PERFORMING TIME SERIES FORECAST ON STREAMFLOW DATA

ABSTRACT

Effectively producing and analyzing monthly streamflow variability is essential for water resource management and development, which includes amongst others hydropower generation, flood defense and controlling, drought alleviation, river basin planning and management. This study, performs time series forecasting on streamflow data using an autoregressive integrated moving average (ARIMA) model to fit streamflow data as well as forecasting models in R. These different methods are compared to see which model fits the respective data better.

Klaus Schroder

Table of Contents

Introduction

Accurate prediction of streamflow demonstrates a crucial role for effective reservoir system operations. More specifically, streamflow forecasting provides valuable information for reservoir operators to make critical decisions on water release amount to maximize reservoir storage benefits considering tradeoffs among flood control, municipal water supply, irrigation, hydropower, etc. However, this task has presented difficult challenges due to the complex apparatus of the physical-based processes, in addition to the impact of uncontrollable factors. Hence, performing time series forecasting on streamflow data in tandem with the supervision of proficient hydrologists for validation purposes is to ensure precise forecasting of discharge flows that could be of primary importance. To this end, an autoregressive integrated moving average (ARIMA) model is applied to fit streamflow data and implemented to perform time-series forecasts, for an accurate prediction of streamflow is presented and evaluated - without losing any generality - for two United States Geological Survey (USGS) gauge stations.

To achieve this, this report will briefly describe the data acquisition processes in addition to the study area. Subsequently, the aims and objectives are listed and the methodology to be performed is outlined and discussed. Thereafter, the results will be presented followed by a concise conclusion.

Data Acquisition

The data utilized for the analysis is monthly surface-water flow discharge data in cubic feet per second for the selected sites. This data was collected from the USGS's website and has a temporal coverage from 1990 through to the most recent available data. Moreover, the information for each station was acquired from https://waterdata.usgs.gov/nwis/nwismap/?site_no=08017200&agency_cd=USGS and https://waterdata.usgs.gov/nwis/inventory/?site_no=08116650&agency_cd=USGS. The USGS station number 08116650 (Brazo's River) had data up until the end of February 2020, and the USGS station number 08017200 (Cowleech Fork Sabine River) had data up until the end of September 2020. It is considered here that the USGS's website provided the option to select the data period of the relevant data so it was not necessary to subset the data in R program, as it is done on the website itself.

Study Area

This study is conducted on two USGS streamflow stations situated in the state of Texas, US. The first station, USGS station no. 08116650, is located along the Brazos River near Rosharon, TX. It has a contributing drainage area of 35,773 square miles. The second station, USGS station no. 08017200, is located along the Cowleech Fork Sabine River at Greenville, TX. This station has a contributing drainage area of 81.0 square miles. Their absolute locations are Latitude 29°20'58", Longitude 95°34'56" NAD27 and Latitude 33°07'58", Longitude 96°04'36" NAD27 respectively and is presented in Figure 1 below.

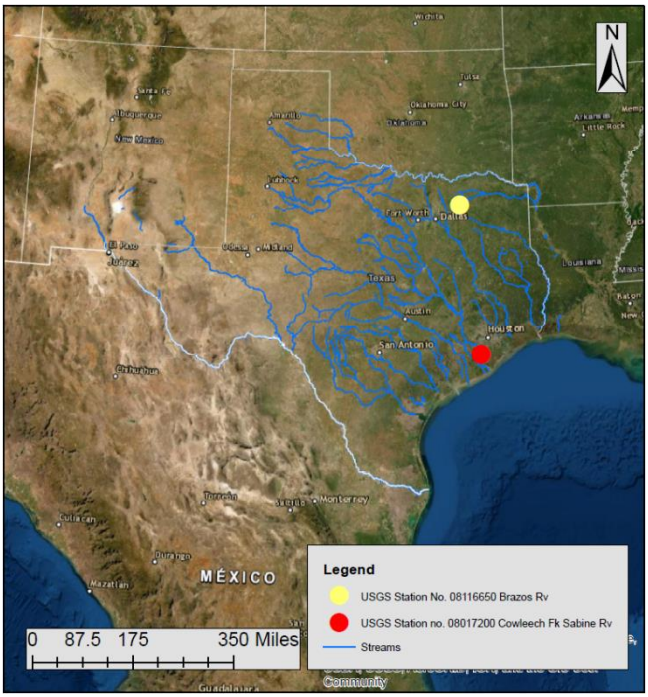


Figure 1: Study area showing the location of the two stations used for this study

Aims, Objectives, and Methodology

This section describes the methodology that was followed for this report to achieve the aim and objectives. This study aims to perform time-series forecasts on streamflow data watershed both for short- and long-term purposes based on past hydrological observations. Towards fulfilling the aim of these projects, the objective of this report (as also outlined by the project guidelines listed on CANVAS) is to:

- Subset the time series for both streams for 1990 through the present
- To compute the skew of each

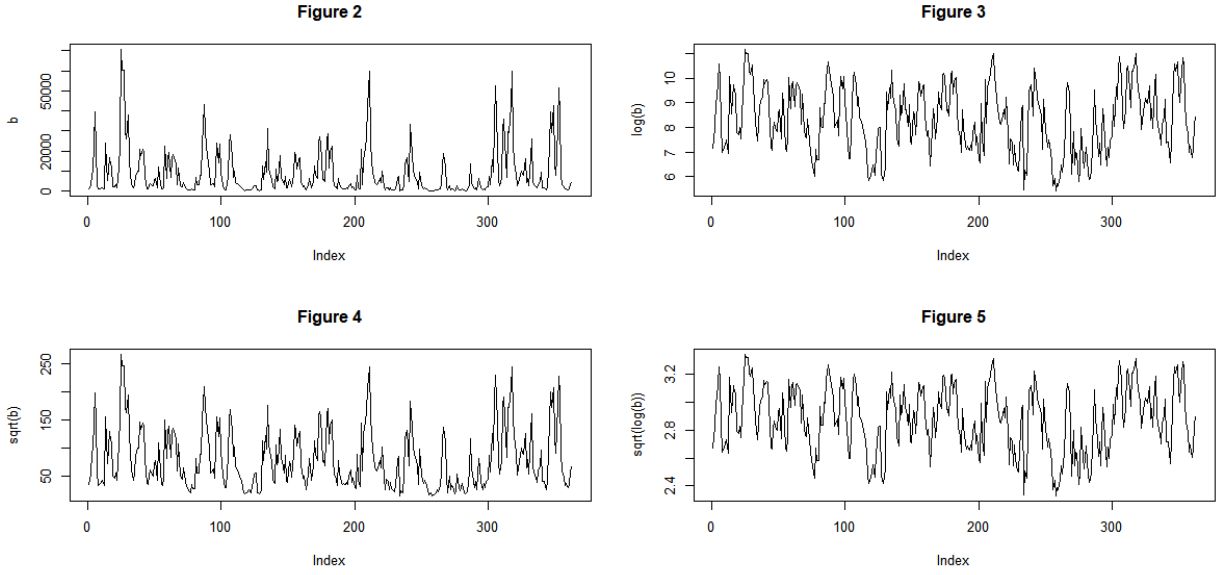
- Experiment with detrending and applying back-difference to the log-transformed series if needed or relevant to the data set.
- Fit the series using an ARIMA model.
- and present the parameters and goodness of fit.

Naturally, the methodology follows that of the objectives outline above. When considering hydrological time series, monthly streamflow series typically displays periodic behaviors in the mean and variance, and in general, periodic autoregressive models are used in de analysis of the data (see, for example, Modal & Wasimi, 2006). In such instances, studies have shown that such series usually have normal or log-normal distributions (see, for example, Tesfaye et al., 2006; Wang et al., 2009). To analyze the data this study firstly analyzed the skewness of the data, it transforms the data and assumes a normal distribution for the transformed data. The data was transformed using a logarithmic transformation. The Auto Correlation Function (ACF) Plots of the original data and that of the transformed data were then plotted to confirm the transformation's' functionality. In this study data sets consisting of the time series of monthly averages of natural streamflows, measured in the year ranging from 1990 to 2020, was considered. The decomposition of the data sets was then observed towards investigating the seasonality of the data. Due to the streamflow series having periodic behavior in the mean and variance. Therefore, as in the case of Modal & Wasimi, 2006, a general periodic autoregressive model was assumed in the analysis of the time-series data. The Auto.arima function in R was utilized towards this end, the Holtwinters model in R was also utilized as a comparison to the ARIMA. A summary of the parameters of the logarithmic data and its specific ARIMA function was then presented. Subsequently, the goodness of the fit for the model was checked by forecasting for the years collected from 1990 through 2015. Finally, the stationarity was tested before the ARIMA model was fitted to derive the optimal ARIMA model. This process was repeated and completed for both datasets. The results of this methodological process are presented in the next section.

Results – Brazos River

There are a few common ways of transforming data, including the logarithm of the data, the square root of the data, and even more complicated transformations such as the square root of the logarithm of the data.

For the data collected from the Brazos River monthly discharge, the transformations would have the following effects as seen in Figures 2 to 5.



Figures 2 -5: Monthly Discharge of the Brazos River Station Transformed

Although the square root of the logarithm of the data, has the least variance, it is sufficient to work with the simpler transformation, the logarithm of the data. These transformations are needed when the skewness of the original data is more than 1 or less than -1.

The skewness of the Brazos data is 2.24, which is sufficiently large to use a transformation to interpret results and predictions from data. The skewness of the logarithm transformation is -0.047, which is remarkably close to zero, making this transformation ideal to work with. The density and the skewness of the of the original data vs the transformed data are shown in figures 6 and 7. It is important to note that the logarithmic transformation follows a relatively more symmetric and normal distribution than the original data.

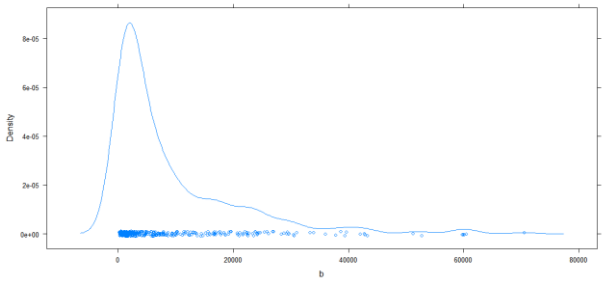


Figure 6: Density of the original data

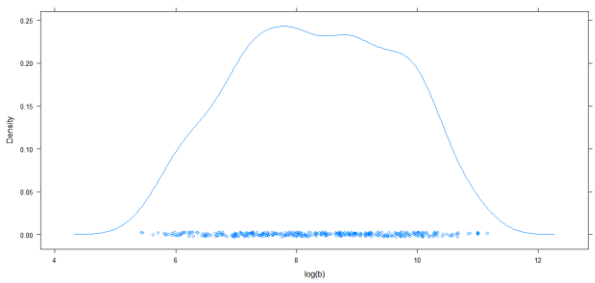


Figure 7: Density of the logarithmic transformation of the data

The following data and conclusions will be made following the assumption made by the previous results that the logarithmic transformation of the data is the better data set to work with, with a skewness close to zero. The plotting of the time series is shown in Figure 8.

