Final Project

Water Temperature

Zhanylai Turatkhan kyzy, Julia Kagiliery, Yilun Zhu

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x lubridate::stamp() masks cowplot::stamp()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following object is masked from 'package:purrr':
##
##     lift
##
##
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
##
##
## Attaching package: 'zoo'
##
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
##
## Loading required package: greybox
##
## Registered S3 method overwritten by 'greybox':
##   method     from
##   print.pcor lava
##
## Package "greybox", v2.0.0 loaded.
##
##
```

```
##
## Attaching package: 'greybox'
##
##
## The following object is masked from 'package:caret':
##
##     MAE
##
##
## The following object is masked from 'package:lubridate':
##
##     hm
##
##
## The following object is masked from 'package:tidyr':
##
##     spread
##
##
## This is package "smooth", v4.0.0
```

Introduction Marine microbe account for 50% of marine primary production and dominate the biomass of the ocean. These organism are vital for life on earth and ocean health but are at serious risk in high temperatures worsened by climate change. The first step in supporting these ecological communities and hence mitigating the damages done by climate change is to understand what future climate scenarios look like. Predictions for future climate and temperature can varry by up to 10 degrees (if not more). With so much uncertainty surrounding climate future, it is of the utmost importance to have as much acurate modeling for future climate scenarios and hence our group hopes to make reasonable forecasts of surface water temperature at the Duke University Marine Lab research facility.

Motivation/relevance of the study Global ocean temperature is on the rise which has significant implications for this dynamic and productive biome. Among the many services the ocean provides, primary production (of oxygen) is among the most important. In fact, approximately 50% of primary production of oxygen comes from marine phytoplankton. It has been proven (largely through work by Duke University Marine Lab's Dr. Zacakry Johnson and Dr. Dana Hunt) that these microbial systems change significantly with seasonal changes (which include dominantly temperature changes, but also isolation and day length changes). These changes alter the way that the ocean cycles nutrients, stores carbon, and produces oxygen. Hence, temperature is an important variable in the consideration of much ocean modeling. This highlights the importance of accurate temperature prediction. Understanding what future climate and temperature looks like allows for better prediction of what other ocean cycles will look like and how we may excpet the serives we recieve from the ocean to change with the climate.

Objectives

Our objective is to accurately model the monthly temperature of Piver's Island Coastal Observatory and produce reasonable forecasting at appropriate time scales.

Dataset Information

The following data comes from a long running time series study out at Piver's Island Coastal Observatory (PICO) which is located at the Duke University Marine Lab conducted by Drs. Zackary Johnson and Dana Hunt. The sytudy monitors the ambient conditions such as turbidity, pH, temperature, and salinity. For this study, only temperature was able to be included. The temperature is reported as mean monthly temperature.

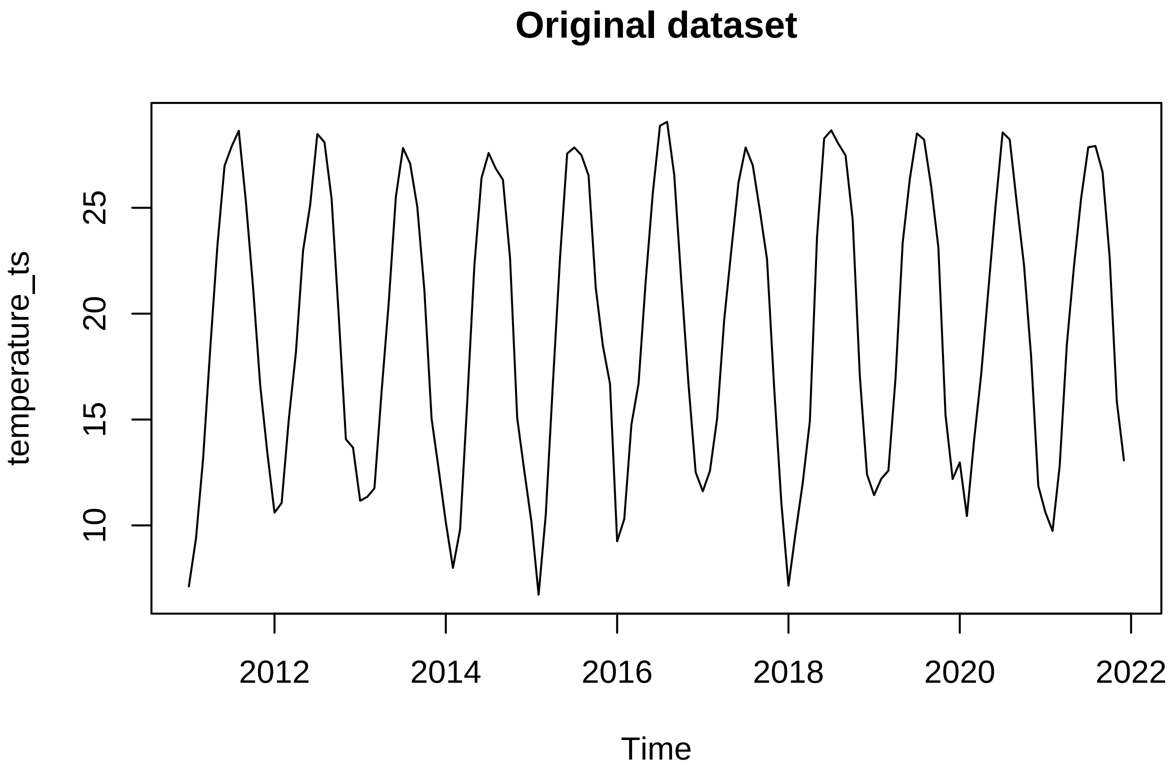
Methodology / Analysis

There are first a few important considerations as we get into modeling. The first is that this data is actually measured approximately weekly, meaning sometimes the lab samples multiple times in a week or skips a week.

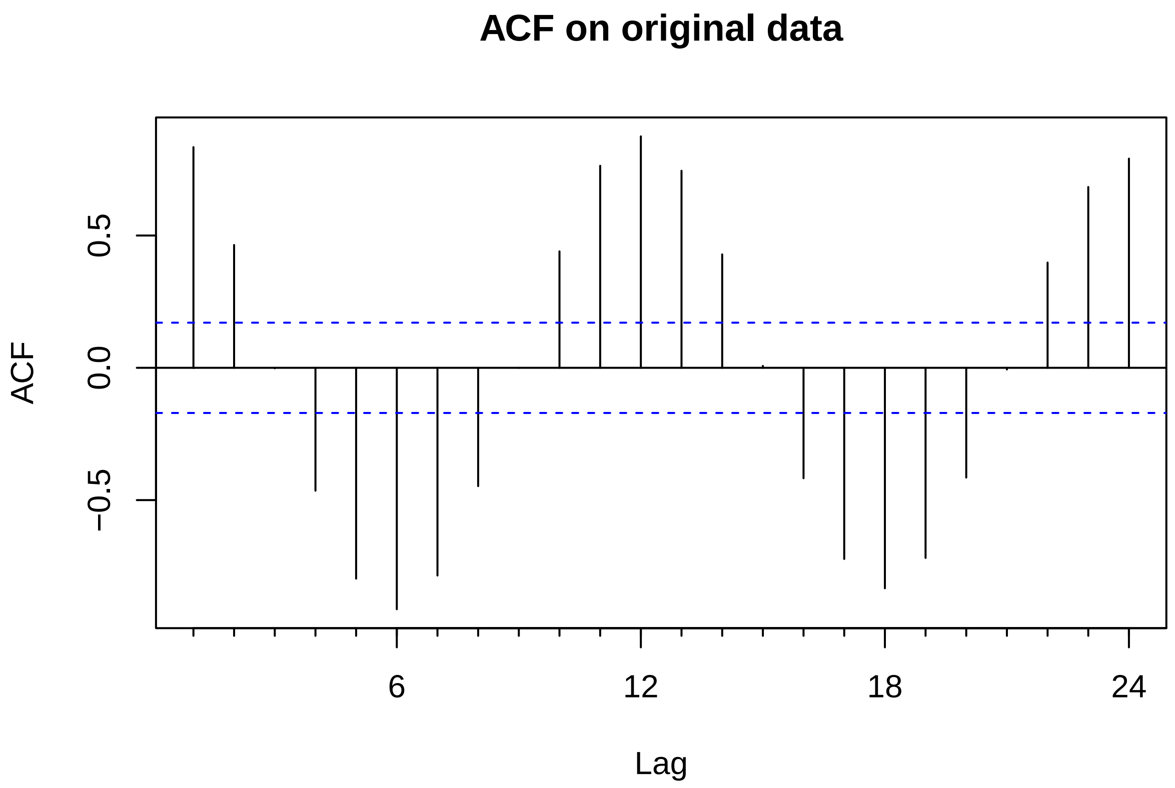
In order to account for this irregular time series frequency, we aggregated the data into monthly means to avoid unnecessary complications and misalignment of our time series. The second consideration we would like to acknowledge that climate change is certainly a a factor that plays a role in shifting trends, seasonal components, and general variability which may not be accurately reflected in our current data set which spans only approximately 11 years. Though the temperature in the region looks highly predictable, future predictions must be cognizant that large prediction horizons are unreasonable.

Description

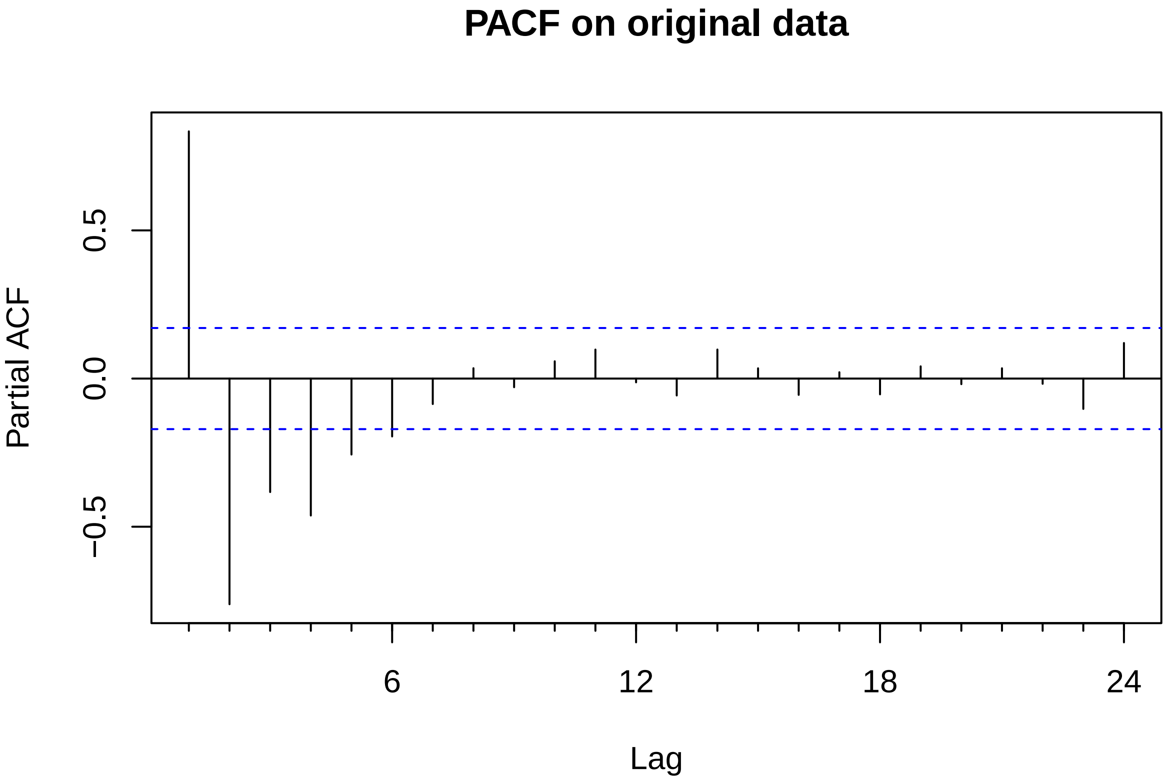
1. SARIMA model: The Seasonal Autoregressive Integrated Moving Average (SARIMA) model expands upon the ARIMA framework to address seasonality in univariate data sets. It incorporates both non-seasonal and seasonal elements in its predictions, allowing it to capture complex patterns that recur over fixed periods. **2. Arima+Fourier:** This approach combines the ARIMA model with Fourier series to enhance the modeling of time series with complex seasonal patterns. ARIMA captures the autocorrelations in the data, while the Fourier terms allow for the approximation of seasonal cycles of various lengths and complexities. This hybrid model is especially useful when the seasonality is not strictly periodic or involves multiple frequencies. **3. TBATS model:** Designed for forecasting time series with intricate seasonal patterns, the TBATS model employs exponential smoothing as its core technique. It thoroughly explores a variety of specifications, including those with and without a Box-Cox transformation, the presence or absence of a trend, trend damping options, and an ARIMA(p,q) component for the residuals. The model also evaluates different harmonic levels for seasonalities. The best-fitting model is determined by the lowest Akaike Information Criterion (AIC). **4. ETS model:** This univariate forecasting method, known as Exponential Smoothing State Space Model (ETS), emphasizes trend and seasonality components within the time series data. It is particularly adept at capturing patterns that evolve over time. **5. SSES model:** The State Space Exponential Smoothing (SSES) model extends the classic exponential smoothing approach by incorporating distribution assumptions about the error terms, which aids in the computation of prediction intervals. It considers both additive and multiplicative error structures within the state space modeling framework. **6. Neural Network:** The Neural Network model facilitates the identification of complex and nonlinear relationships between the dependent variable and its predictors. Its versatile architecture can adapt to a wide array of data patterns.



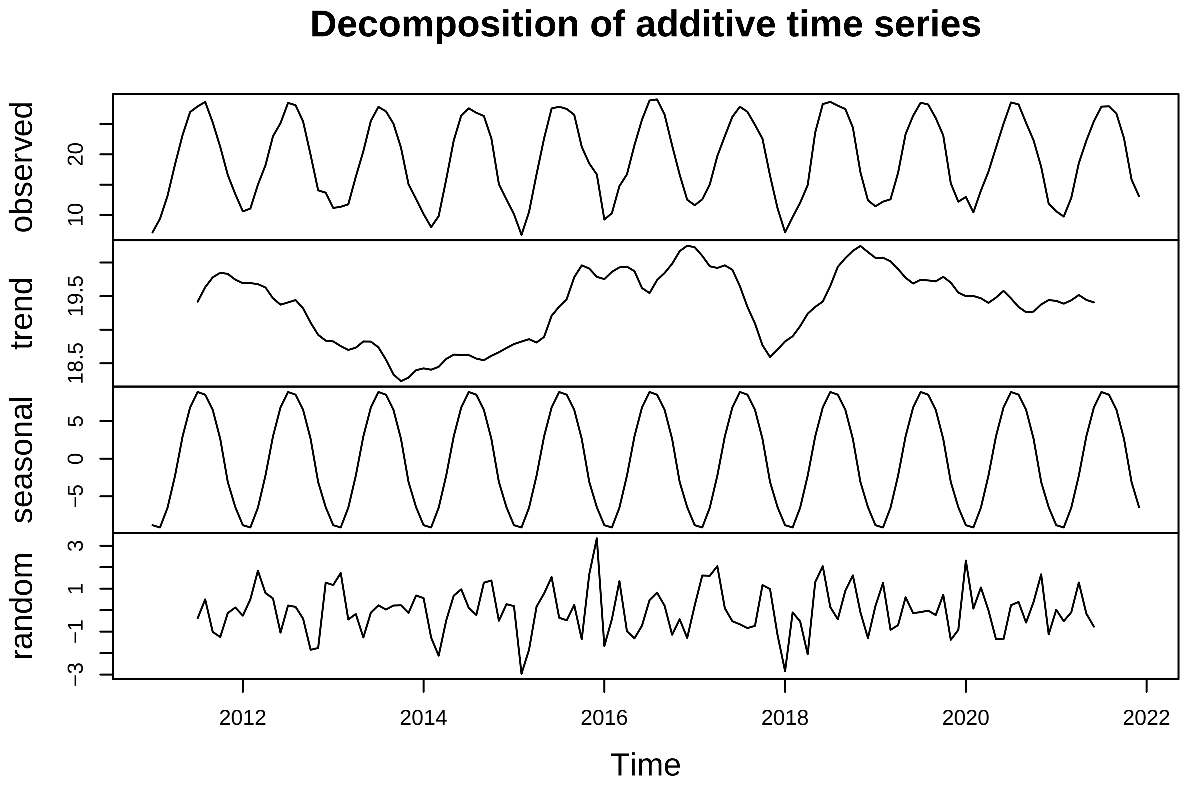
The temperature shows cyclical behavior with peaks and troughs that correspond to expected seasonal variations. The amplitude of these oscillations appears consistent over the years, indicating a stable seasonal effect without significant year-over-year changes in peak or trough temperatures. From a visual inspection, there are no apparent outliers or disruptions in the seasonal pattern, suggesting that the dataset is clean and well-maintained. The regularity of the pattern would likely make it suitable for forecasting using seasonal models, such as SARIMA or TBATS, which could exploit the periodicity inherent in the data.



The graph illustrates the Autocorrelation Function (ACF) applied to the original temperature data set. Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. The ACF is plotted against various lag values, which span from 0 to 24. The y-axis represents the autocorrelation coefficient, ranging from -0.5 to 0.5, while the x-axis represents the lag in terms of the number of time units. Each vertical bar corresponds to the autocorrelation coefficient at a specific lag. Notably, the graph shows a pattern of spikes at regular intervals, which suggests a seasonal pattern in the data. The presence of these spikes at consistent intervals can be indicative of the seasonality in the dataset, which correlates with the seasonal fluctuations seen in the time series plot of the original dataset. The blue dashed lines represent the significance bounds. Any spike that extends beyond these bounds is considered statistically significant. In this ACF plot, several lags have autocorrelation values that cross the significance threshold, confirming that the data exhibit a strong seasonal component at these lags.



In this PACF plot, most of the spikes are within the significance bounds, suggesting that most of the autocorrelations in the original data can be accounted for by the immediate preceding values.



