \mathbf{0.4635}(\mathbf{Y_{t-2}} - \mathbf{Y_{t-3}}) - \mathbf{0.4831}(\mathbf{Y_{t-3}} - \mathbf{Y_{t-4}}) - \mathbf{0.8788}\varepsilon_{t-1} - \mathbf{0.5856}\varepsilon_{t-2} + \mathbf{0.5799}\varepsilon_{t-3} + \varepsilon_t$$

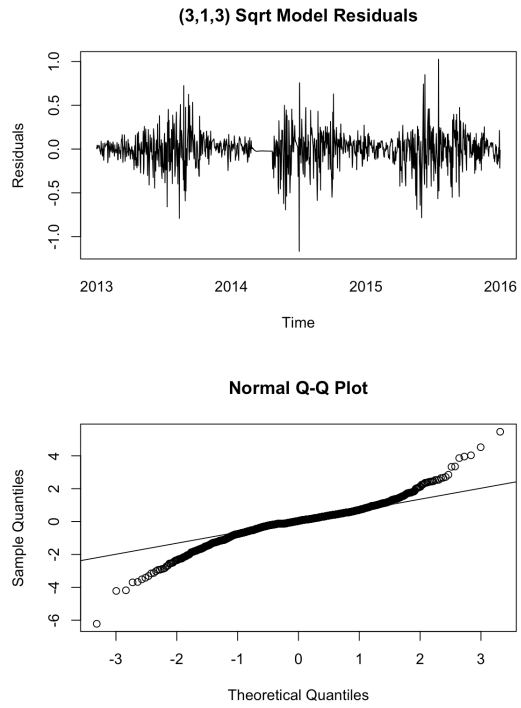
Eq. 1 ARIMA(3,1,3) Model

2.2 Model Fitting & Diagnostics

In order to determine statistically if the ARIMA(3,1,3) model is a well-fitting model, a Ljung-Box test was evaluated. This test suggests a null hypothesis that our model *does not* show a lack of fit. With a p-value equal to 0.5339 we are unable to reject the null hypothesis in this case and therefore can confirm that this model *does not* show a lack of fit, or rather, the model is an appropriate one.



Figures 9(a) ACF (top) and (b) PACF (bottom) of residuals on ARIMA(3,1,3)



Figures 10(a) Residual Plot (top) and (b) QQ Plot (bottom) of residuals on ARIMA(3,1,3)

In order to confirm whether or not the trend was modeled properly, leaving behind only white noise, the ACF and PACF plotted of the residuals on the ARIMA(3,1,3) model can be seen in **Figures 9(a) and (b)**. It can be determined that the autocorrelation function at most lag values is much better contained within the 95% confidence interval than prior to the model fitting, but still slightly surpasses the bounds. This indicates that marginally high correlations at later lags still exist so perhaps the residuals are not completely white noise.

In Figures 10(a) and (b) the residuals have been plotted against time and a normal QQ plot has been constructed to test the normality assumption of the error terms. It can be seen in **Figure 10(a)** that some volatility clusters exist and suggest that this model is not a flawless one. The normal QQ plot shown in **Figure 10(b)** reveals moderately heavy tails and ultimately a betrayal from the normality assumption in the residuals. Nonetheless, prediction was attempted.

2.3 Forecast

It is possible that since the ARIMA(3,1,3) model does not completely produce white noise residuals, a forecast using this model would lead to inaccurate results. Nevertheless, prediction was performed utilizing the ARIMA (3,1,3) model to predict average daily oxygen levels for the first 15 days of January 2016 and validated against a test set of already recorded data from that same time frame.

Figure 11 shows the plotted time series of dissolved oxygen with the square root transformation narrowed in on the end of 2015 and early part of 2016 to make each specific prediction value easier to see. The blue x's in **Figure 11** indicate the 15 predicted values of the square root of dissolved oxygen in January of 2016. The dark grey area represents the 80% confidence bounds on the prediction levels and the lighter grey represents the 95% bounds. Most all of the red triangles, which represent the corresponding

observed values from our test set, exist within the prediction interval. Therefore, we can consider the forecast an acceptable one.