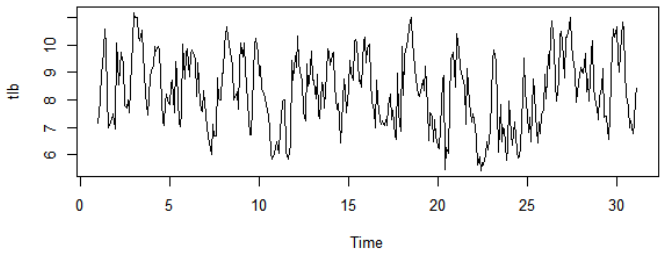


Figure 8: Time series of the logarithmic transformation of the data

The data follows a multiplicative trend and not an additive trend, so a decomposition of the multiplicative model will give a visual representation of the trend and seasonality as seen in figures 9 and 10.

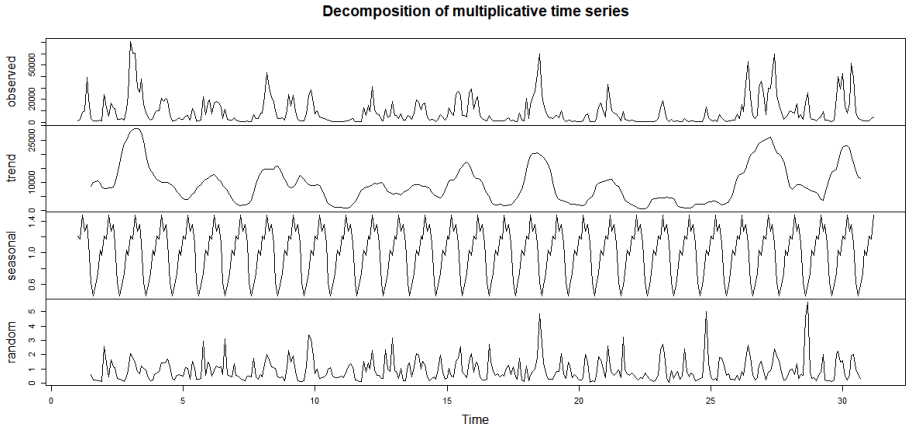


Figure 9: Decomposition of the time series of original data

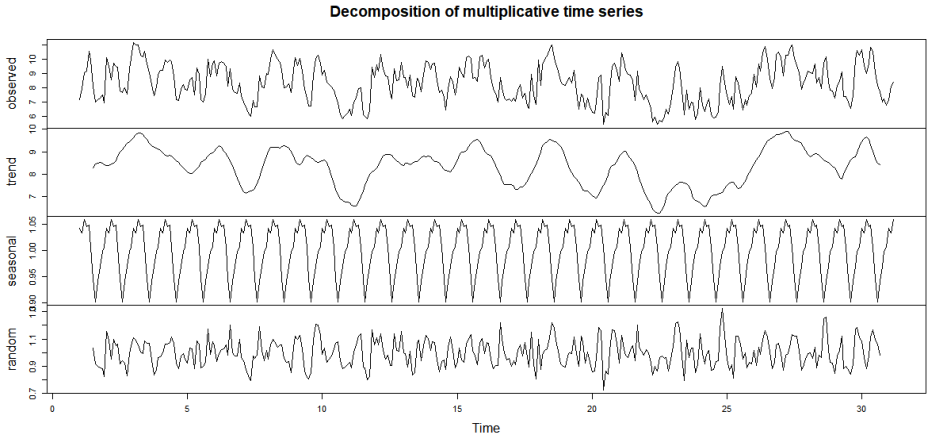


Figure 10: Decomposition of the time series of transformed data

```
> #Question 2 skewness
> par(mfrow=c(2,2))
> b=Assignment_data$Brazon[1:363]
> skewness(b)
[1] 2.240486
attr(,"method")
[1] "moment"
> skewness(log(b))
[1] -0.04719579
attr(,"method")
[1] "moment"
> #use logb
> tlb=ts(log(b),frequency = 12)
> plot(tlb)
> densityplot(b)
> densityplot(log(b))
> #Question 3 ARIMA, parameters
> plot(b,type="l",main="Figure 1")
> plot(log(b),type="l",main="Figure 2")
> plot(sqrt(b),type="l",main="Figure 3")
> plot(sqrt(log(b)),type="l",main="Figure 4")
>
> #heroscedactity
> plot(decompose(ts(b,frequency = 12),type = c('multiplicative')))
> plot(decompose(ts(log(b),frequency = 12),type = c('multiplicative')))
```

The 6-month moving averages and the 12 month moving averages are shown on the curves of the following figures 11 and 12, as well as the ACF, figures 13 and 14.

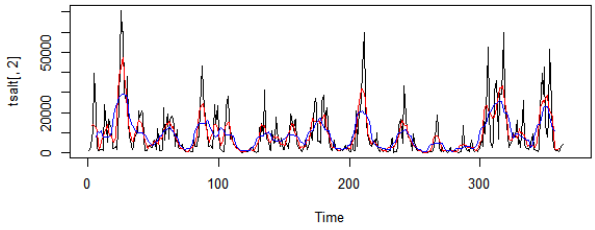


Figure 11: 6 and 12-month moving average of original data

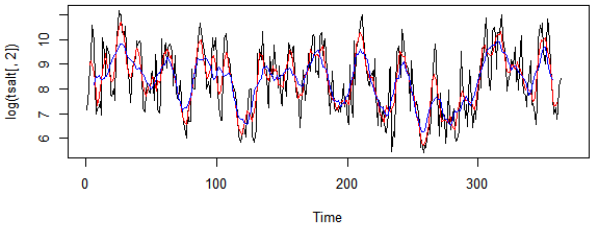


Figure 12: 6 and 12 months moving average of the transformed data

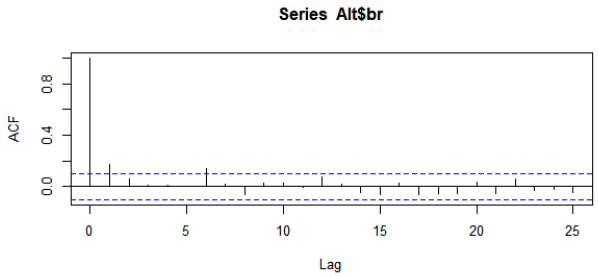


Figure 13: ACF of Brazos data

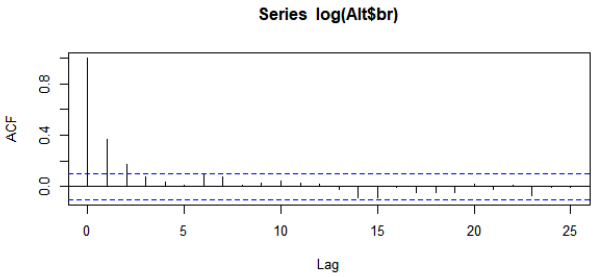


Figure 14: ACF of logarithmic transformation of data

```
names(Alt)=c("no","tp","br","month","year")
tsalt=ts(subset(Alt,select = c(-br)))
plot(tsalt[,2])
qaltl6=filter(tsalt[,2],filter = rep(1/6,6))
lines(qalt6,col="red")
qalt12=filter(tsalt[,2],filter = rep(1/12,12))
lines(qalt12,col="blue")
acf(Alt$br)
plot(log(tsalt[,2]))
qaltl6=filter(log(tsalt[,2]),filter = rep(1/6,6))
lines(qaltl6,col="red")
qaltl12=filter(log(tsalt[,2]),filter = rep(1/12,12))
lines(qaltl12,col="blue")
acf(log(Alt$br))
```

An optimal technique for testing the accuracy of a model and suitable parameters are estimating parameters for the model from 1990 to 2015 and predicting from 2015-2020. This gives the option to test the goodness of fit using r methods and using a logical method of comparing the predicted/forecasted data to the actual observed data from 2015-2020.

The ACF and PACF of the data from 1990-2015 will be presented in Figures 15 and 16.

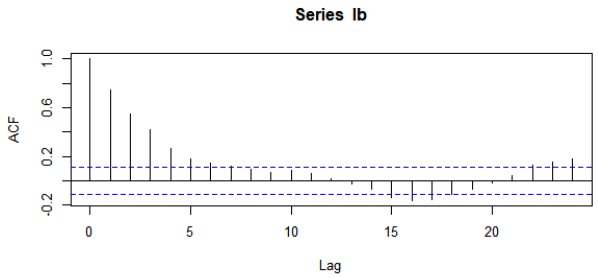


Figure 15: ACF of data 1990-2015

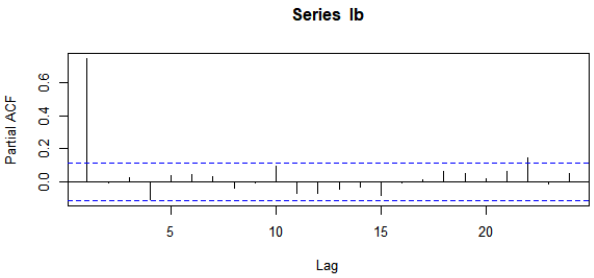


Figure 16: PACF of data 1990-2015

Using forecasting and auto.arima methods on r that compute predictions with regard to seasonality, the 2015-2020 forecasted data is modeled in Figure 17 below.

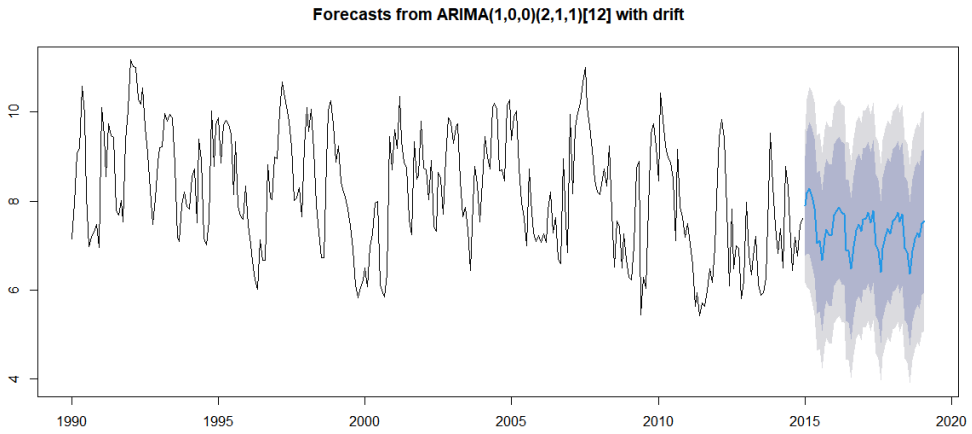


Figure 17: The forecasted values done by the auto.arima model in R

Comparing this with Figure 8, it is relatively accurate with lower variance.

Testing data can also be done by the box test. As the data set is relatively symmetric and normal (transformed), a p-value of bigger than 0.05 is needed.