```
## Score =  407 , Var(Score) = 258418.3
## denominator =  8645.5
## tau = 0.0471, 2-sided pvalue =0.42448
## NULL
```

```
## Warning in adf.test(temperature_ts, alternative = "stationary"): p-value
## smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  temperature_ts
## Dickey-Fuller = -11.153, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

The Augmented Dickey-Fuller Test result indicates a Dickey-Fuller statistic of -11.153 with a lag order of 5, and a p-value of 0.01, suggesting that the null hypothesis of a unit root can be rejected and the time series is stationary.

```
##  Time-Series [1:132] from 2011 to 2022: 7.12 9.38 13.16 18.37 23.2 ...
```

For the forecasting purposes, the data was split into training and test data following a proportion of 80/20.

```
## [1] 2011    1
```

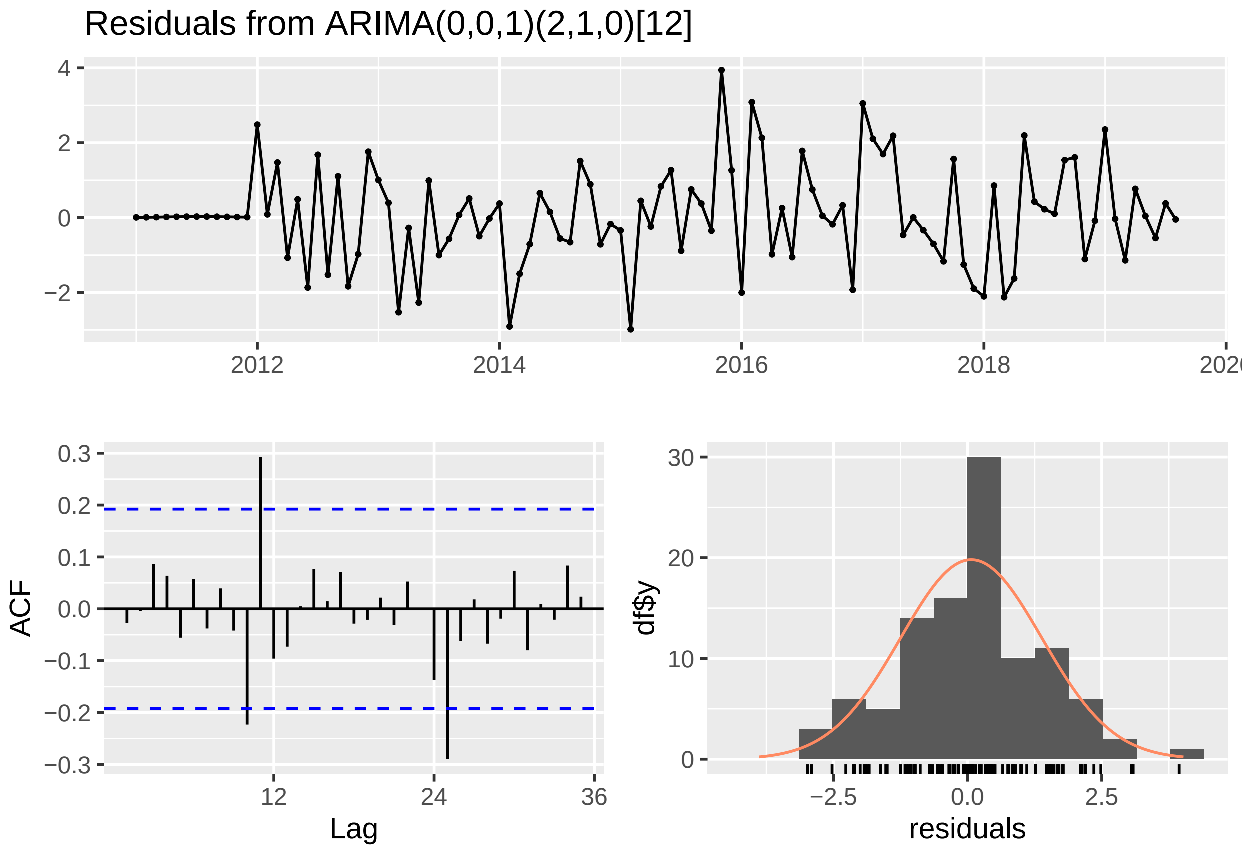
```
## [1] 2019    8
```

```
## [1] 2019    9
```

