**The assignment 2 report**

Name: Haiyue Wang

Content

[Dataset of Clements Gap Wind Farm Output 3](#_Toc147606176)

[1.1 Find the best *ARMA(p,q)* model for the 2011 data. 3](#_Toc147606178)

[1.2 Take the noise Zt from that model and check its *SACF*. 4](#_Toc147606179)

[1.3 Calculate squared noise and show that it has the *ARCH* effect. 4](#_Toc147606180)

[1.4 Find the best *ARCH* or *GARCH* model for it. 5](#_Toc147606181)

[1.5 Apply all models to the 2012 output data. 6](#_Toc147606182)

[1.6 Evaluate the performance 7](#_Toc147606183)

[1.7 Compare the results 8](#_Toc147606184)

[Dataset of Melbourne Airport Rain 9](#_Toc147606185)

[2.1 Test normality for December, January, February, July, August 9](#_Toc147606188)

[2.2 Test for Gamma fit. 9](#_Toc147606189)

[2.2.1 Get the parameters for gamma 9](#_Toc147606190)

[2.2.2 Visualize the distribution 10](#_Toc147606191)

[2.3 Test the correlation 10](#_Toc147606192)

[2.4 Synthetic and CDFs 10](#_Toc147606193)

[2.5 Synthetic and CDFs for July, August. 11](#_Toc147606194)

[Datasets of Mt Gambier By Months Temperature and Mt Gambier Rainfall 11](#_Toc147606195)

[3.1 Model the seasonality and then get residuals. 11](#_Toc147606199)

[3.2 Use exponential smoothing to see the overall trend 12](#_Toc147606200)

[3.3 Find the trends for the smoothed data of whole series and various sections 13](#_Toc147606201)

[3.3.1 Find the trend of whole dataset. 13](#_Toc147606202)

[3.3.2 Split whole dataset into multiple sections. 14](#_Toc147606203)

[3.3.3 find the trends for the two sections 14](#_Toc147606204)

[3.4 Take the data for the month of December and the Annual mean temperature from MtGambierByMonthsTemperature.xlsx and find the trend over time. 15](#_Toc147606205)

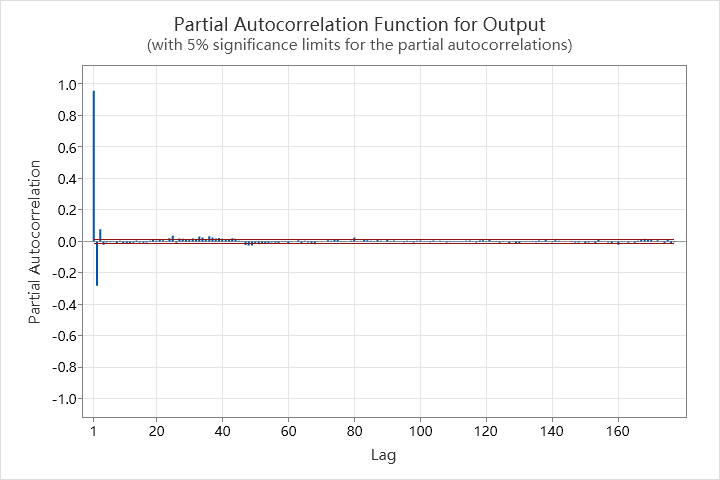
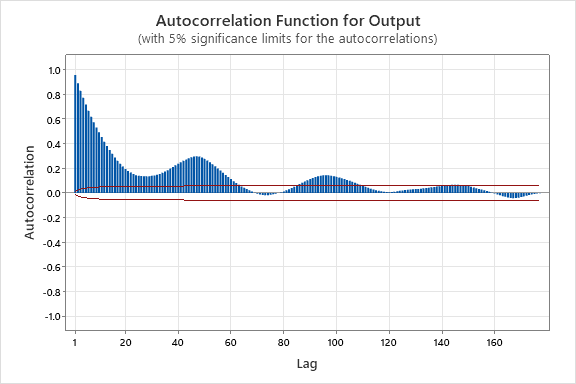
[3.5 How much has the mean temperature changed over time in each case? 15](#_Toc147606206)

# Dataset of Clements Gap Wind Farm Output



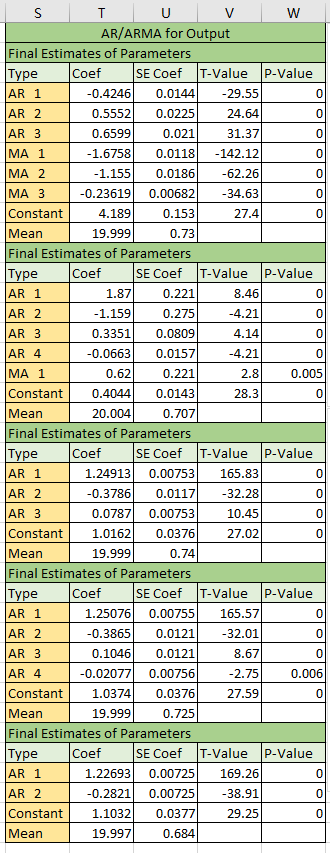
## Find the best ARMA(p,q) model for the 2011 data.