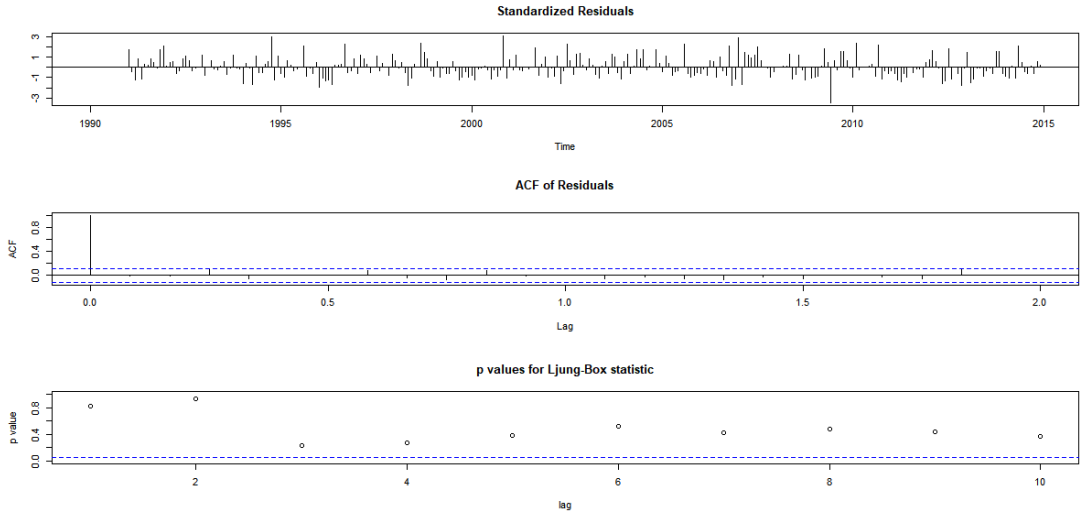


Figure 18: Residuals, ACF of Residuals, and p-values for the box test respectively

This figure shows a standard ACF for residuals and a normal distribution for residuals as the mean tends to zero and the variance to 1. The lag that needs to be used in the Box-test is 1 as it can be visually seen that it is where the ACF is the highest.

```
> #working with transformed data
> tsdiag(aab)
> Box.test(aab$residuals,lag = 1) #whicn is good as p value larger than 0.05

        Box-Pierce test

data:  aab$residuals
X-squared = 0.057827, df = 1, p-value = 0.81
```

This gives a p-value of 0.81>0.05.

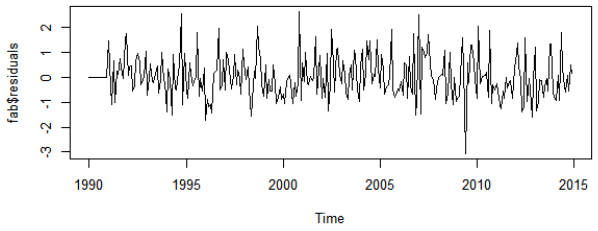


Figure 19: Time series of the residuals

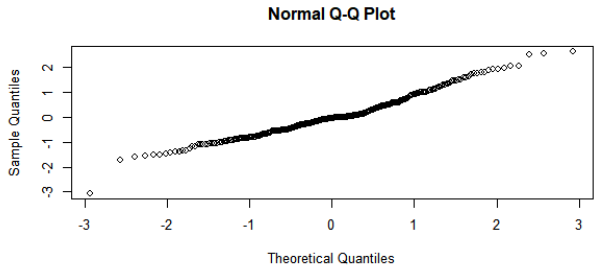


Figure 20: Normal plotting of residuals

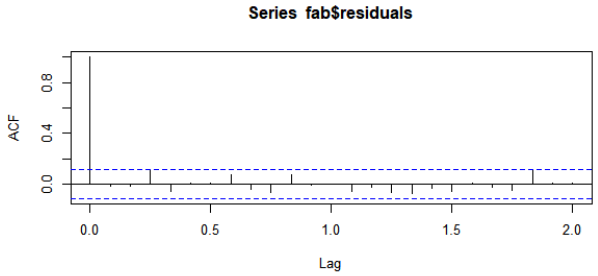


Figure 21: ACF of residuals

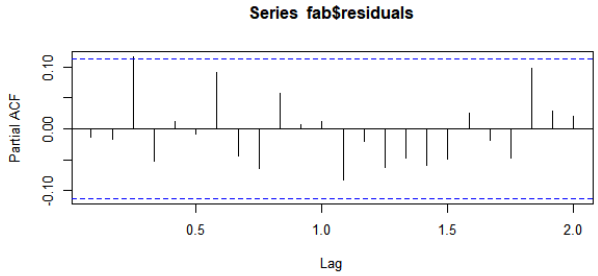


Figure 22: PACF of residuals

```
> par(mfrow=c(2,2))
> plot(fab$residuals)
> qqnorm(fab$residuals)
> acf(fab$residuals)
> pacf(fab$residuals)
```

As can be seen from figures 19 to 22 the residuals are normally distributed with optimal ACF and PACF graphs which indicates the goodness of fit.

The estimated parameters are given by the summary of the forecasted model in r. The forecasted model also shows how the forecasted values are constantly changing from increasing to decreasing, which shows that the seasonality effect is incorporated into the predictions.

```
> summary(fab)

Forecast method: ARIMA(1,0,0)(2,1,1)[12] with drift

Model Information:
Series: tslb
ARIMA(1,0,0)(2,1,1)[12] with drift

Coefficients:
         ar1     sar1     sar2     sma1    drift
      0.7117  -0.1772  -0.1295  -0.8641  -0.0048
s.e.  0.0420   0.0731   0.0711   0.0544   0.0022

sigma^2 estimated as 0.7614:  log likelihood=-378.47
AIC=768.95   AICc=769.24   BIC=790.92

Error measures:
                     ME      RMSE       MAE         MPE     MAPE      MASE        ACF1
Training set 0.02120659 0.8474941 0.6447274 -0.7051777 8.001393 0.4203058 -0.01303724

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95     Hi 95
Jan 2015       7.887185 6.768796 9.005573 6.176757  9.597613
Feb 2015       8.181784 6.809061 9.554506 6.082385 10.281182
Mar 2015       8.281260 6.796228 9.766292 6.010099 10.552420
Apr 2015       8.100040 6.561239 9.638840 5.746648 10.453431
May 2015       7.845756 6.280423 9.411089 5.451786 10.239726
Jun 2015       7.047914 5.469310 8.626517 4.633648  9.462179
Jul 2015       7.103884 5.518601 8.689168 4.679403  9.528366
Aug 2015       6.661552 5.072896 8.250208 4.231913  9.091192
Sep 2015       7.170817 5.580456 8.761177 4.738570  9.603063
Oct 2015       7.355005 5.763783 8.946226 4.921442  9.788568
Nov 2015       7.238799 5.647145 8.830452 4.804575  9.673022
Dec 2015       7.230178 5.638314 8.822042 4.795632  9.664724
Jan 2016       7.688535 6.096460 9.280609 5.253667 10.123402
Feb 2016       7.779622 6.187431 9.371813 5.344576 10.214668
Mar 2016       7.855941 6.263691 9.448191 5.420805 10.291077
```

The MAPE is 8, which means that this model is more or less 92% accurate. Which makes it a good fit. The RMSE is also 0.845 compared to an RMSE of 8501 for the original data and a MAPE of 187. This once again shows the impact of using a logarithmic transformation of the data and the goodness of fit of this modeling method.

```
28 aab=auto.arima(tslb,D=1) #parameters
29 aab$coef
30
29:9  (Top Level)                                   R Scrip

Console  Terminal  R Markdown  Jobs
~/
> aab=auto.arima(tslb,D=1) #parameters
> aab$coef
         ar1          sar1          sar2          sma1         drift
 0.711740093 -0.177235052 -0.129542032 -0.864066199 -0.004803419
>
```

The estimated parameters are ar1=0.712, sar1=-0.178, sar2=-0.13, sma1=-0.864 and the drift=-0.005.

```
~/
> auto.arima(ts(b),D=1) #which shows bayesian inference and aic down with a lot using log
Series: ts(b)
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1     ma1      mean
      0.6107  0.1407  9073.567
s.e.  0.0599  0.0751  1301.281

sigma^2 estimated as 72868539:  log likelihood=-3799.8
AIC=7607.6   AICc=7607.71   BIC=7623.18
> aats=auto.arima(ts(b),D=1) #which shows bayesian inference and aic down with a lot using log
> summary(aats)
Series: ts(b)
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1     ma1      mean
      0.6107  0.1407  9073.567
s.e.  0.0599  0.0751  1301.281

sigma^2 estimated as 72868539:  log likelihood=-3799.8
AIC=7607.6   AICc=7607.71   BIC=7623.18

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 17.5228 8500.96 5704.546 -161.521 186.9301 0.9967993 0.001683454
>
```

Another method that can be used is the method to normalize the data. The outcome of that is presented in figures 23 and 24 below.

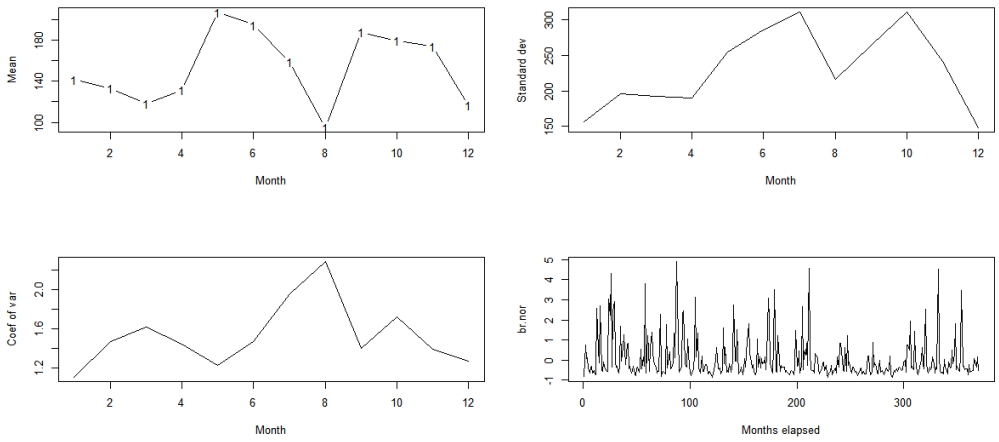


Figure 23: Normalized data mean, standard deviation, coefficient of variation:

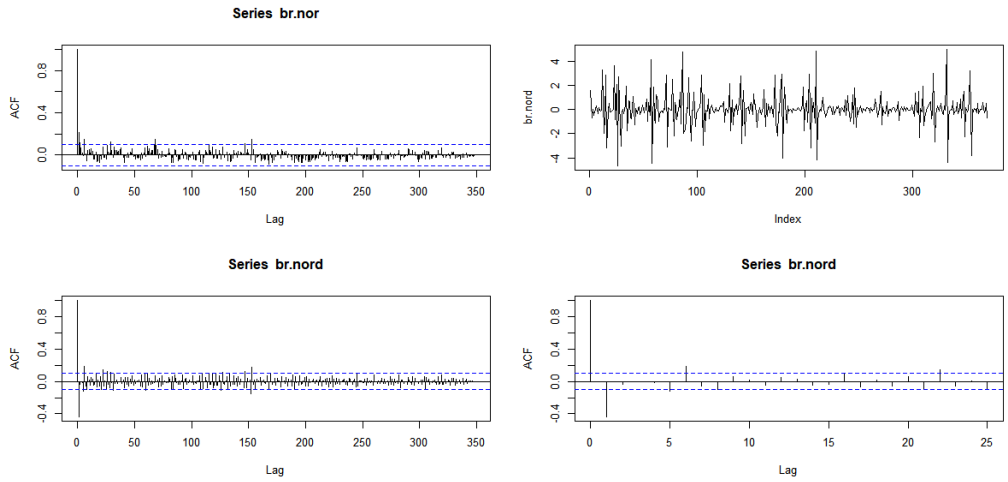


Figure 24: ACF and normalized data

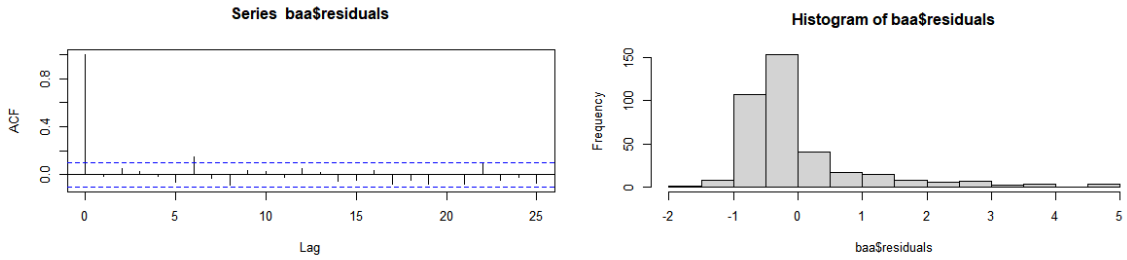


Figure 25: Distribution of residuals

From the visual inspection of figure 25, the skewness of the residuals is not acceptable, so the model is not used to forecast from as the normality assumptions would not hold. Thus, other methods prove more accurate in this instance.

The testing of the actual observed data and the predicted values still need to be analyzed to test the goodness of fit.

```
> dfb=data.frame(log(b[301:350]),fab)
> colb=c("Actual","Forecasted")
> names(dfb)=colb
> peb=(df$Actual-df$Forecasted)/(df$Actual)
> meanpeb=(mean(peb))
> rmse(df$Actual,df$Forecasted)
[1] 245.0407
> rmsle(df$Actual,df$Forecasted)
[1] 1.361294
```

The RMSE in this instance is higher than when the accuracy based on the fitted ARIMA model was test and in a more theoretical sense. The goodness of fit of the model when compared to actual data is not as good as in a theoretical sense. This makes sense as there are large standard deviations in the data from 2015-2020. In a theoretical sense, an ARIMA model adjusted for seasonality is a good fit with the given parameters, but in a real-world sense, it is not a good enough fit when modeling future monthly discharge from Brazos.

Holt-Winters is an alternative way of predicting values of a multiplicative model, and is presented with that of the actual series and ARIMA Forcast in figure 26 below.

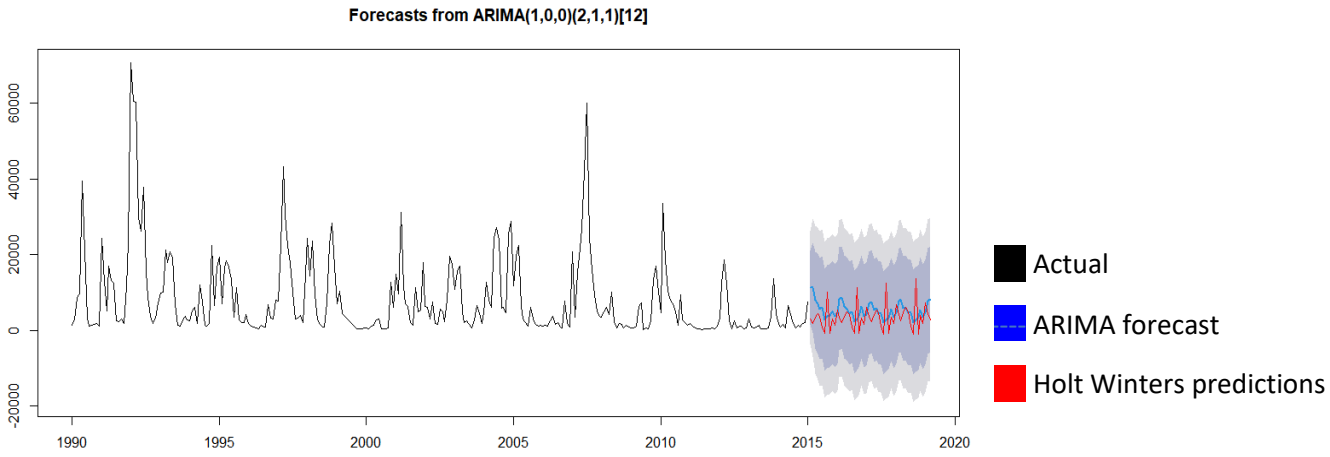


Figure 26: ARIMA and Holt-Winters forecasting compared to observed data from 1990-2020

The parameters of the Holt Winter model is given by R as:

```
> hwltp$coefficients
           a            b           s1           s2           s3           s4           s5
4633.4488522   49.7346451    0.6255573    0.3940572    0.7059497    0.9049963    0.7473546
          s6           s7           s8           s9          s10          s11          s12
   0.1639396   -0.1538639    2.0064180   -0.1361018    0.5598361    0.2684138    1.0422522
>
```

Visually, the Holt-Winters model proves to be better fitted to the data, as it takes higher variance into account, but the spikes differ from the spikes of the forecasting ARIMA model. The conclusion of the preferred model then depends on how good it fits the observed data. These models were tested on the original data as seen in figures 27 and 28.

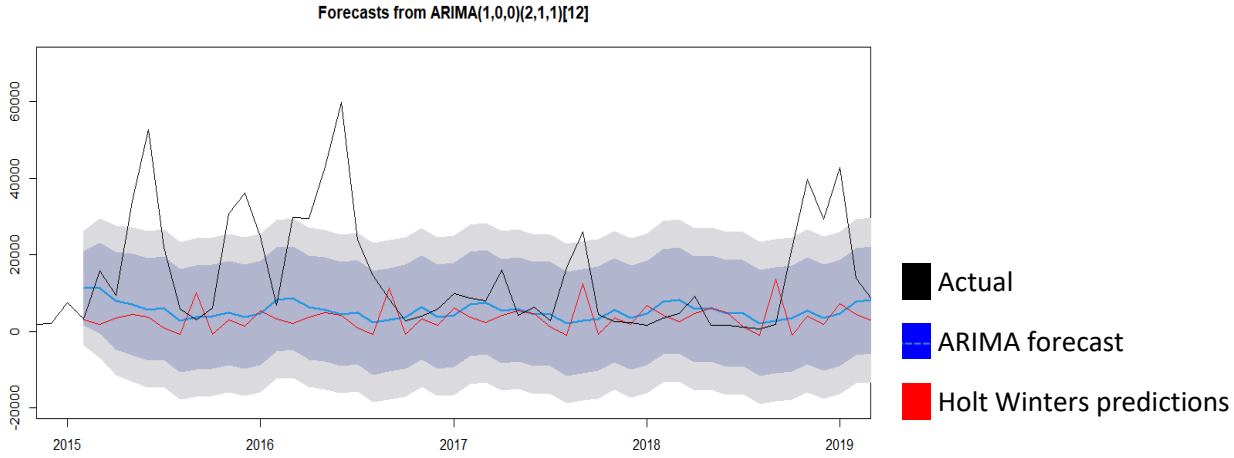


Figure 27: ARIMA and Holt-Winters forecasting compared to observed data from 2015-2020

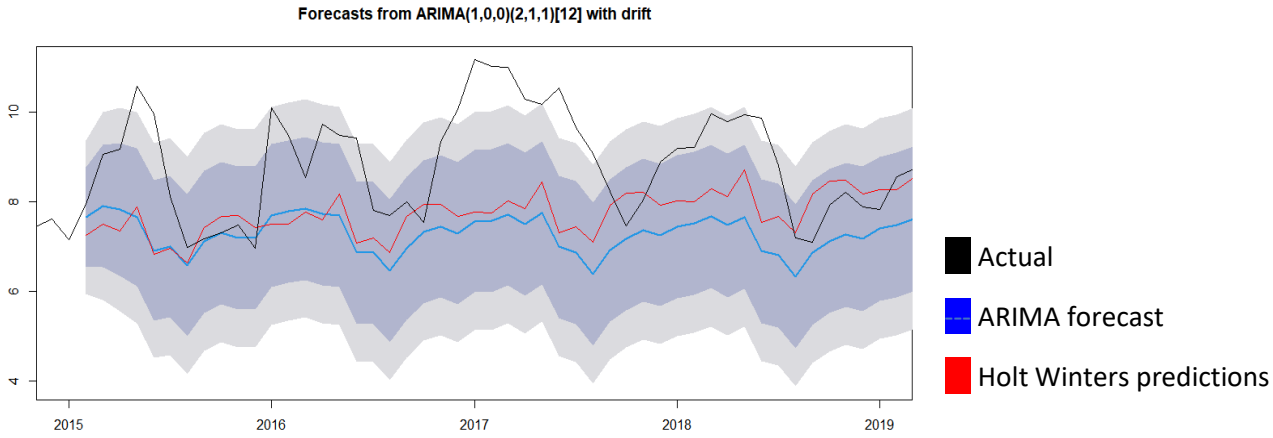


Figure 28: ARIMA and Holt-Winters forecasting compared to observed data from 2015-2020 of the logarithmically transformed data

HoltWinters and the ARIMA forecast model are both bad fits for the original data, but a better fit for the transformed data.

The parameters of the Holt-Winters model are given by R as:

```
> hwltp$coefficients
         a          b         s1         s2         s3         s4         s5         s6
7.64691715 0.02224633 0.94563141 0.97553045 0.95252972 1.02049290 0.88132561 0.89479300
        s7         s8         s9        s10        s11        s12
0.85060474 0.94870530 0.97809576 0.97729866 0.94097021 0.94870453
>
```

Visually the Holt-Winters model is a better fit as it is closer to the observed data. The residuals also have a normal distribution making it the best-fitted model as seen in figure 29 below.