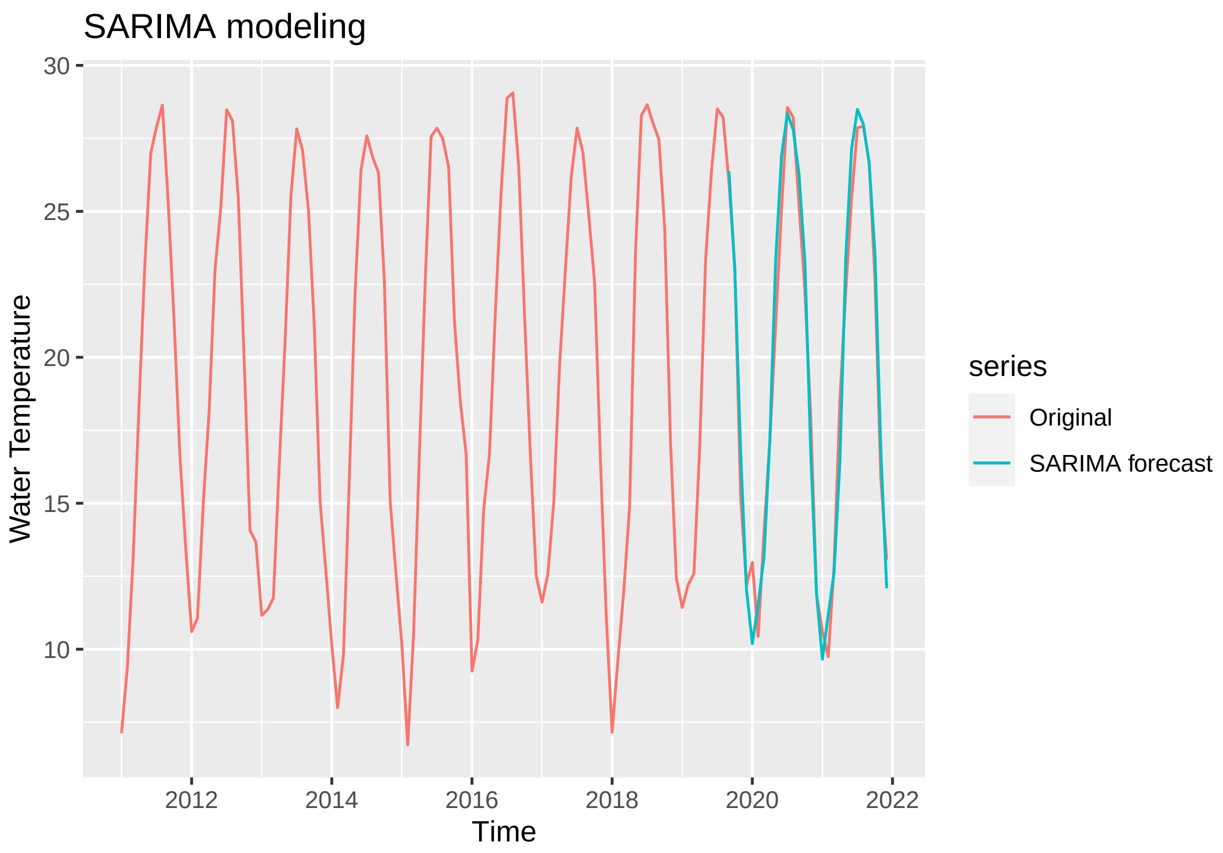
```
## [1] 2021   12
```

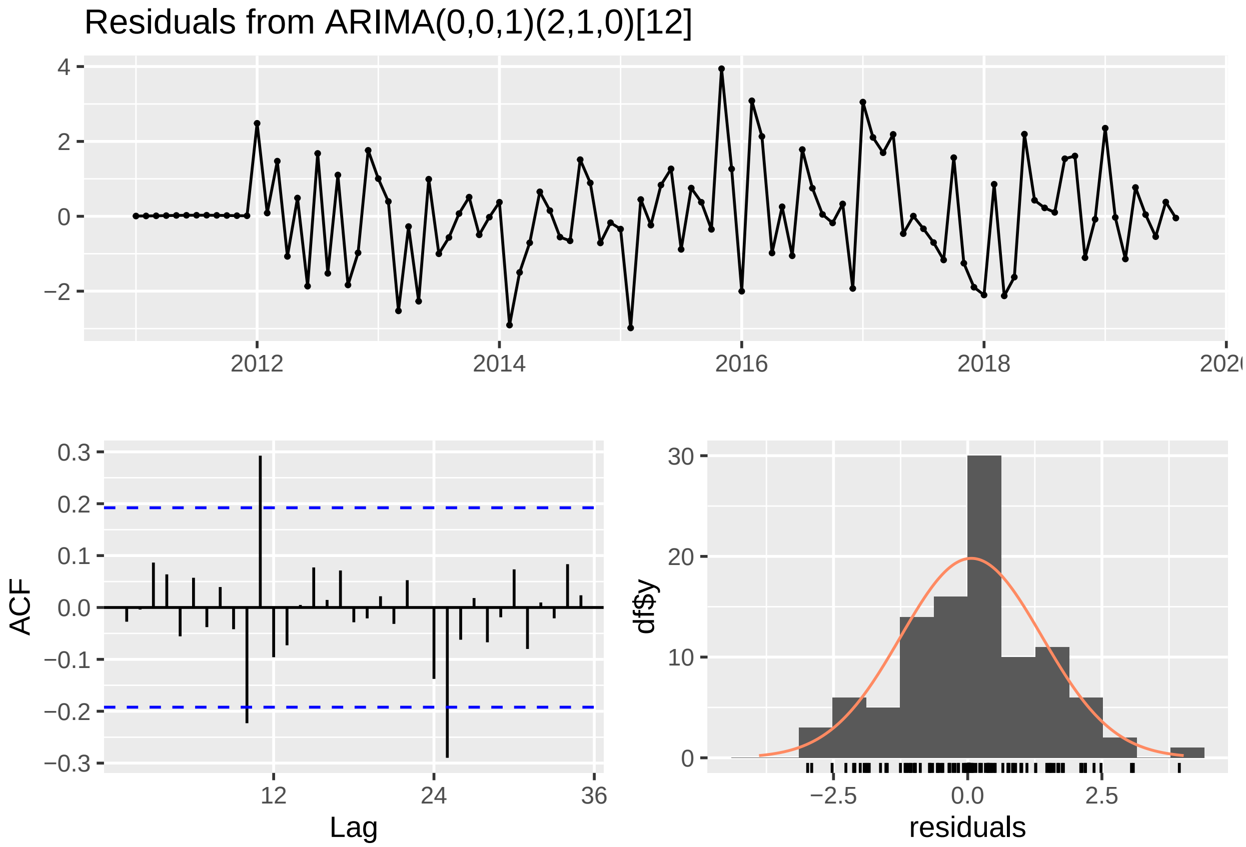
SARIMA Modeling

```
## Series: train_ts
## ARIMA(0,0,1)(2,1,0)[12]
##
## Coefficients:
##          ma1     sar1     sar2
##       0.6362  -0.6013  -0.2986
## s.e.  0.0947   0.1102   0.1104
##
## sigma^2 = 2.017:  log likelihood = -164.12
## AIC=336.24   AICc=336.7   BIC=346.33
##
## Training set error measures:
##                      ME     RMSE       MAE        MPE     MAPE      MASE
## Training set 0.06803095 1.313794 0.9753956 -0.2685469 6.627143 0.6805806
##                     ACF1
## Training set -0.02742599
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,1)(2,1,0)[12]
## Q* = 22.093, df = 18, p-value = 0.2279
##
## Model df: 3.   Total lags used: 21
```


SARIMA modeling
Water Temperature
30
25
20
15
10
2012
2014
2016
2018
2020
2022
Time
series
Original
SARIMA forecast



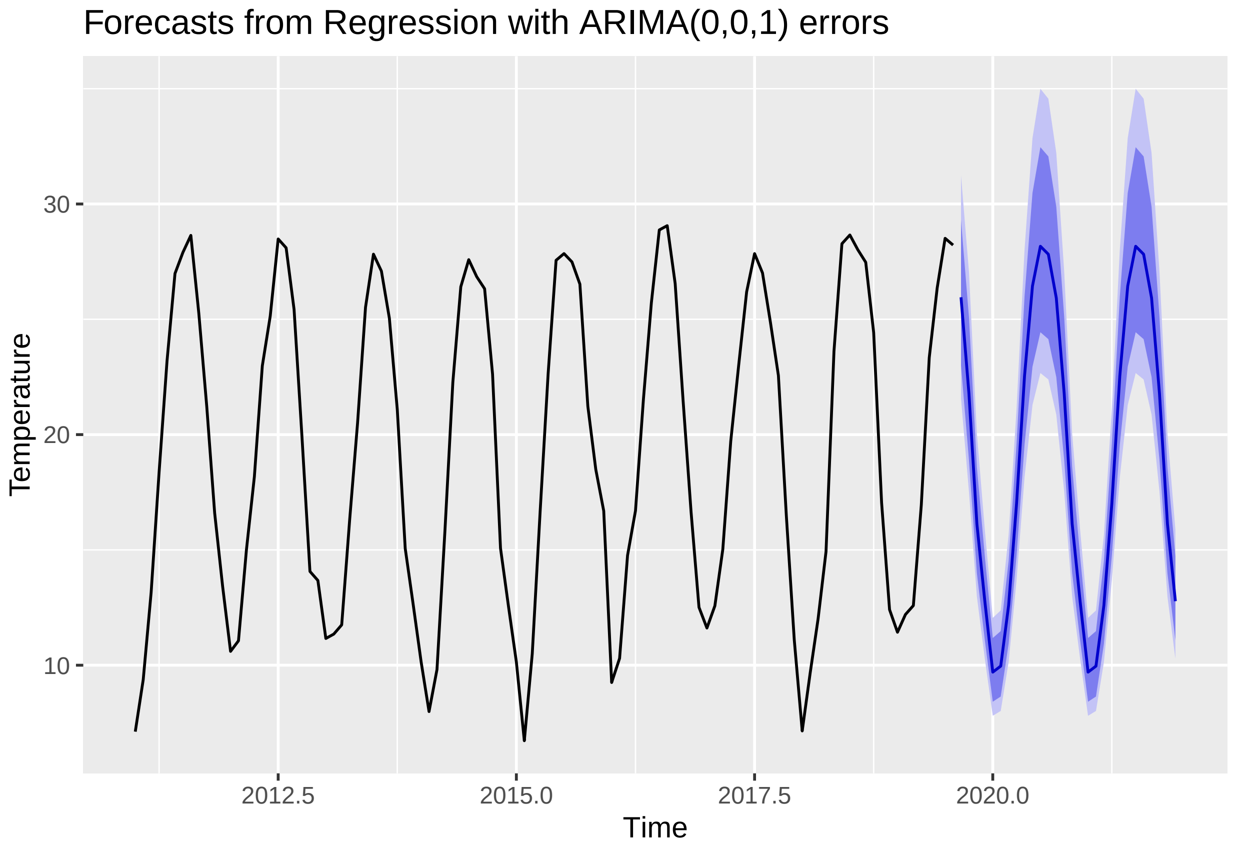
```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,1)(2,1,0)[12]
## Q* = 22.093, df = 18, p-value = 0.2279
##
## Model df: 3.   Total lags used: 21
```

```
##                      ME     RMSE       MAE        MPE     MAPE      MASE
## Training set 0.06803095 1.313794 0.9753956 -0.2685469 6.627143 0.6805806
##                     ACF1
## Training set -0.02742599
```