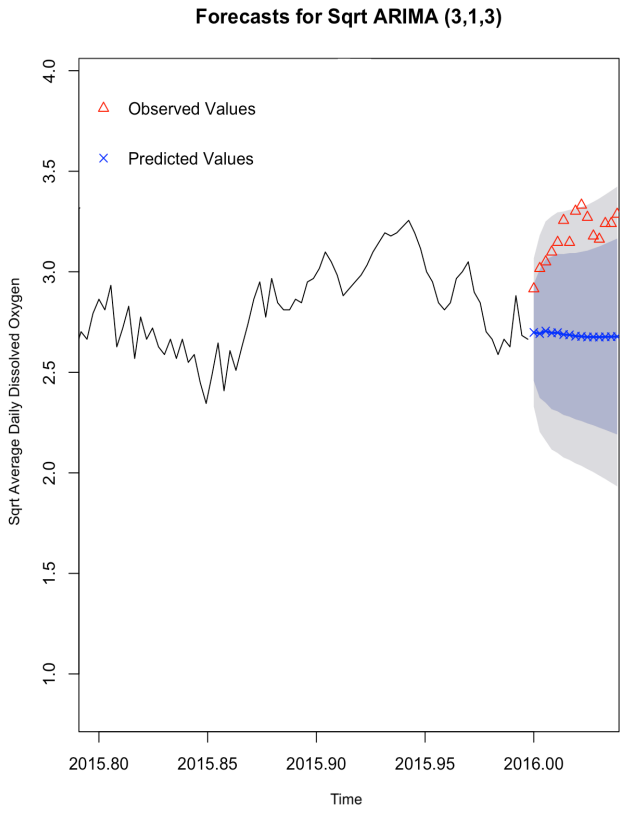


Figure 11 Forecast vs. Observed Values

To expand on interpretability, **Table 2** has been developed with all the observed and predicted values of transformed dissolved oxygen converted back to their original form (mg/L). The second to last column in **Table 2** signifies the sum of the squares of the forecast values minus the observed values averaged over five, ten and fifteen predictions. Additionally, the last column provides the percent error for each of the groups of predictions five, ten and fifteen.

It is clear from the table that with percent errors between 20 and 30%, prediction could be improved. When predicting the first five observations, the error is best however as we predict further out in the series, as expected, the percent error increases.

Dissolved Oxygen		$\frac{\sum(\text{Forecast}-\text{Obs.})^2}{\text{No. of Obs.}}$	% Error
Forecast	**Observed**		
7.280872	8.5	**Averaged over 5 Observations** 4.232189	**21.36 % error**
7.250453	9.1		
7.318485	9.3		
7.267636	9.6		
7.275591	9.9		
7.225283	10.6	**Averaged over 10 Observations** 8.152729	**26.83% error**
7.215579	9.9		
7.181220	10.9		
7.175070	11.1		
7.159221	10.7		
7.160994	10.1	**Averaged over 15 Observations** 8.909309	**28.19 % error**
7.157955	10.0		
7.164141	10.5		
7.166537	10.5		
7.172681	10.8		

Table 2 Forecast vs. Observed Values for ARIMA(3,1,3) Model

3. Results and Conclusion

In general, it appears as though average daily dissolved oxygen increases in the winter months when water temperature is low. The ARIMA(3,1,3) model without seasonal adjustment did an unexceptional job at predicting average daily dissolved oxygen levels for January 2016.

Although the ACF and PACF plots for the model show minimal lag values outside the confidence bounds, the model diagnostics still indicate slightly volatile residual clusters and a departure from normality, as shown from the QQ plot.

Perhaps these departures from model assumptions lead to the limitations and inaccuracy in our forecasted data. Future work could include further exploration in the modeling specification section with revisitation to the seasonal ARIMA component in order to find a model with white noise residuals. This could improve the model diagnostics and ultimately provide more accurate forecasting values.

Since we have seen a pattern of decreasing dissolved oxygen through time, working to produce a more accurate and substantial forecast would be highly beneficial in protecting wildlife inhabiting the Peconic River going forward.

References

[1] Data Description: USGS 01304562 Peconic River at County HWY 105 AT Riverhead NY. (n.d.). Retrieved May 06, 2020, from https://waterdata.usgs.gov/ny/nwis/inventory/?site_no=01304562

[2] Long Island's Estuaries. (n.d.). Retrieved May 06, 2020, from http://www.seagrassli.org/conservation/our_estuaries.html

[3] Coastal Fish & Wildlife Habitat Assessment Form. (2002). Retrieved May 06, 2020, from https://web.archive.org/web/20050506053024/http://www.nyswaterfronts.com/downloads/pdfs/sig_hab/LongIsland/Peconic_River.pdf

[4] Indicators: Dissolved Oxygen. (2016, August 16). Retrieved May 06, 2020, from https://www.epa.gov/national-aquatic-resource-surveys/indicators-dissolved-oxygen