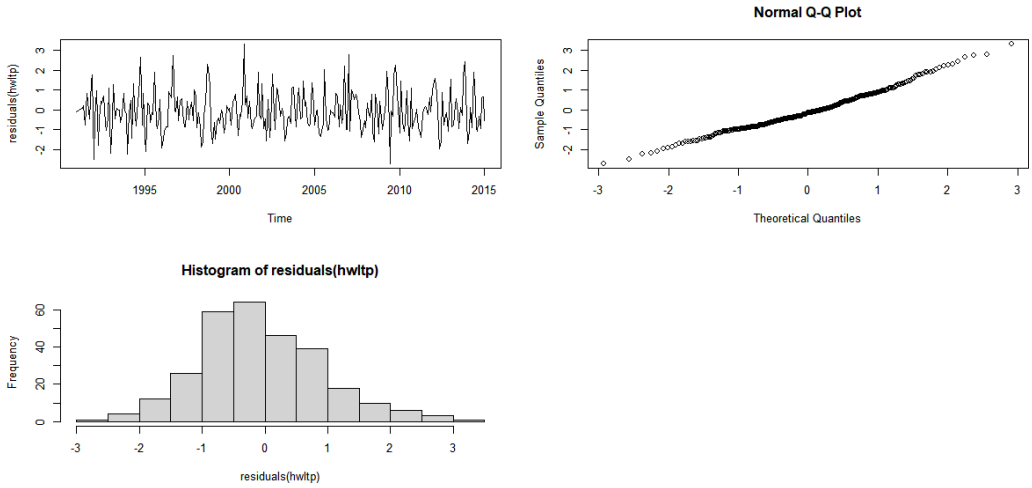


Figure 29: Holt-Winters residual distribution

Results – Cowleech Fork Sabine River

Cowleech Fork Sabine River station data transformations:

The null data was transformed to 0.001 for the purpose of working with log transformations. Mean cannot be used, as seasonality and trends have a big influence on the data and predictions. It is possible to also interpolate the corresponding months to hold the effect of seasonality, but the null data in this particular case is more useful to be interpreted as a value close to zero as there are other small values (less than 3) and this could give information on droughts, etc (Martin & Ruhl, 1993).

Transformations of data can make data easier to work with or interpret. The same transformations as with the previous Brazos river has been studied. In that case, it seemed like a good idea to work with the square root of the logarithmic transformation, but it was an unnecessary and more complicated

transformation. In this instance it is impossible to work with that transformation as the data for this specific model does not fit well with that transformation as seen in the figures below. This is due to all the values smaller than 1, where the log would be a negative value and the square root cannot be computed in a simple manner. The log transformation is sufficient to work with and it seems like it has a mean near zero just from a visual perspective, but the mean is in fact 1.7.

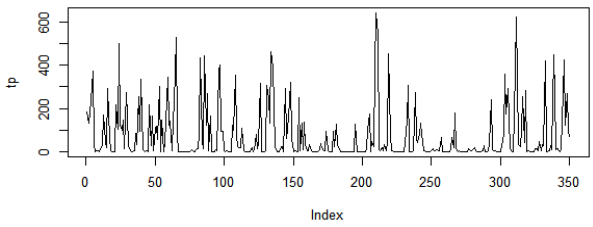


Figure 30: Original data

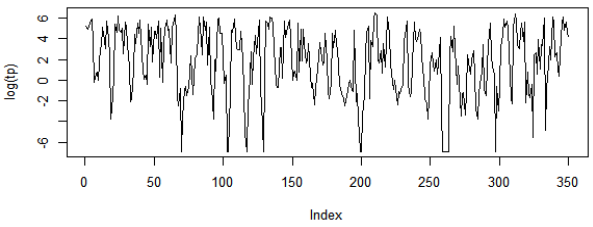


Figure 31: Square root of data

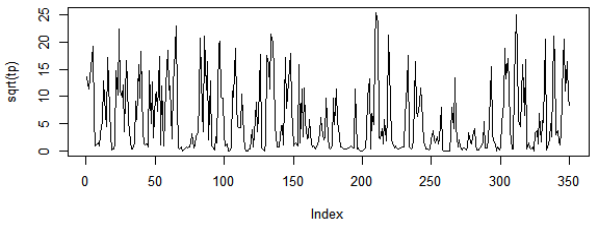


Figure 32: Logarithmic transformation of data

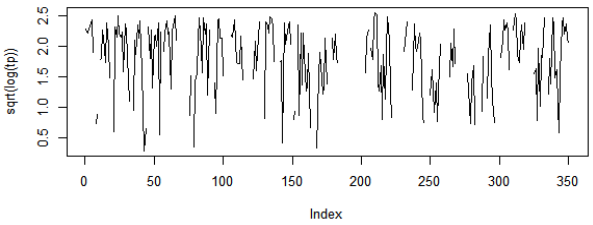


Figure 33: Square root of the logarithmic transformation of the data

The skewness of the data is 2.1 which is a high positive skewness. The data does not have a normal distribution. The logarithm transformation of the data has a skewness of -0.667 (Figure 35), which is a lot closer to 0 than the data without transformation (Figure 34). The transformed data (Figure 35) also has a relatively normal distribution, which makes it the preferred data to work with. Even though the density plot of the original data seems to have a mean of 0, its calculated mean is higher than 70, where the calculated mean of the transformed data is just higher than 1. The transformed data is presented again the data before transformation in figures 34 to 39 below.

```
> skewness(tp)
[1] 2.09998
attr(,"method")
[1] "moment"
> skewness(log(tp))
[1] -0.6665525
attr(,"method")
[1] "moment"
>
```

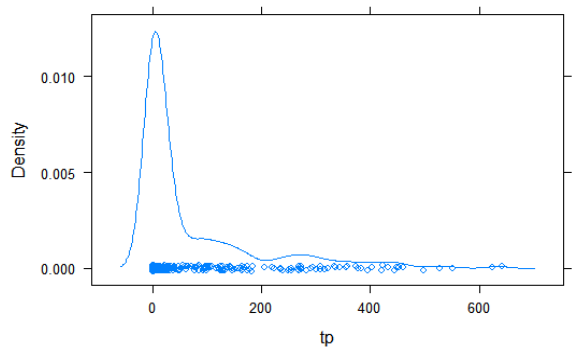


Figure 34: Density of the original data

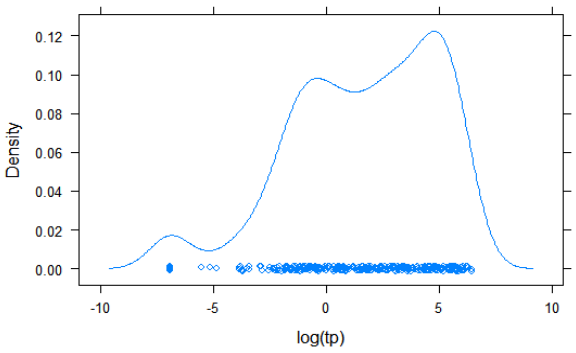


Figure 35: Density of the logarithmic transformation of the data

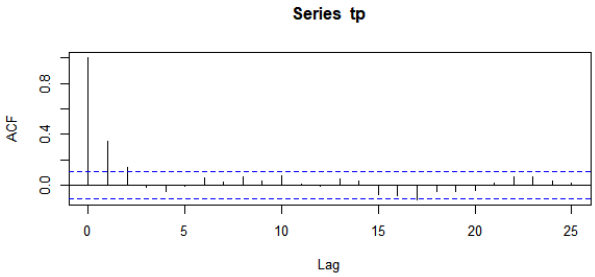


Figure 36: ACF of the original data

vs

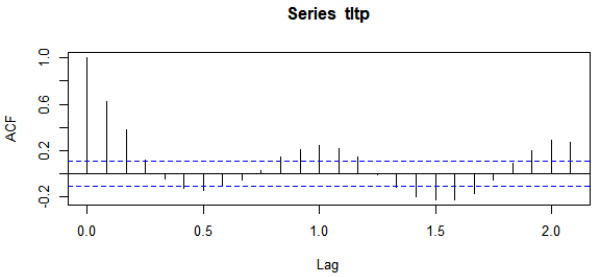


Figure 37: ACF of the transformed data

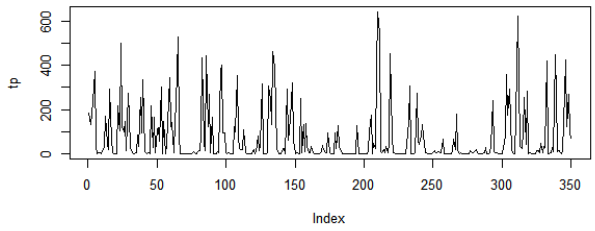


Figure 38: plotted original data

vs

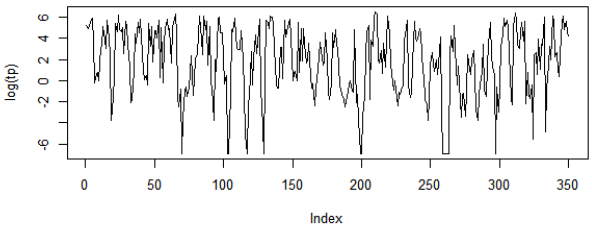


Figure 39: plotted transformed data

An important observation to make is the difference between variance in the original data and logarithm transformation of the data. Another reason why the transformed data is the preferred data set to make predictions with.

A decomposition of the time series when working with the original data was done: It shows the effect of seasonality, which plays a big role in predictions as all collected data is influenced by the seasonality of the area and predictions will also need to consider seasonality in certain models. In the decomposition, you can also see how the seasonality, trend, and observed values correlate. The model is also a multiplicative model and not an additive one. An example of an additive model would be if the data

showed a clear increase over the years caused by climate change or erosion for example, but it is only influenced by seasonality hence the multiplicative model. The decompositions of the multiplicative time series of the original data (Figure 40) and the transformed data (Figure 41) is presented below.