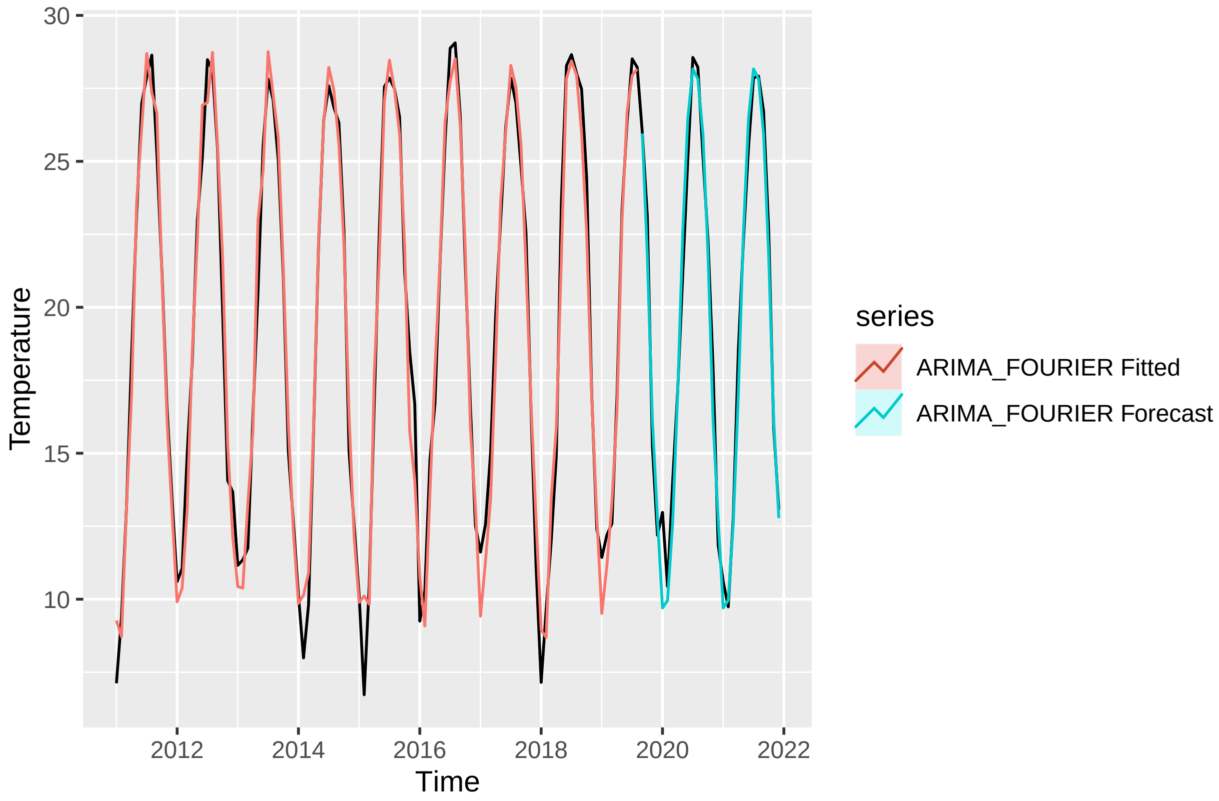
ARIMA+fourier Modeling

```
ARIMA_Four_fit <- auto.arima(train_ts,
                             seasonal=TRUE,
                             lambda=0,
                             xreg=fourier(train_ts,
                                          K=c(6))
                             )

ARIMA_Four_for <- forecast(ARIMA_Four_fit,
                           xreg=fourier(train_ts,
                                        K=c(6),
                                        h=h),
                           h=h
                           )
```

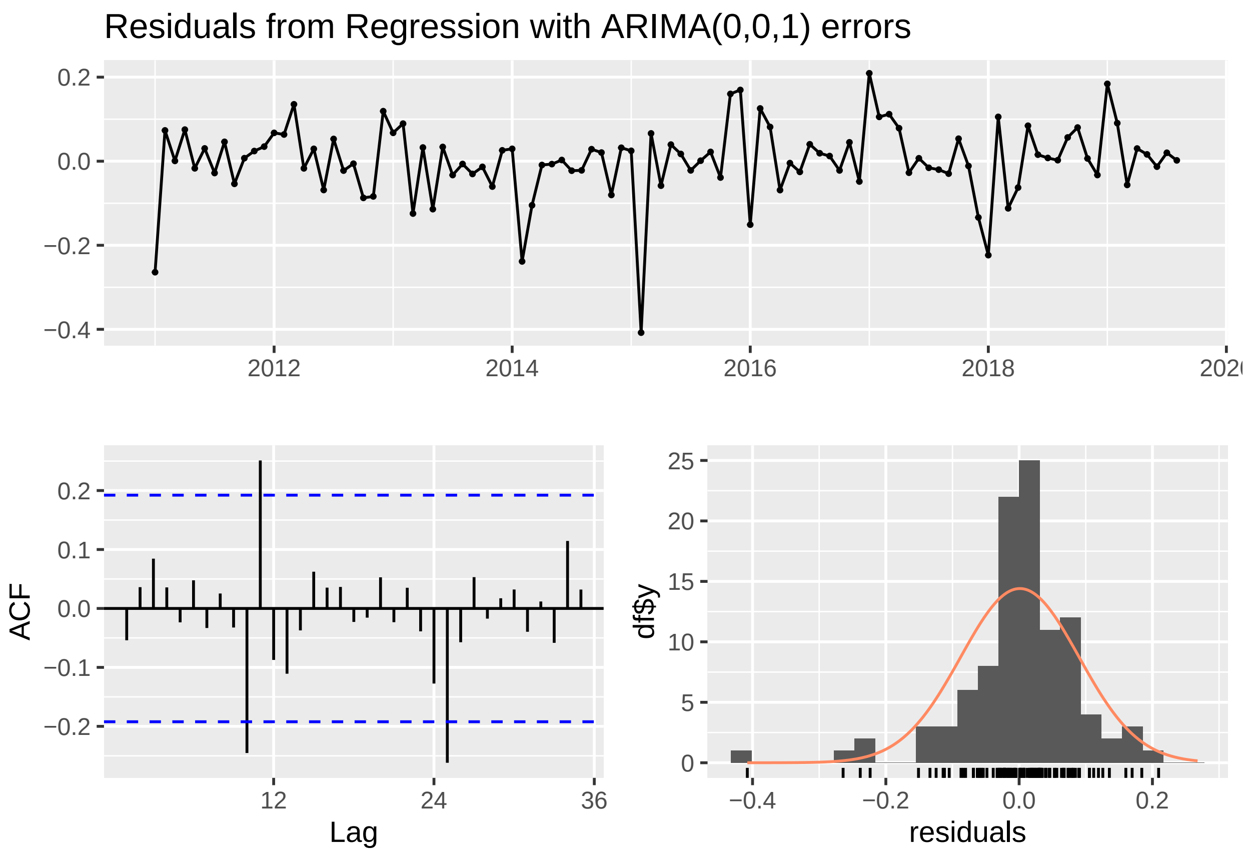
```
autoplot(ARIMA_Four_for) + ylab("Temperature")
```



```
autoplot(temperature_ts) +
  autolayer(ARIMA_Four_fit$fitted, series="ARIMA_FOURIER Fitted",PI=FALSE) +
  autolayer(ARIMA_Four_for, series="ARIMA_FOURIER Forecast",PI=FALSE) +
  ylab("Temperature")
```



```
checkresiduals(ARIMA_Four_for)
```



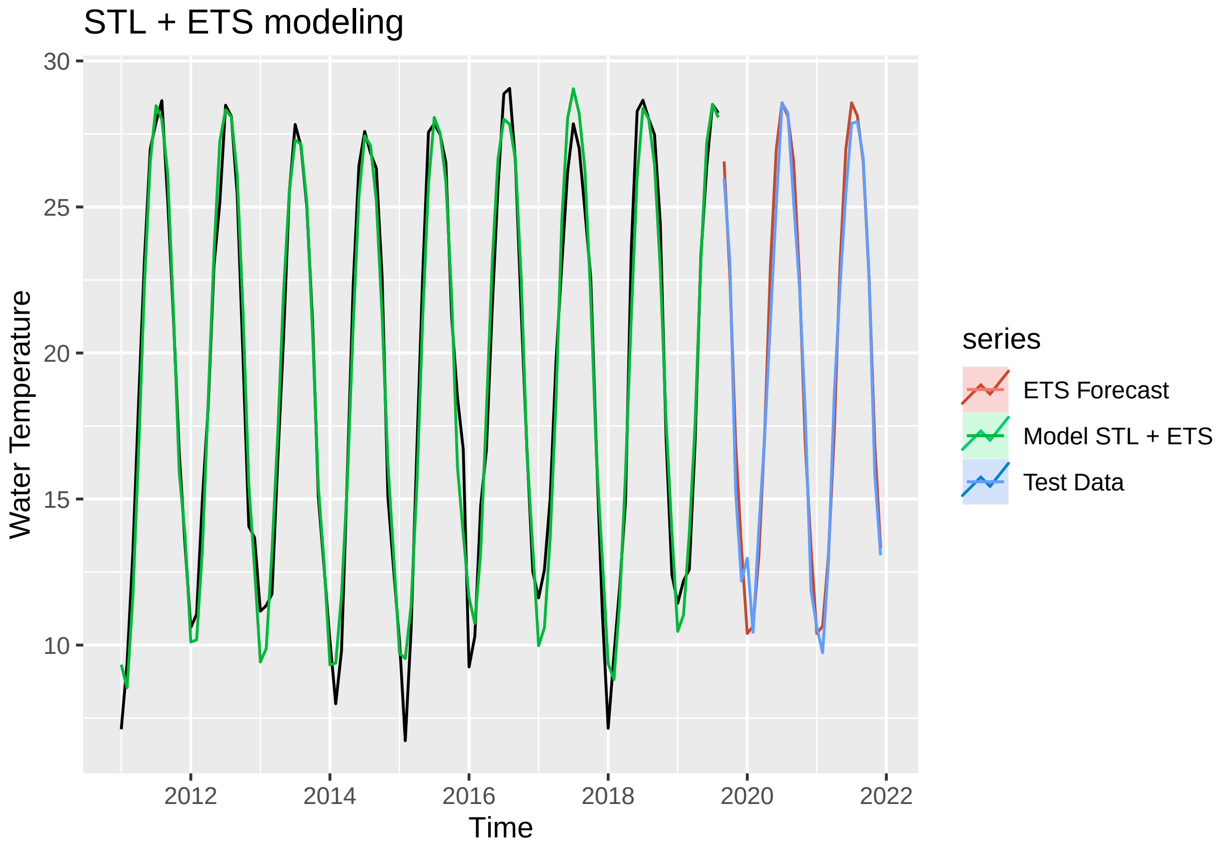
```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,0,1) errors
## Q* = 20.441, df = 20, p-value = 0.4306
##
## Model df: 1.   Total lags used: 21
```

```
forecast_accuracyARIMA_Four <- accuracy(ARIMA_Four_for)
print(forecast_accuracyARIMA_Four)
```