The first step is to check whether the dataset suit for ARMA model via ACF and PACF. The result like the pictures below.

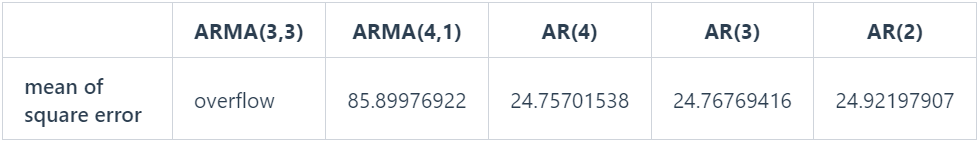


According to the graphs, it’s reasonable to use ARMA model for this dataset.

Next step is to find a proper ARMA(p, q) model for the dataset. There are five models that could be found for the dataset. The parameters like the picture below.



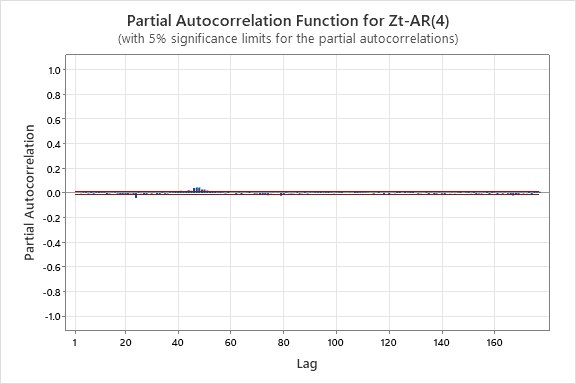
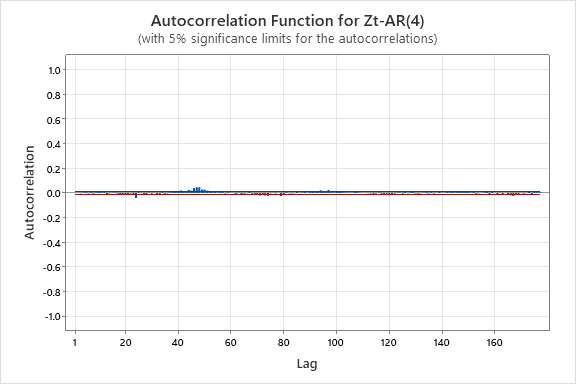
For selecting a proper model, we should compare the mean of squared error (MSE) and p-value of the parameters for each model, the smaller the better. The table below lists all the MSEs.



According to the mean of square error, the performance of AR(3) and AR(4) are similar, I select the **AR(4)** as the best model.

## Take the noise Zt from that model and check its *SACF*.

The residuals based on the model AR(4) could get, and name the residuals as Zt-AR(4), then, we could see the SACF according to the ACF and PACF, like the pictures below.

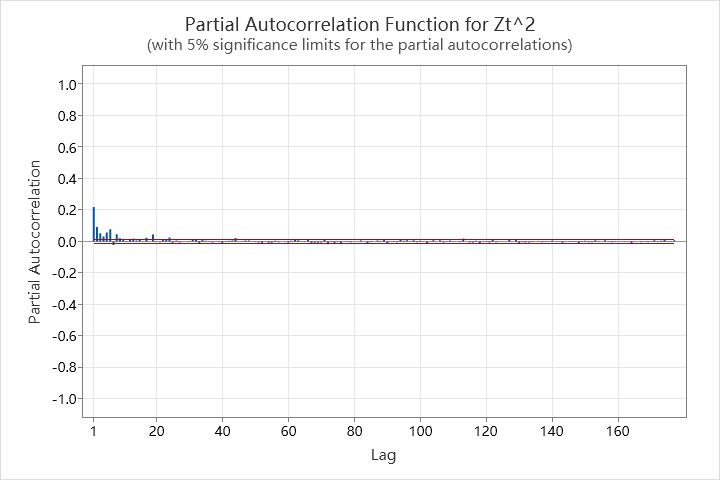
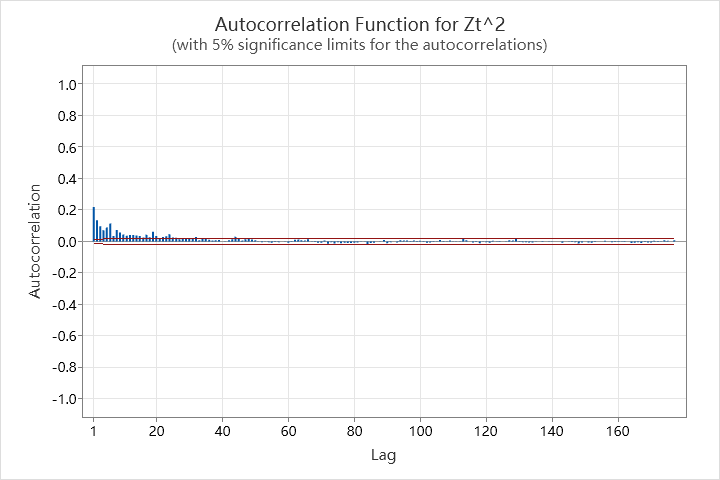


According to the result from ACF and PACF, the **Zt-AR(4)** is not suit for ARMA model.