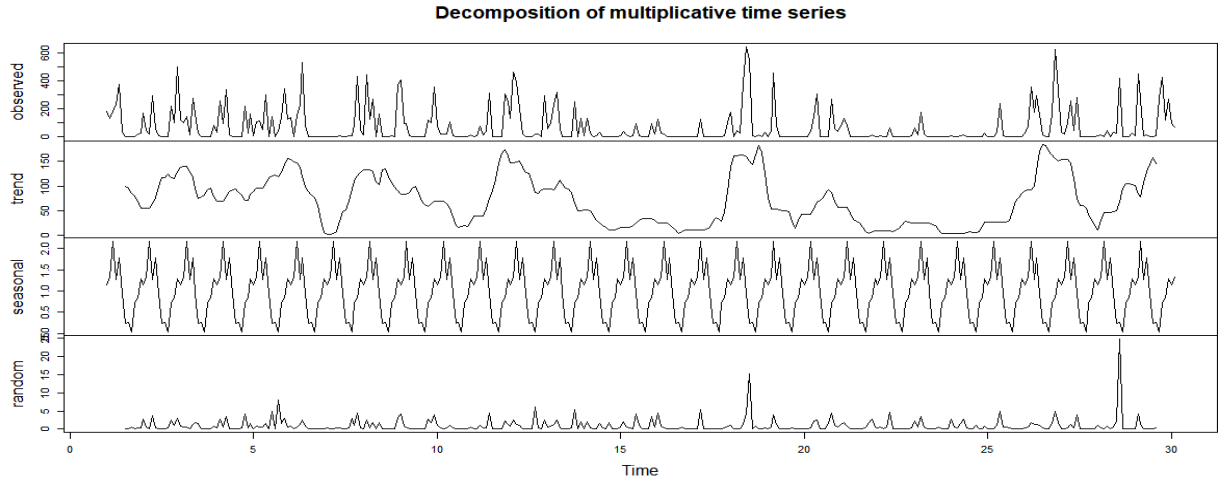


Figure 40: Decomposition of a multiplicative time series of the original data

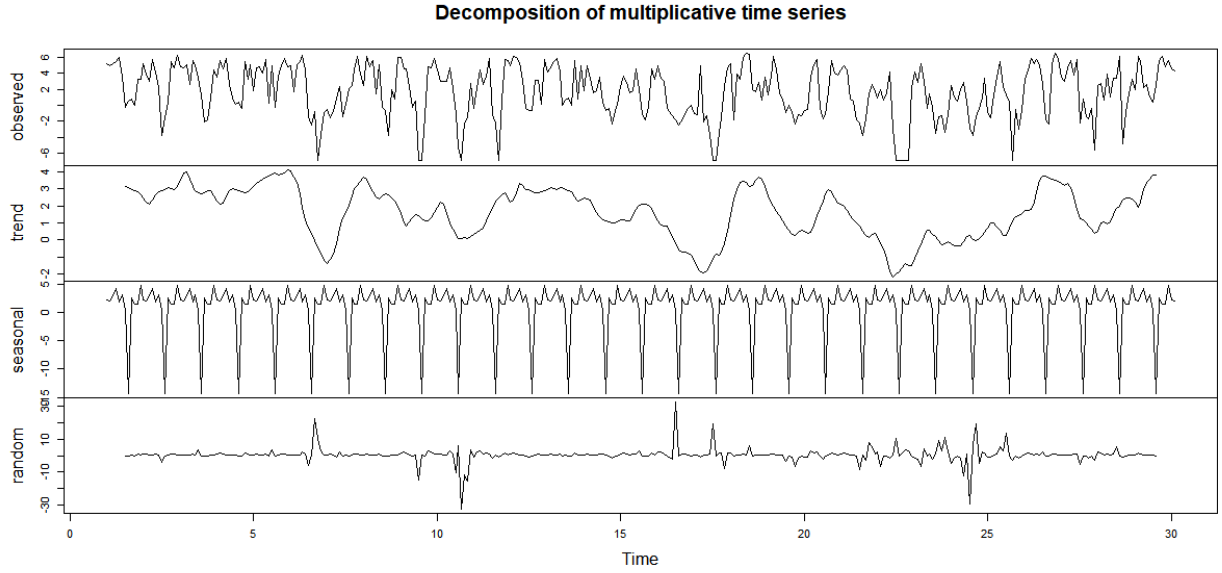


Figure 41: The decomposition of a multiplicative time series of the transformed data

The randomness is normalized with the transformation. Visually it seems to have reduced. The trends also seem easier to interpret with the transformation compared to the original data. The seasonality is assumed to be 12 months in cases where data is dependent on weather. It can be clearly seen and safely assumed as in the visual representation of the seasonality in the log transformation, there are 5 'bump curves' in every 5-year period.

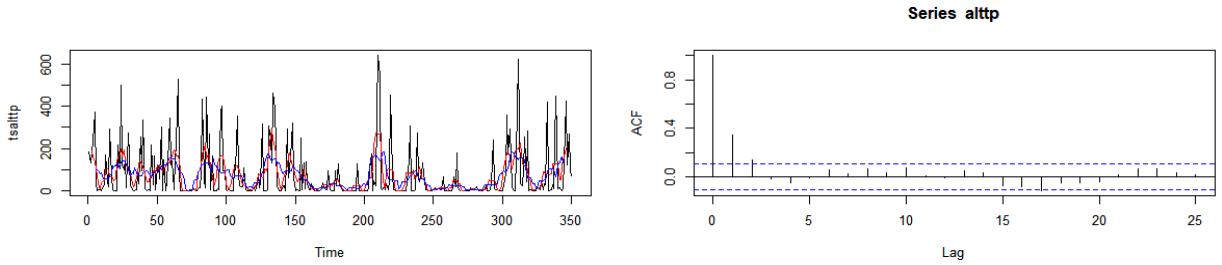


Figure 42: 6 month and 12-month moving averages of the original data and the plotted ACF

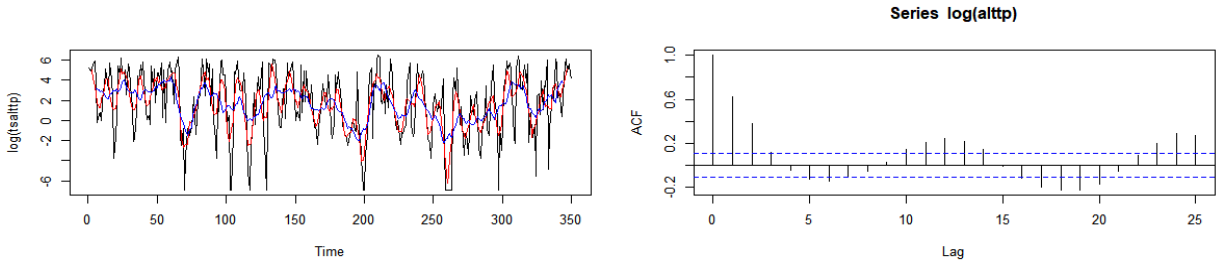


Figure 43: 6 month and 12-month moving averages of the transformed data and plotted ACF

In Figure 44, forecasting was done using the forecast and auto.arima functions in R, putting in a seasonality value. It can be seen that neither of the graphs takes the extent of variance into account. The function is used for the values of 1990-2015, to predict the values of 2015-2020 and compare them with the actual observed values. Figure 44 shows the forecasted values of the transformed data and original data respectively. The top graphs show the forecasting when seasonality is taken into account and the bottom graphs shows the forecast when seasonality is not taken into account. For the original data, this did not matter when forecasting, showing it is a poor model. For the transformed data, the forecasting done with seasonality is a clear better fit as it takes into account some level of variance.

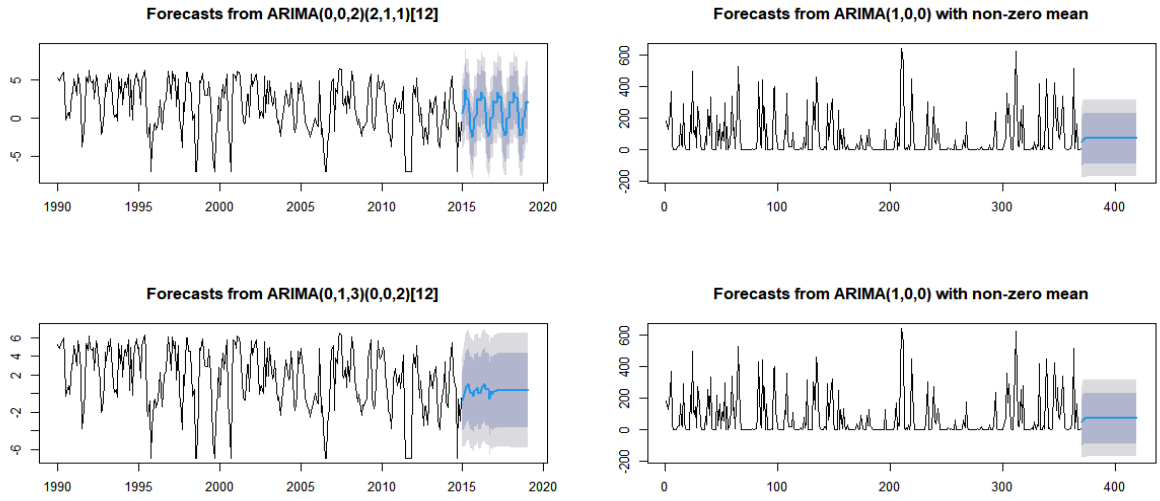


Figure 44: Plotted forecasted values done with auto.arima and forecasting in R.

Holt-Winters is a function in R that also predicts values. Here it can be seen in figure 45 that Holt-Winters is more accurate than the ARIMA model as it takes higher variance into account. The residuals still need to be taken into account as it shows the goodness of fit for the models. Here the forecasting of the original data vs the Holt-Winters prediction of original data is shown, showing just how accurate both these models are, taking seasonality into account and relevant variance (without needing the transformed data). The Holt-Winters model has the following parameters:

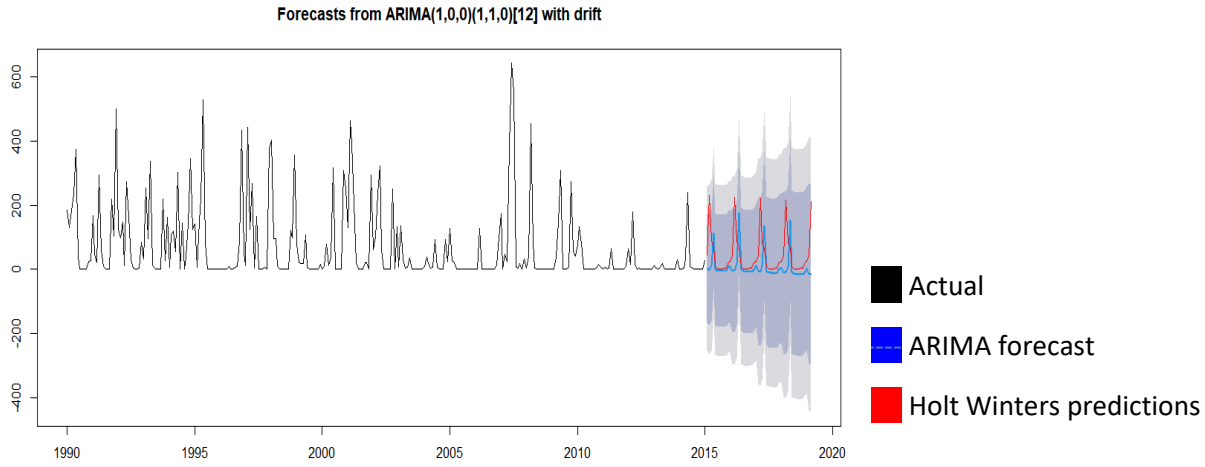


Figure 45: Plotted predicted values with Holt Winters in R vs the forecasted values from R

```
> hwltp$coefficients
             a              b             s1             s2             s3             s4
 9.589937e+02 -1.783878e+00  4.623550e-02  2.419845e-01  9.353012e-02  7.303078e-02
            s5             s6             s7             s8             s9            s10
 9.447900e-03  8.352378e-04  3.770958e-04  4.993405e-04  3.244492e-03  4.642765e-03
           s11            s12
 2.376472e-02  2.660376e-02
> |
```

Using these forecasted values from the logarithmic transformation, the normality of residuals and correlation between residuals can be shown in figures 46 to 43 to be sufficient to use in a forecasting model:

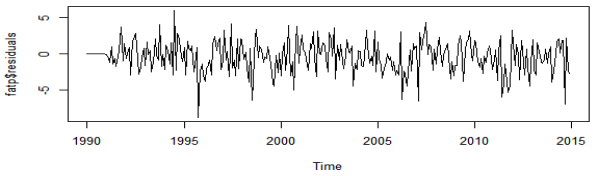


Figure 46: Time series of the residuals

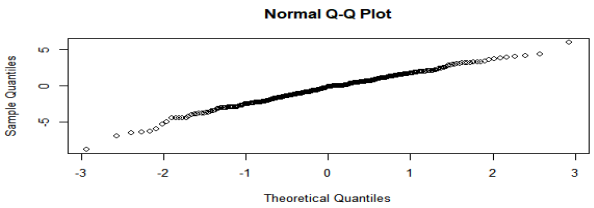


Figure 43: Normal plotting of residuals

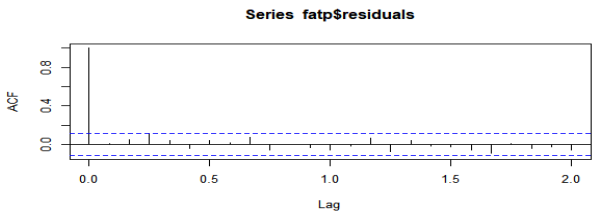


Figure 47: ACF of residuals

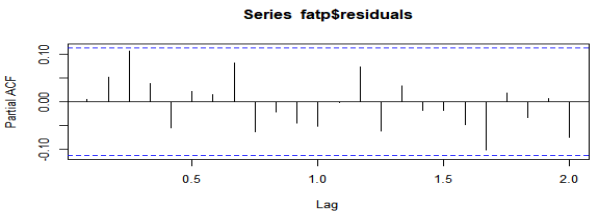


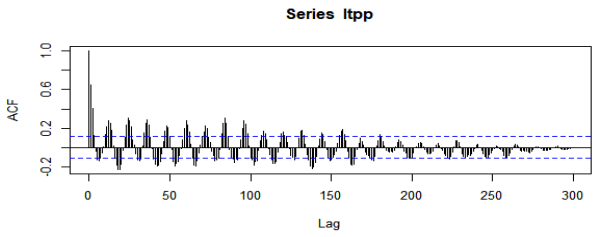
Figure 48: PACF of residuals

A summary of the logarithmic data and its specific ARIMA function, fitted to be ARIMA(0,0,2)(2,1,1)[12], gives parameters ma1=0.52, ma2=0.34, sar1=-0.13, sar2=-0.06, sma1=-0.88. The RMSE is 2.23, which is relatively high. The MAPE value is also very high, but lower with the transformed data than with the original data, which is an optimal outcome.

```
> summary(fatp)

Forecast method: ARIMA(0,0,2)(2,1,1)[12]

Model Information:
Series: tsltp
ARIMA(0,0,2)(2,1,1)[12]

Coefficients:
         ma1     ma2     sar1     sar2     sma1
      0.5146  0.3387  -0.1303  -0.0578  -0.8808
s.e.  0.0566  0.0513   0.0705   0.0678   0.0495

sigma^2 estimated as 5.268:  log likelihood=-656.47
AIC=1324.95   AICc=1325.25   BIC=1346.93

Error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set -0.4096592 2.229148 1.729645 578.6329 677.7606 0.5629659 0.005458918

Forecasts:
         Point Forecast       Lo 80     Hi 80     Lo 95    Hi 95
Jan 2015     0.56594342 -2.37607224 3.5079591 -3.933481 5.065368
Feb 2015     1.66522459 -1.64337370 4.9738229 -3.394840 6.725289
Mar 2015     3.76681094  0.31147330 7.2221486 -1.517672 9.051294
```

Again, good way to check the goodness of fit for the model would be forecasting from data in the years collected in 1990-2015 to predict values for 2015-2020 and to check the accuracy.

An analysis needs to be done between the observed values of 2015-2020 and the forecasted/expected values of 2015-2020 (presented figure 50).



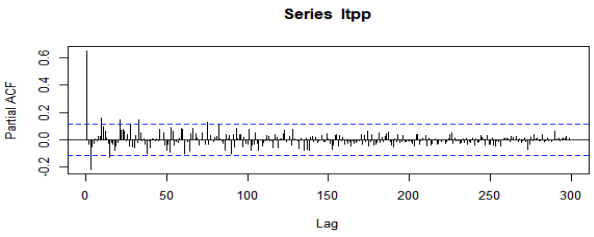


Figure 49: The plotted ACF and PACF (of the logarithmic data from 1990-2015) respectively:

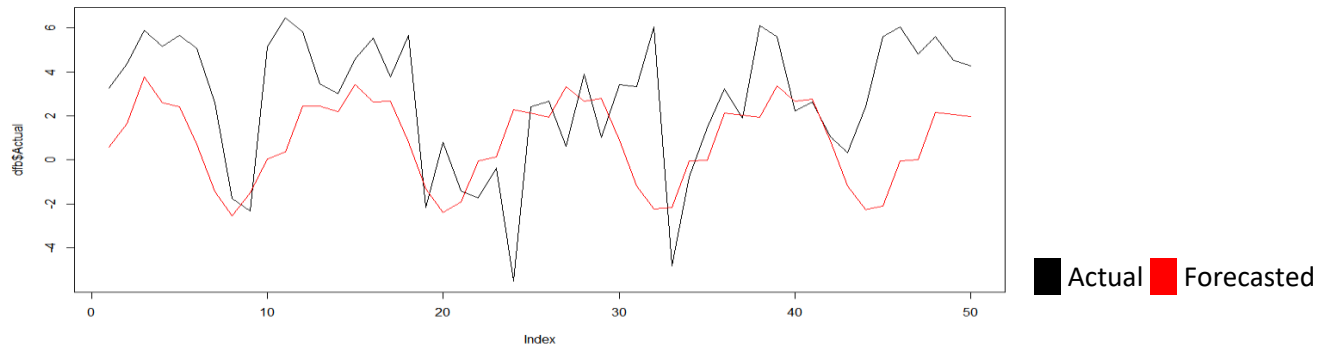
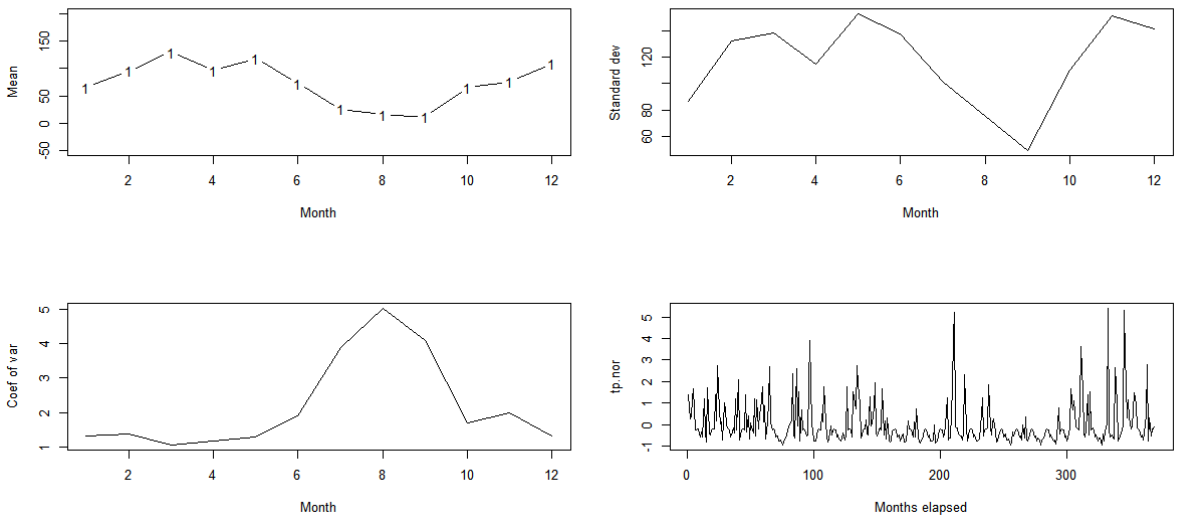


Figure 50: The observed and ARIMA forecasted data

The mean percentage error is 0.625. The RMSE of the Actual data vs the Forecasted data is 3.366. Which is higher than the RMSE of the previous forecasting ARIMA model. So, although the model seems visually accurate if assessing the goodness of fit of the ARIMA forecast, the RMSE is higher than previously calculated.

Another method that can be used is to normalize the data. This method takes means of different months into account, giving a visual representation of lower discharge levels in summer months. Sketches of the data are provided in figure 51 to show an alternative method (normalizing data) but will not be explained in detail as it is not a fitting model for the data.



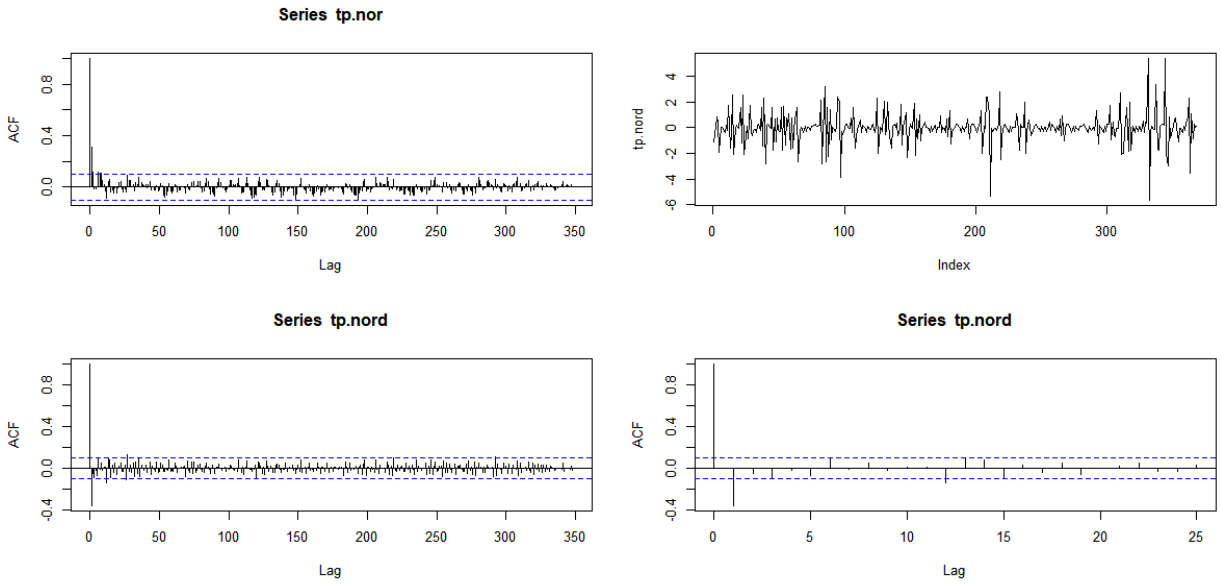


Figure 51: Normalized data

This method will not be discussed in detail, as its residuals are not normally distributed as can be seen in the following figure 52. There is no sufficient correlation between the residuals, which makes it a tempting method to use, but the distribution of the residuals must be normal.

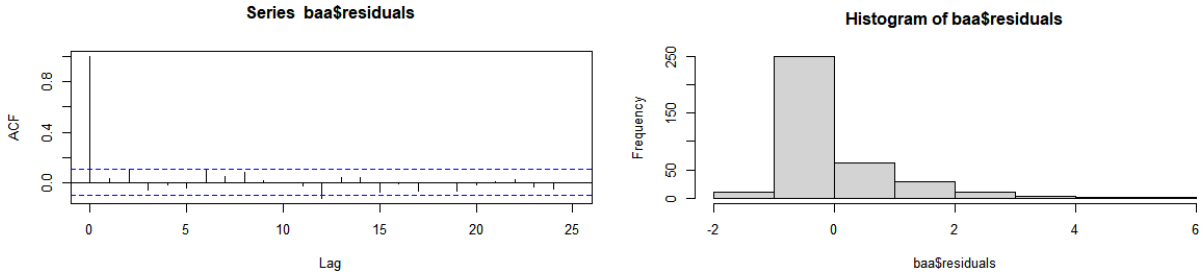


Figure 52: Residuals of normalized data

For the Cowleech Sabine River, the Holt-Winter model would be the best fit, but not the ARIMA nor the Hol-tWinters would be a good fit as can be seen in Figure 53 when looking at the original data. The logarithmic transformation ARIMA forecast would be the best model in this instance.