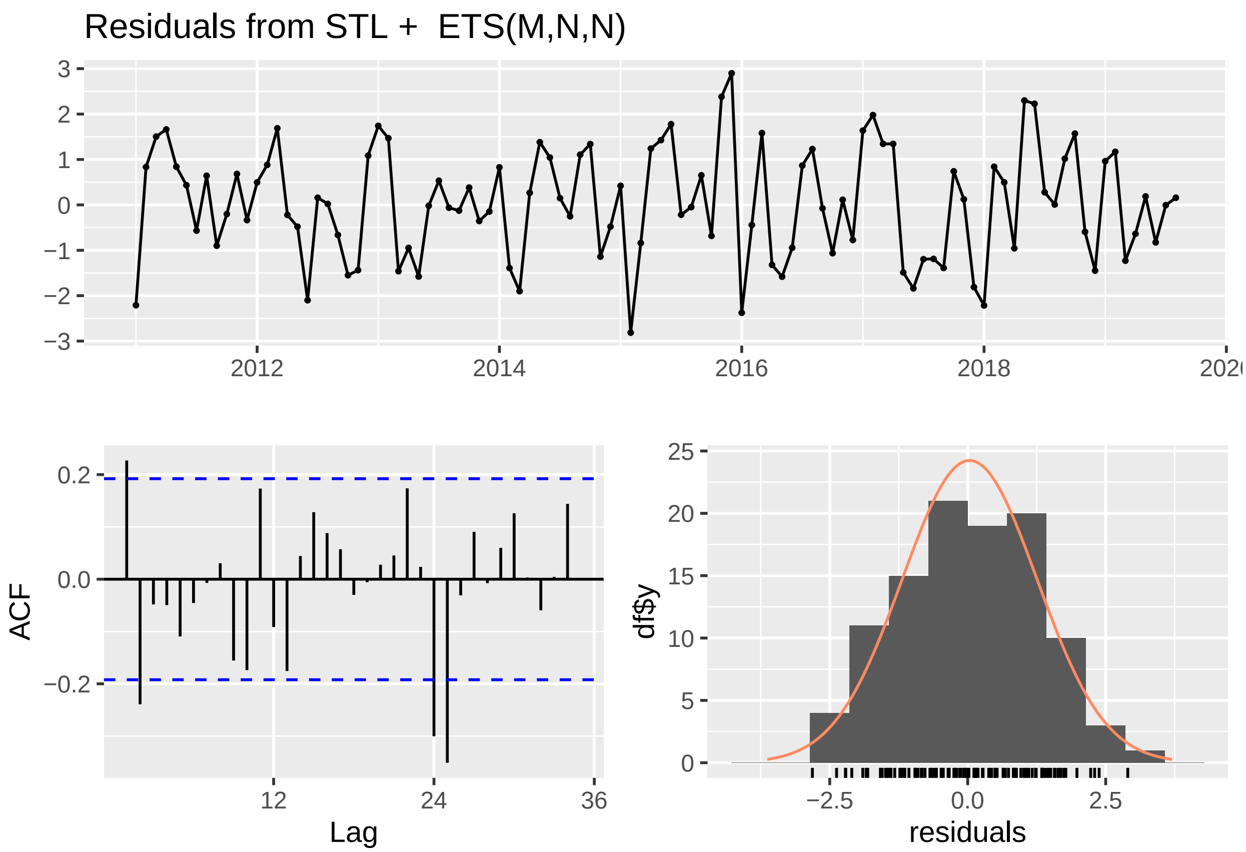
```
##                      ME     RMSE       MAE        MPE    MAPE      MASE
## Training set 0.05958683 1.128032 0.9064195 -0.2954703 6.18649 0.6324527
##                     ACF1
## Training set -0.08224948
```

STL + ETS Modeling

```
## Warning in ggplot2::geom_line(ggplot2::aes(x = .data[["timeVal"]], y = .data[["seriesVal"]], : Ignor
## Ignoring unknown parameters: `PI`
```



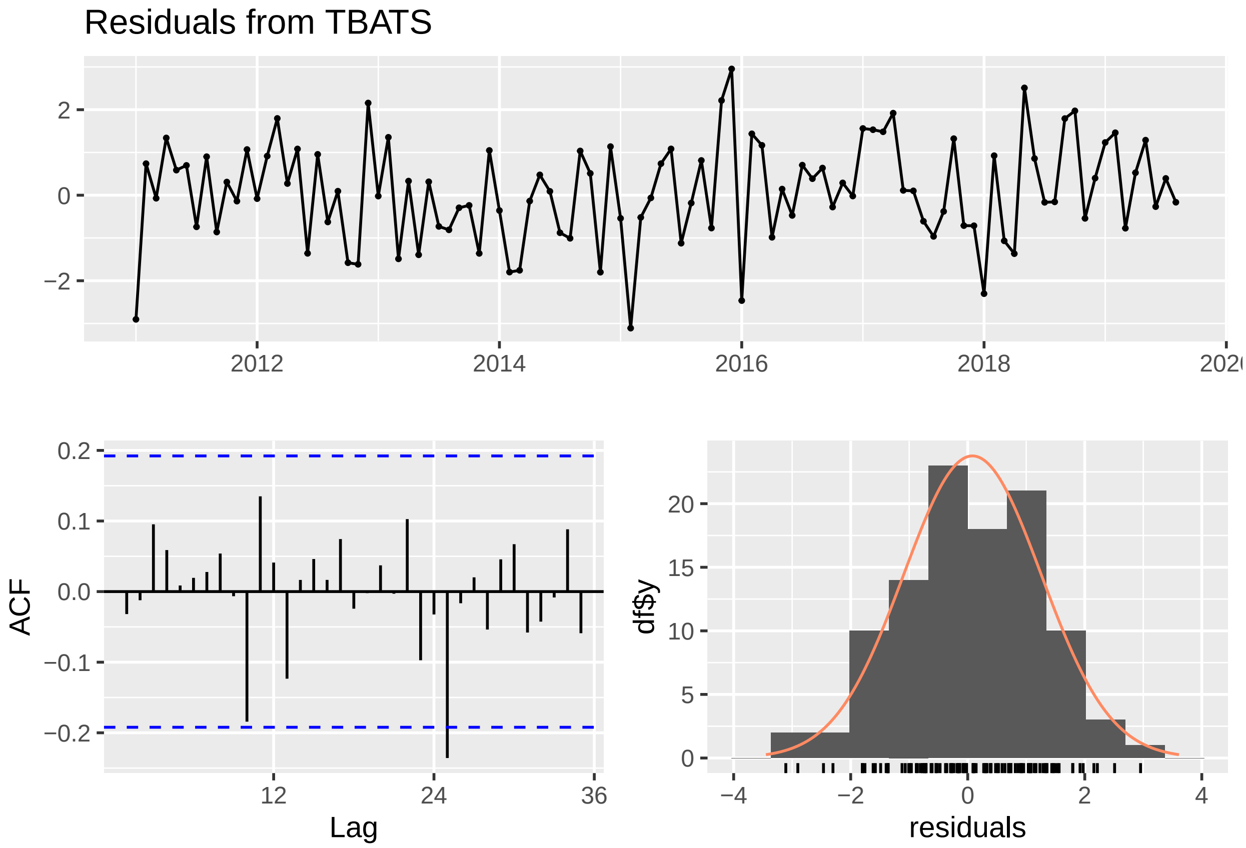
```
##                      ME     RMSE      MAE        MPE     MAPE      MASE
## Training set 0.03446794 1.217042 1.006346 -0.5474264 6.749242 0.7021762
##                   ACF1
## Training set 0.2270183
```



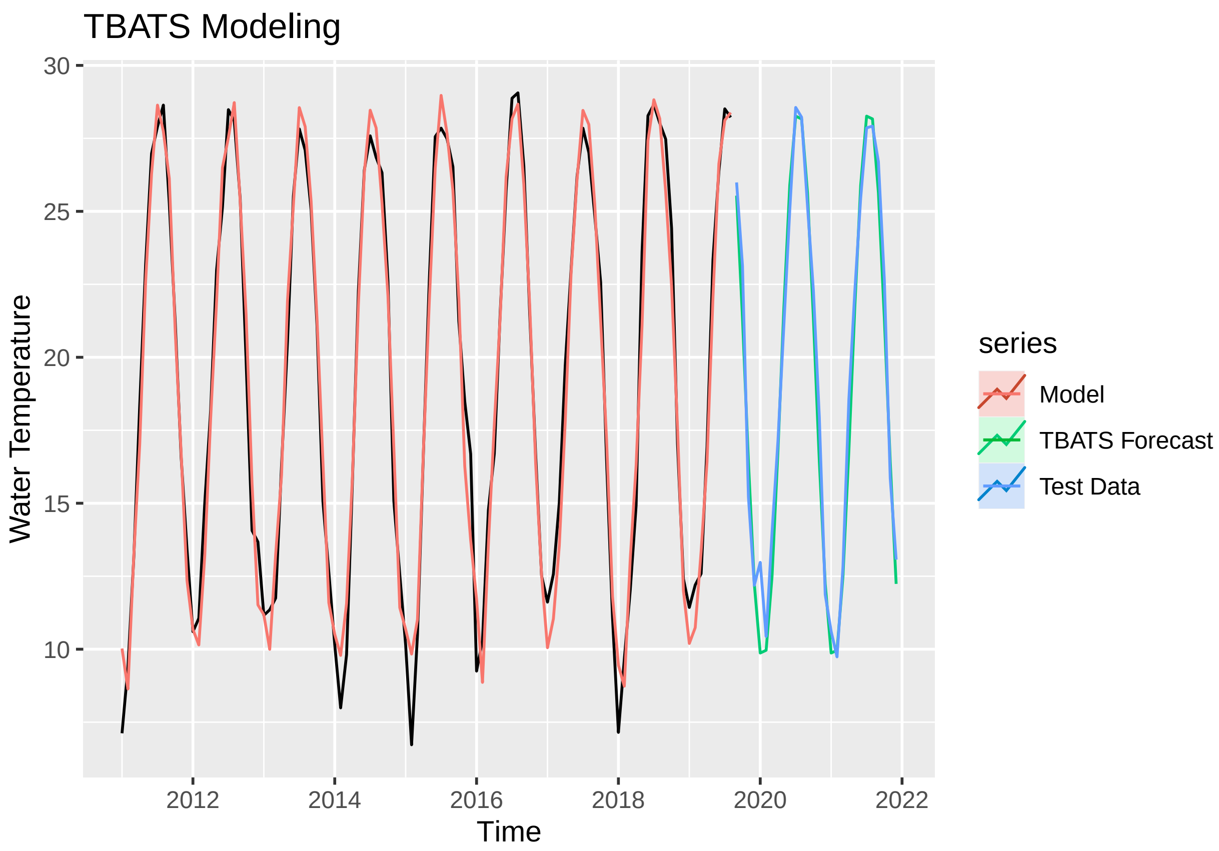
```
##
##  Ljung-Box test
##
## data:  Residuals from STL +  ETS(M,N,N)
## Q* = 32.701, df = 21, p-value = 0.04964
##
## Model df: 0.   Total lags used: 21
```

TBATS Modeling

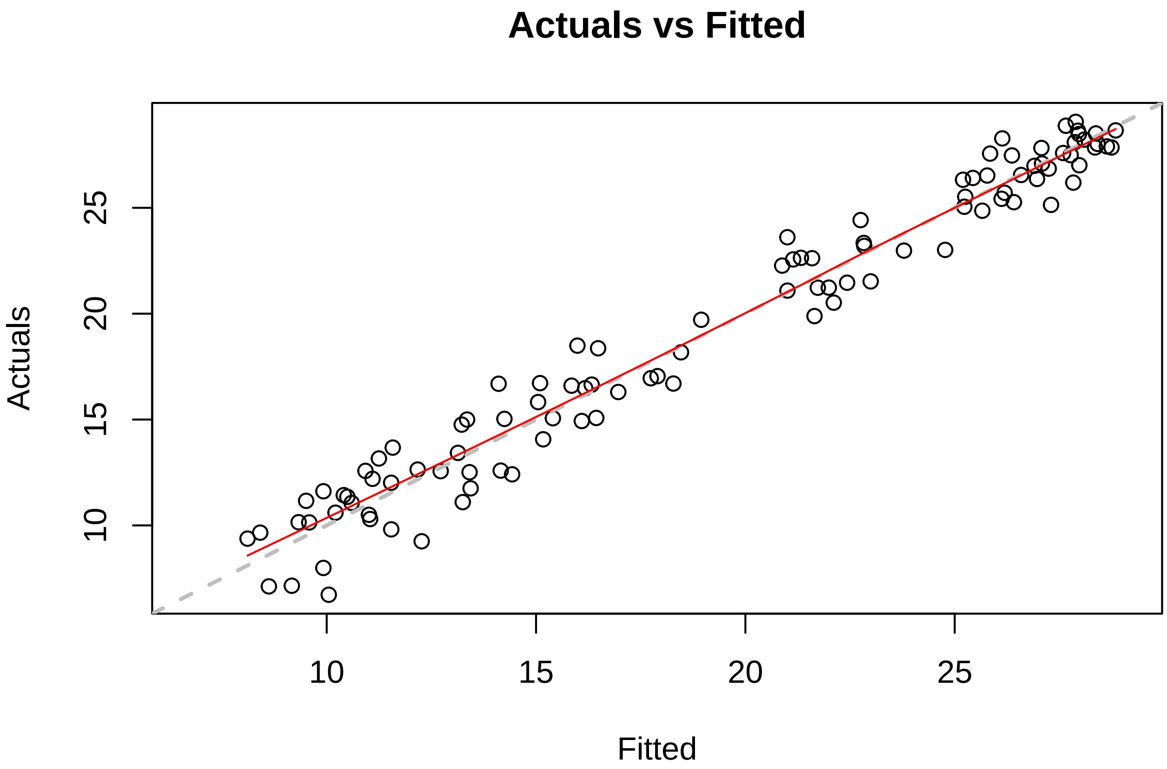
```
##                     Length Class  Mode
## lambda                0    -none- NULL
## alpha                 1    -none- numeric
## beta                  0    -none- NULL
## damping.parameter     0    -none- NULL
## gamma.one.values      1    -none- numeric
## gamma.two.values      1    -none- numeric
## ar.coefficients       0    -none- NULL
## ma.coefficients       1    -none- numeric
## likelihood            1    -none- numeric
## optim.return.code     1    -none- numeric
## variance              1    -none- numeric
## AIC                   1    -none- numeric
## parameters            2    -none- list
## seed.states           4    -none- numeric
## fitted.values       104    ts     numeric
## errors              104    ts     numeric
## x                    416     -none- numeric
## seasonal.periods       1     -none- numeric
## k.vector               1     -none- numeric
## y                    104     ts     numeric
## p                      1     -none- numeric
## q                      1     -none- numeric
## call                   2     -none- call
## series                 1     -none- character
## method                 1     -none- character
```



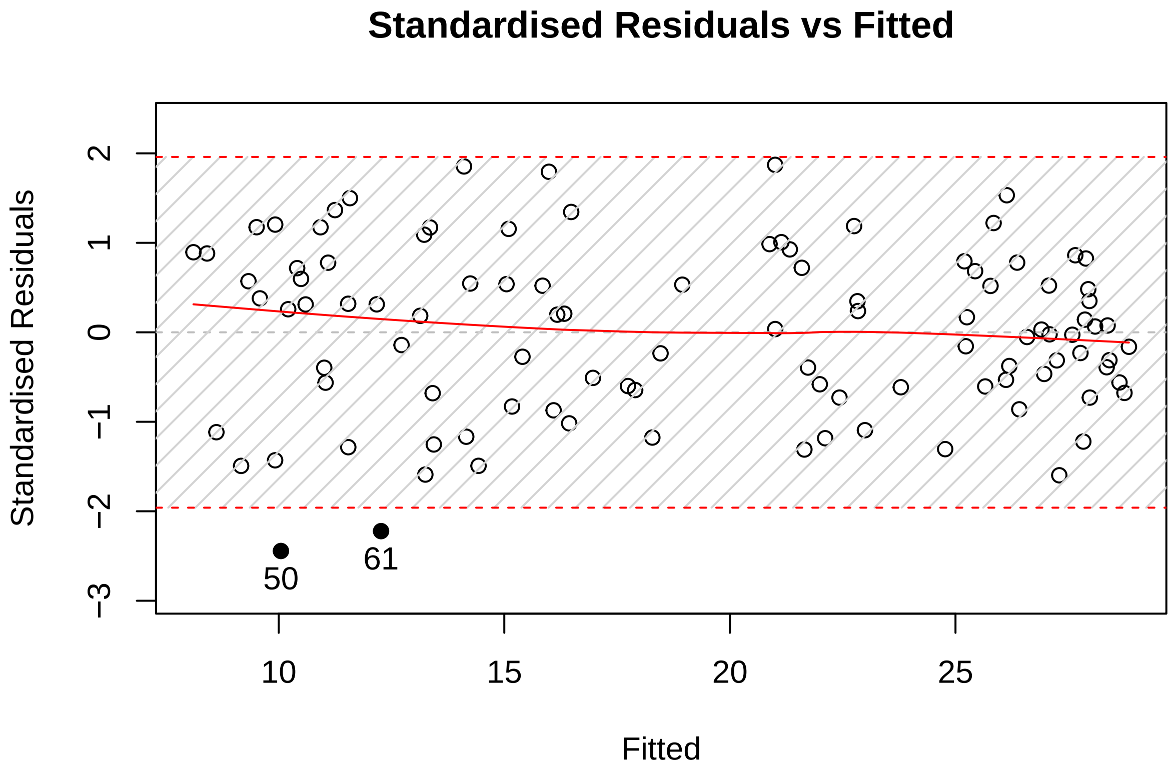
```
##
##  Ljung-Box test
##
## data:  Residuals from TBATS
## Q* = 11.45, df = 21, p-value = 0.9533
##
## Model df: 0.   Total lags used: 21

## Warning in ggplot2::geom_line(ggplot2::aes(x = .data[["timeVal"]], y = .data[["seriesVal"]], : Ignori
## Ignoring unknown parameters: `PI`
```

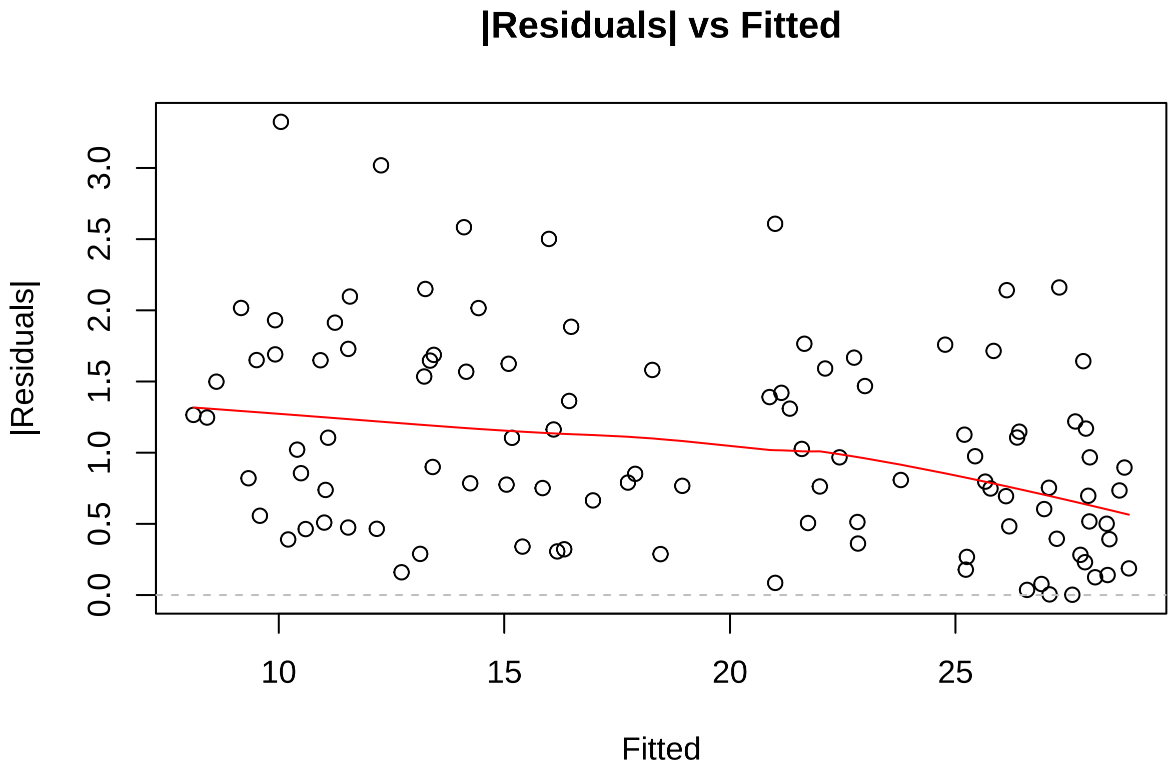


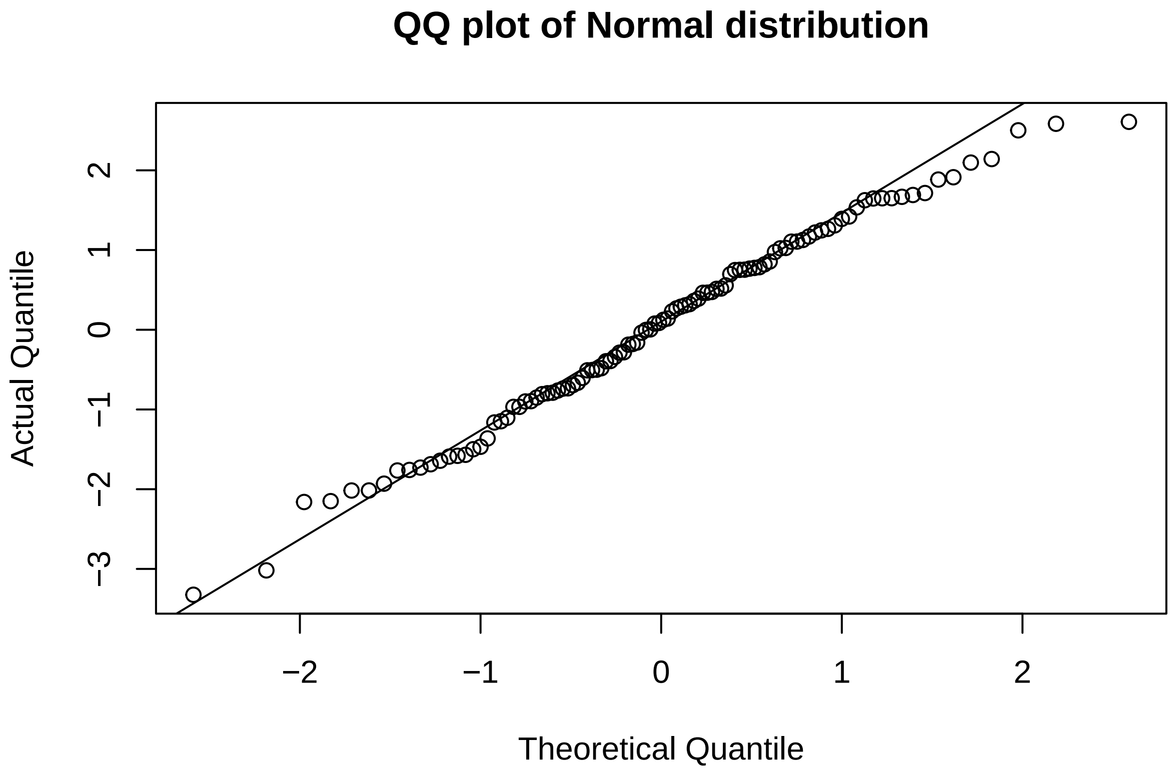
```
##                      ME     RMSE       MAE        MPE     MAPE      MASE
## Training set 0.0810154 1.173607 0.9415607 -0.3983483 6.478958 0.6569724
##                     ACF1
## Training set -0.03184701
```