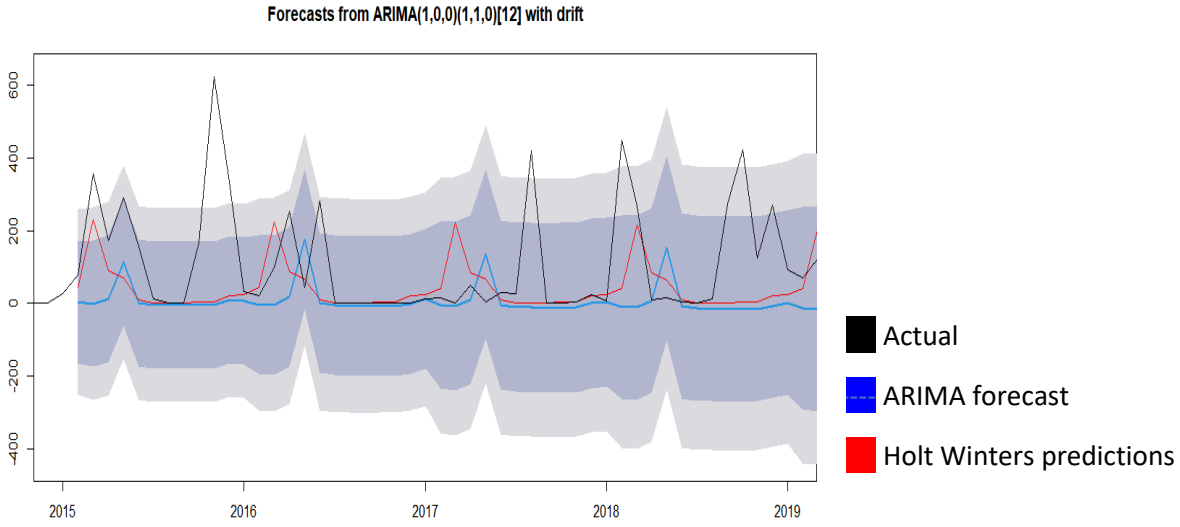


Figure 53: ARIMA and Holt-Winters forecasting compared to observed data from 2015-2020 of the original data

Both forecasting methods have the same shape as the observed data, but the spikes occur at the wrong time and the variance is not fitted. Therefore, it can be concluded the Holt-Winters model is not an acceptable fit.

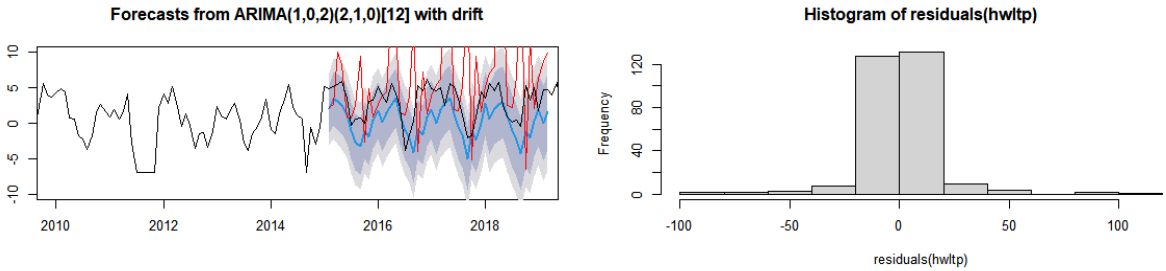


Figure 54: ARIMA and Holt-Winters forecasting compared to observed data from 2015-2020 of the logarithmically transformed data and the residual distribution of Holt-Winters

The residuals have a wide distribution, and the variance is too large. Here the ARIMA model would be a better fit.

Concluding Remarks

The correlation between the two data sets is 0.26, it is not strong enough to assume a method that works for the one data set will hold for the other, but not 0, which means there is some type of correlation which could be because of the seasonality of the data. As a rule, there will always be a correlation between any data that is dependent on weather. The positive correlation shows it is in the same region, experiencing winters and summers at the same time. It would be a negative correlation if not. The fact that the correlation is not very strong, could be due to the altitude of the respective rivers and other location details that are not weather-related.

This is seen by the different ways the data is fitted and the accuracy of the fitted models. The parameters are also of different signs.

Modeling the Brazos river's monthly discharge, the best-fitted model is the Holt-Winters model of the transformed data, with parameters:

```
> hwltp$coefficients
         a          b         s1         s2         s3         s4         s5         s6
7.64691715 0.02224633 0.94563141 0.97553045 0.95252972 1.02049290 0.88132561 0.89479300
        s7         s8         s9        s10        s11        s12
0.85060474 0.94870530 0.97809576 0.97729866 0.94097021 0.94870453
>
```

The residuals are normally distributed, and the predicted values fit the observed values the best. The mean percentage error is 0.178 for the ARIMA forecast and 0.098 for the Holt-Winters model, which is also optimal. The Box test gives a p-value of 0.82 which is bigger than 0.05, which is sufficient. All these values are obtained when the model is compared to the actual data.

Modeling the Cowleech Fork Sabine River's monthly discharge, the best-fitted model would be the ARIMA forecast model, with parameters; ma1=0.52, ma2=0.34, sar1=-0.13, sar2=-0.06, sma1=-0.88.

The RMSE is 2.23, which is relatively low compared to the RMSE of the Holt-Winters model that was calculated as 12.45. The mean percentage error is 0.625, which is also lower than the 2.81 calculated with the Holt-Winters model. The Box test gives a p-value of 0.925 which is sufficient as the p-value needs to be bigger than 0.05.

Both river's best modeling method and parameters have been chosen by a goodness of fit method, comparing the RMSE's and the methods with the lowest RMSE has been chosen for each river, respectively.

The Brazos river's Holt-Winters model could be used to model future predictions, but it should be noted that it does not take the full extent of the variance into account, it shows trends and seasonality, but the model underestimates the monthly discharge of the Brazos river by large amounts.

When using an ARIMA model as required it would be the ARIMA(1,0,0)(2,1,1)[12] model. This model has an optimal RMSE but it has been modeled with all the available data, leaving no room to test it against observed values.

The estimated parameters are given by the summary of the forecasted model in r. The forecasted model also shows how the forecasted values are constantly changing from increasing to decreasing, which shows that the seasonality effect is incorporated into the predictions.

```
> summary(fab)

Forecast method: ARIMA(1,0,0)(2,1,1)[12] with drift

Model Information:
Series: tslb
ARIMA(1,0,0)(2,1,1)[12] with drift

Coefficients:
         ar1     sar1     sar2     sma1    drift
      0.7117  -0.1772  -0.1295  -0.8641  -0.0048
s.e.  0.0420   0.0731   0.0711   0.0544   0.0022

sigma^2 estimated as 0.7614:  log likelihood=-378.47
AIC=768.95   AICc=769.24   BIC=790.92

Error measures:
                     ME      RMSE       MAE        MPE     MAPE      MASE        ACF1
Training set 0.02120659 0.8474941 0.6447274 -0.7051777 8.001393 0.4203058 -0.01303724

Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95     Hi 95
Jan 2015       7.887185 6.768796 9.005573 6.176757  9.597613
Feb 2015       8.181784 6.809061 9.554506 6.082385 10.281182
Mar 2015       8.281260 6.796228 9.766292 6.010099 10.552420
Apr 2015       8.100040 6.561239 9.638840 5.746648 10.453431
May 2015       7.845756 6.280423 9.411089 5.451786 10.239726
Jun 2015       7.047914 5.469310 8.626517 4.633648  9.462179
Jul 2015       7.103884 5.518601 8.689168 4.679403  9.528366
Aug 2015       6.661552 5.072896 8.250208 4.231913  9.091192
Sep 2015       7.170817 5.580456 8.761177 4.738570  9.603063
Oct 2015       7.355005 5.763783 8.946226 4.921442  9.788568
```

The MAPE is 8, which means that this model is more or less 92% accurate. Which makes it a good fit. The RMSE is also 0.845 compared to an RMSE of 8501 for the original data and a MAPE of 187. This once again shows the impact of using a logarithmic transformation of the data and the goodness of fit of this modeling method. The estimated parameters are ar1=0.712, sar1=-0.178, sar2=-0.13, sma1=-0.864 and the drift=-0.005. This is the best model according to r when it is not compared to observed data.

The Cowleech river's ARIMA model visually shows the goodness of fit.

The ARIMA model can be used to forecast the trends and seasonality and to some extent the discharge levels, but it should be noted that this model also underestimates the monthly discharge and in rare cases overestimates it by large which can be problematic. Visually it is a good fit for predicting future discharge.

It is worth noting that even if the RMSE is higher for the Cowleech ARIMA model than for the Brazos ARIMA model, the RMSE is a relative measurement and should be seen that way. It does not make either model a better fit, because it is calculated using different data, but rather be compared to the size of the transformed data. The Brazos river's size of its logarithmic data is about 3, where the Cowleech river's logarithmic data size is about 11.

Visually and theoretically the Cowleech Fork Sabine River's future discharge could be relatively accurately modeled and it has a fitted ARIMA model(ARIMA(0,0,2)(2,1,1)[12]). It can be used as an accurate expectation but must be fitted every few years to maintain accuracy, decisions around drought preparation should also not only depend on this model as the rare cases where it overestimates, it is by a large amount. The Brazos river has an ARIMA(1,0,0)(2,1,1)[12] model as the best model out of the options, but it is generally not a good fit and should not be used for predicting future floods or any modeling that needs possible maximum values but can be used to predict possible future droughts or any modeling that would need possible minimum values.

Reference List

Cepeda Cuervo, E., Achcar, J. A., and Andrade, M. G. (2018). "Seasonal hydrological and meteorological time series." *Earth Sciences Research Journal*, 22(2), 83–90.

Martin, G. R., & Ruhl, K. J. (1993). Regionalization of harmonic-mean streamflows in Kentucky (Vol. 92, No. 4173). US Department of the Interior, US Geological Survey.

Modal, M. S., & Wasimi, S. A. (2006). Generating and forecasting monthly flows of the Ganges river with a par model. Journal of Hydrology, 323(1-4), 41-66.

Tesfaye, Y. G., Meerschaert, M. M. & Anderson, P. L. (2006). Identification of periodic autoregressive moving average models and their application to the modeling of river flows. Water Resources Research, 42(W01419), 1-11.

Wang, Q. J., Robertson, D. E. & Chiew, F. H. S. (2009). A Bayesian joint probability modeling approach for seasonal forecasting of stream- flows at multiple sites. Water Resources Research, 45(W05407), 1-18.