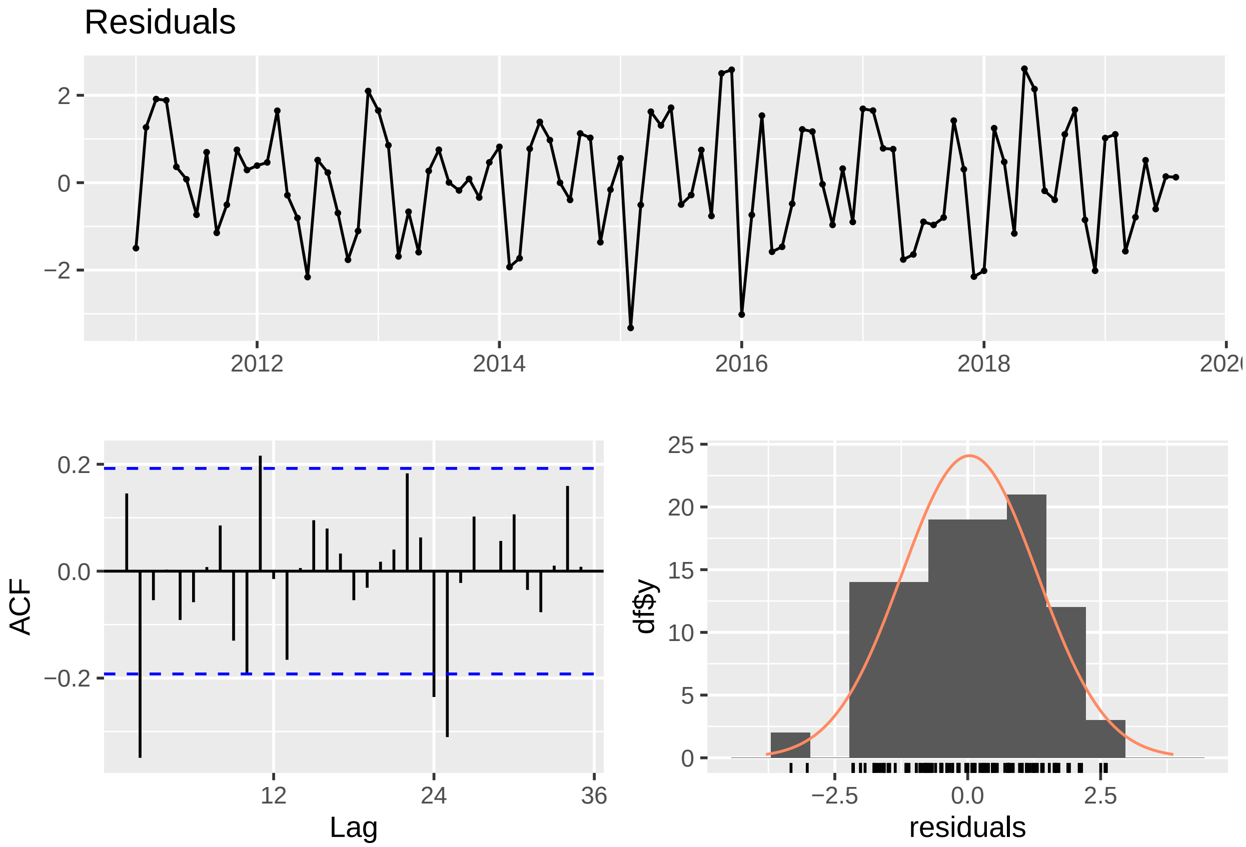


SSES


Standardised Residuals vs Fitted
Standardised Residuals
2
1
0
−1
−2
−3
50
61
10
15
20
25
Fitted

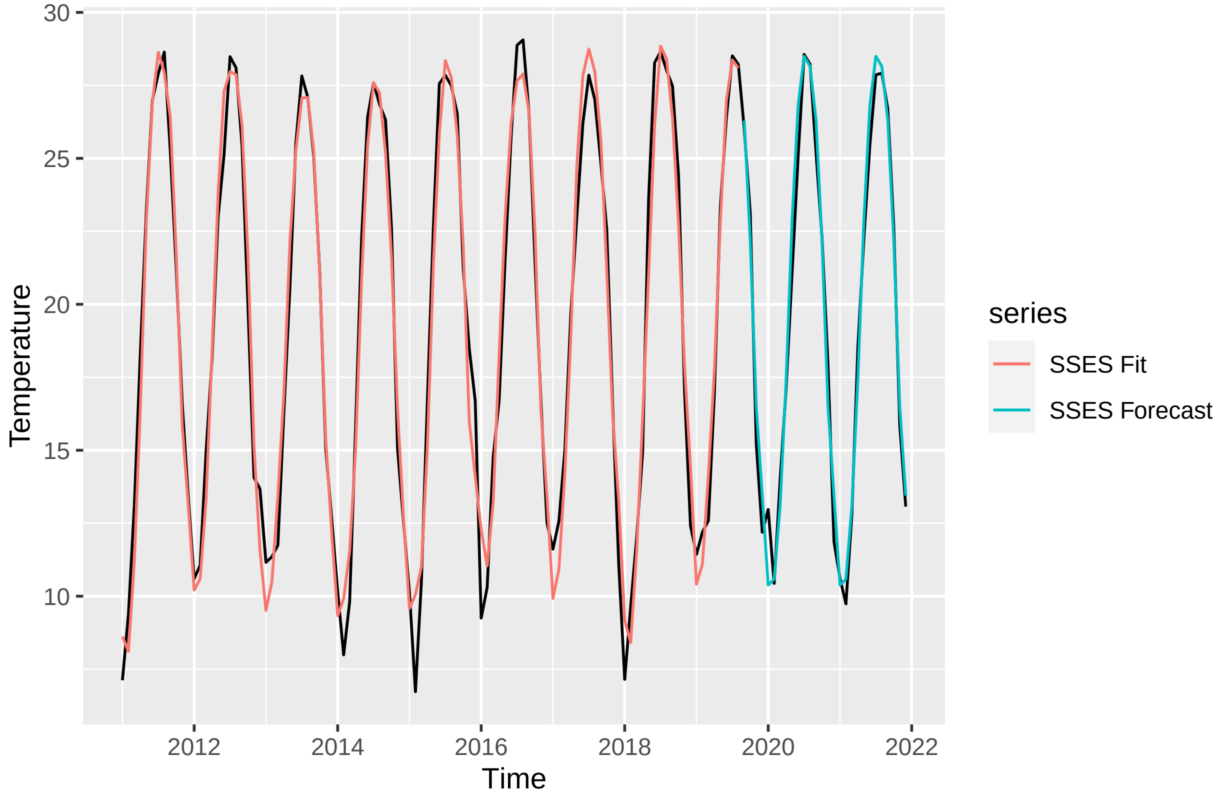

|Residuals| vs Fitted
|Residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
10
15
20
25
Fitted

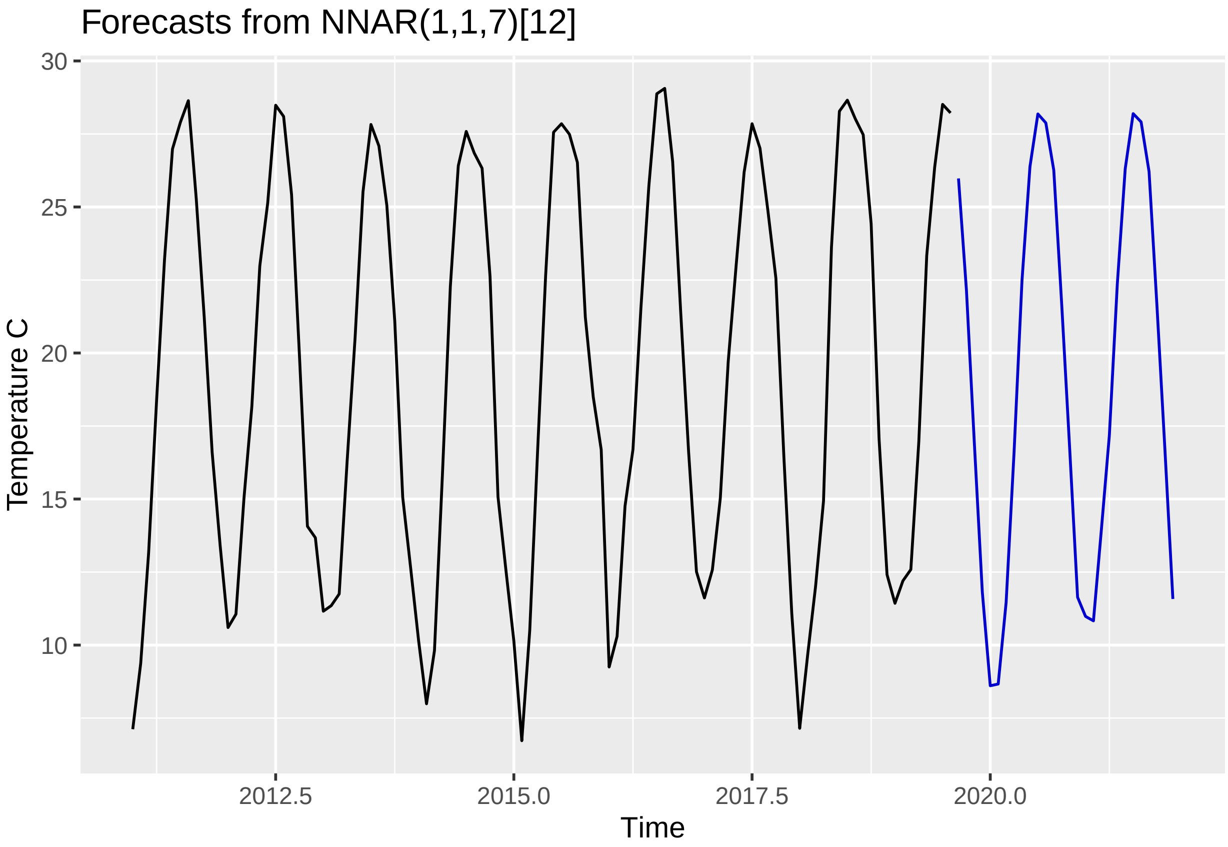

QQ plot of Normal distribution
Actual Quantile
Theoretical Quantile
2
1
0
−1
−2
−3
−2
−1
0
1
2



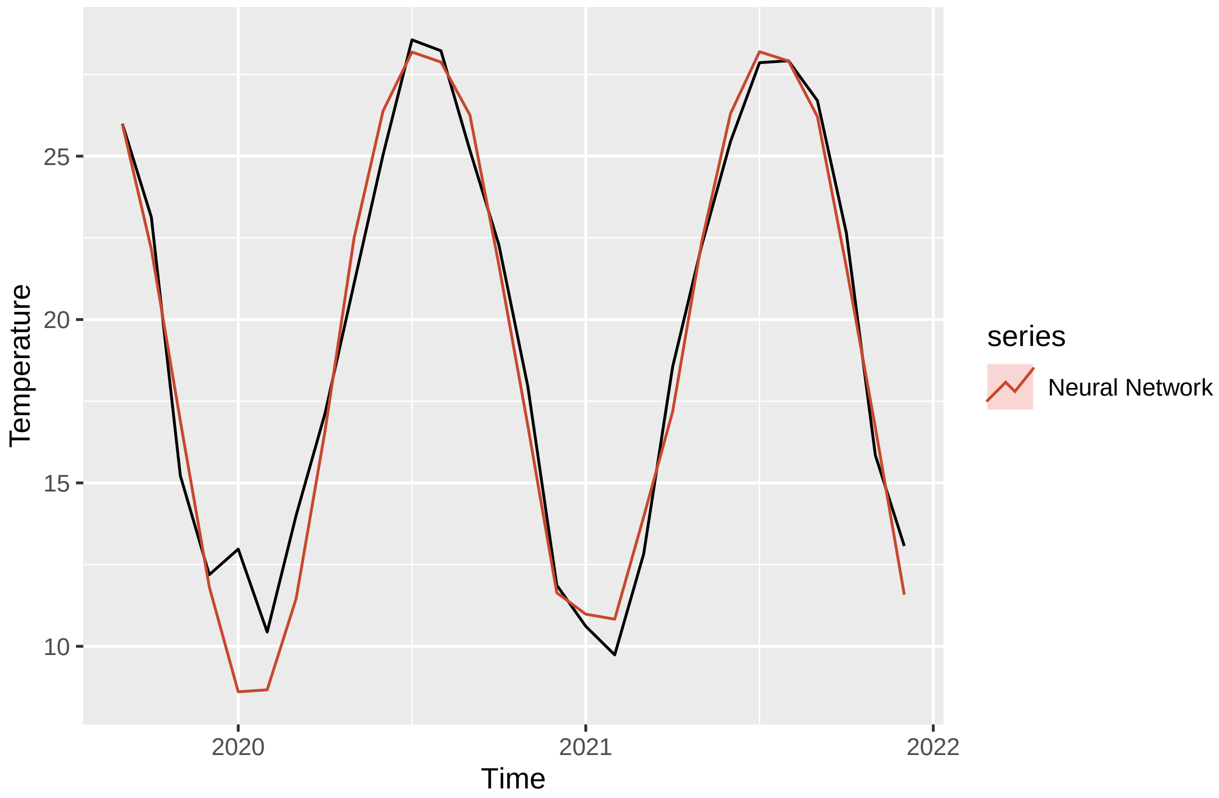
```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 35.778, df = 21, p-value = 0.02315
##
## Model df: 0.   Total lags used: 21
```

```
## Warning in ggplot2::geom_line(ggplot2::aes(x = .data[["timeVal"]], y = .data[["seriesVal"]], : Ignor
## Ignoring unknown parameters: `PI`
```


30
25
20
15
10
Temperature
2012
2014
2016
2018
2020
2022
Time
series
SSES Fit
SSES Forecast

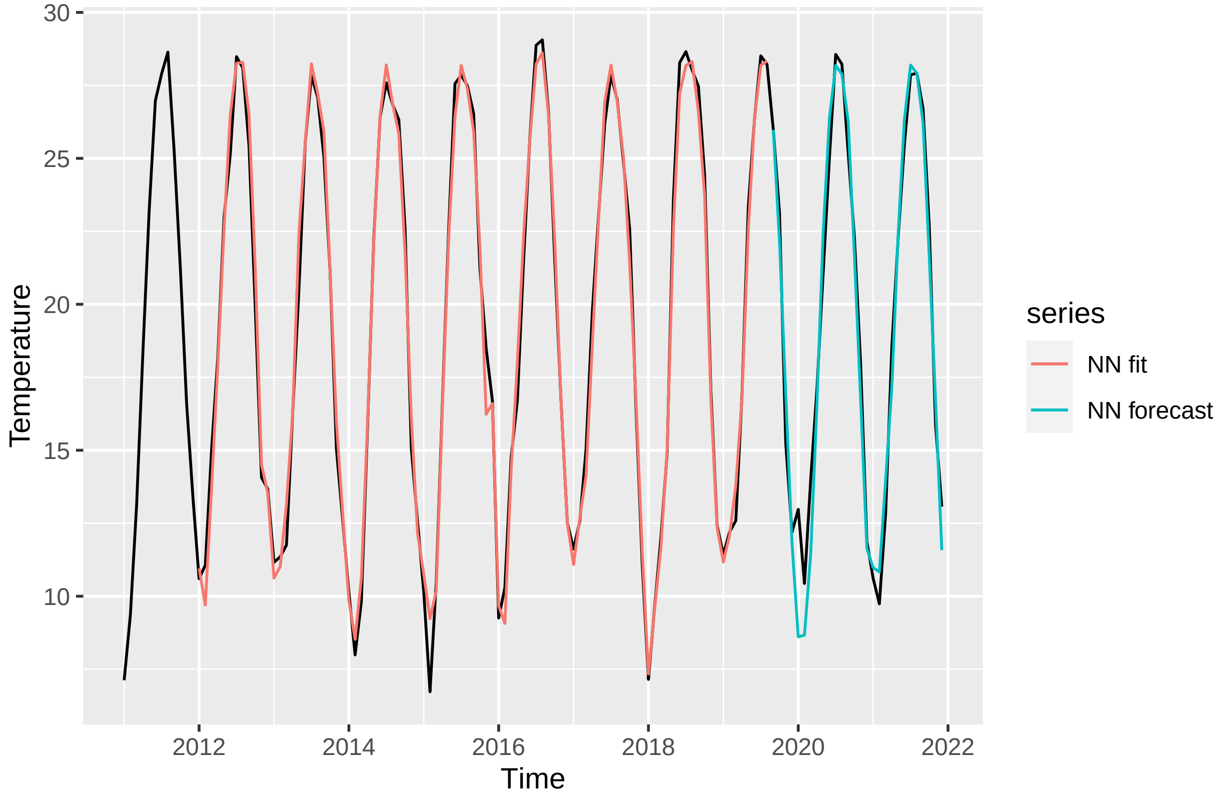


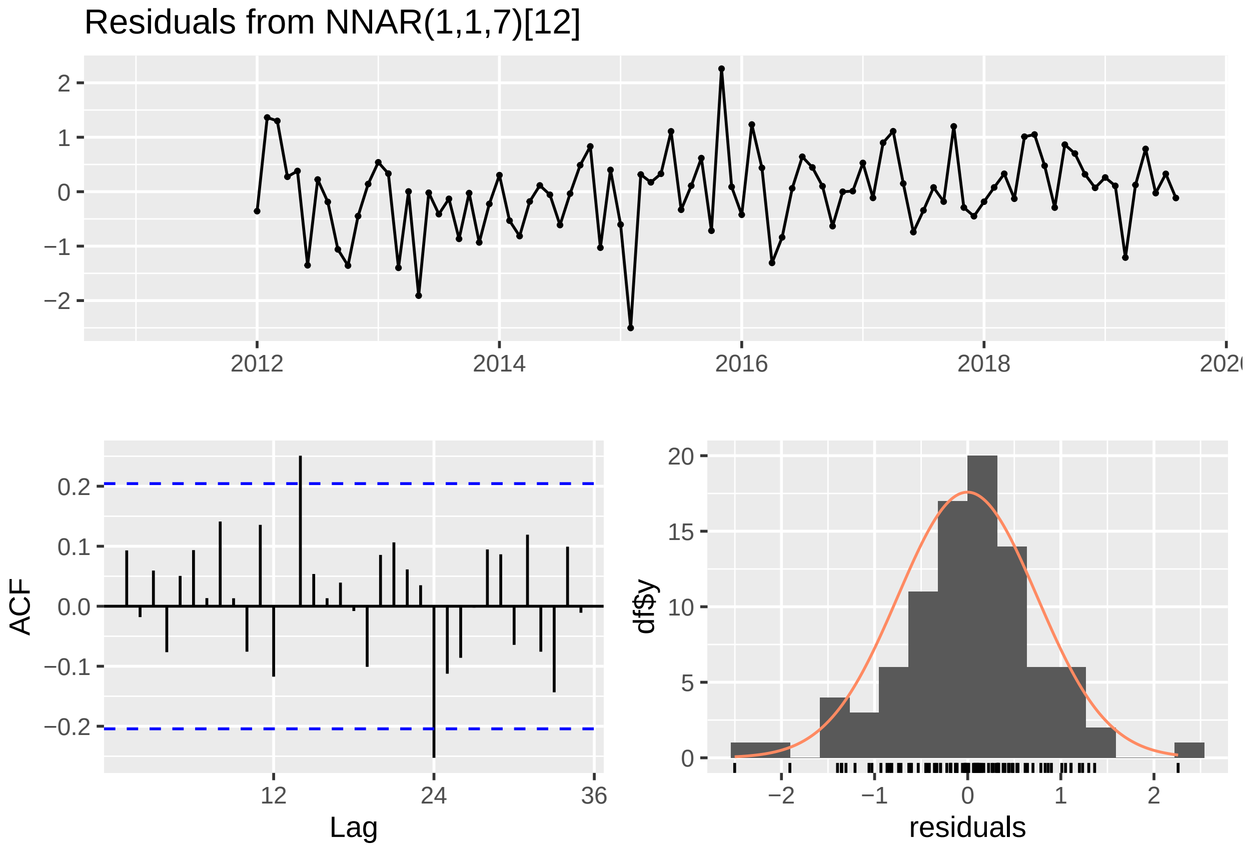
Neural Network



```
## Warning in ggplot2::geom_line(ggplot2::aes(x = .data[["timeVal"]], y = .data[["seriesVal"]], : Ignori
## Ignoring unknown parameters: `PI`
```

```
## Warning: Removed 12 rows containing missing values (`geom_line()`).
```


30
25
20
15
10
Temperature
2012
2014
2016
2018
2020
2022
Time
series
NN fit
NN forecast



```
##
##  Ljung-Box test
##
## data:  Residuals from NNAR(1,1,7)[12]
## Q* = 20.028, df = 21, p-value = 0.5195
##
## Model df: 0.   Total lags used: 21
```

Create Scores

Create Score table

```
## The best model by RMSE is: NN
```

Limitations Though we have produced accurate models and forecasts, there are a few limitations to be aware of. The first is regarding our prediction time frame We only have access to ∼ 10 years of data which likely does not demonstrate enough of a trend to be recognized as the role of climate change in our models. We should expect increasing temepratures in the future as a result of climate change. Even if our model did capture this linear trend, the modeling would be based on the assumtions that the trend is constant (aka that emissions causing climate change are to remain exactly as they are now, neither increasing nor decreasing). This limits further the time frame for which projections and models are viable. Finally, the resolution of our data is not ideal; we have monthly aggregate data so we are projecting out monthly as well. This is not as helpful as a finer resolution may have been given the fact that monthy aggregate data does not accurately reflect the extreme temeprature spikes we may see on specific days. In the future, it would be usely to repeat this modeling ideally with daily data. Even geting multiple temepratures a day could had more complexity as we should except to see patterns of temeprature changes throughout the day.

Selected Refrences Johnson, Z. I., Wheeler, B. J., Blinebry, S. K., Carlson, C. M., Ward, C., & Hunt, D. E. (2013a). Dramatic variability of the carbonate system at a temperate coastal ocean site (Beaufort, North

Carolina, USA) is regulated by physical and biogeochemical processes on multiple timescales. PloS One, 8(12), e85117. https://doi.org/10.1371/journal.pone.0085117

McIntyre, A. D. (2010). Life in the world's oceans. In Wiley eBooks. https://doi.org/10.1002/9781444325508

Wang, Z., Tsementzi, D., Williams, T. C., Juarez, D. L., Blinebry, S. K., Garcia, N. S., Sienkiewicz, B. K., Konstantinidis, K. T., Johnson, Z. I., & Hunt, D. E. (2020a). Environmental stability impacts the differential sensitivity of marine microbiomes to increases in temperature and acidity. The ISME Journal, 15(1), 19–28. https://doi.org/10.1038/s41396-020-00748-2