## Calculate squared noise and show that it has the *ARCH* effect.

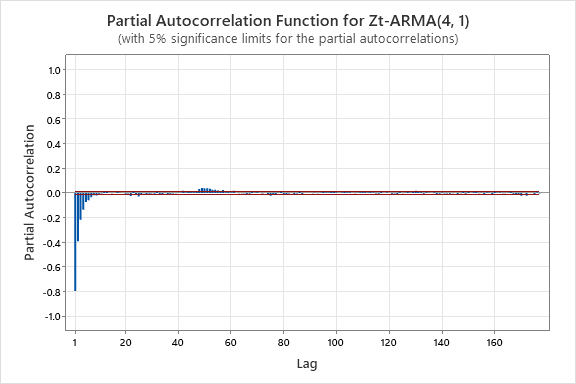
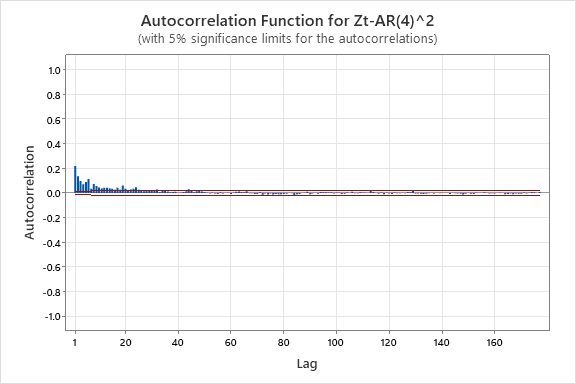
For calculating the **ARCH** effect, it should be seperated into two parts, one for **ARCH** model, another for **GARCH** model.

* + 1. Effect for ARCH model



According to the SACF above, there is ARCH effect for ARCH model.

* + 1. Effect for GARCH model



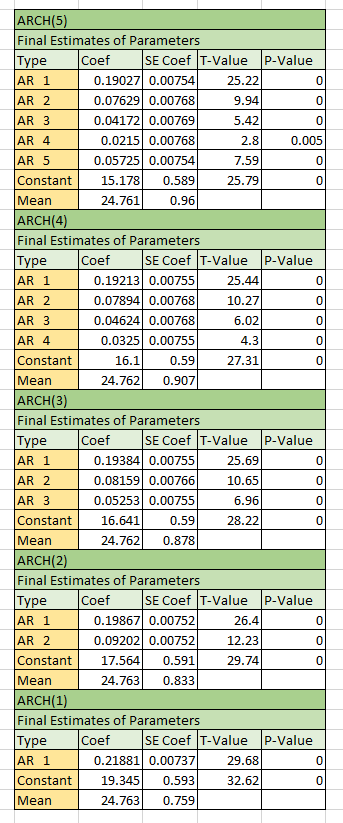
According to the SACF above, there is also ARCH effect for GARCH model.

## Find the best *ARCH* or *GARCH* model for it.

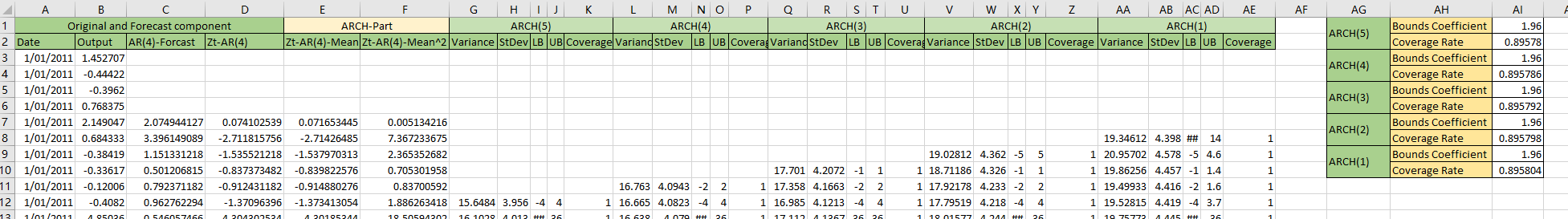
For this part, I try to find all possible ARCH and GARCH models, and then try to compare the results for finalizing the model.

* + 1. **For ARCH model**

According to the squared residuals for ARCH model, five **AR** models for the dataset could be found, the parameters like the picture below.



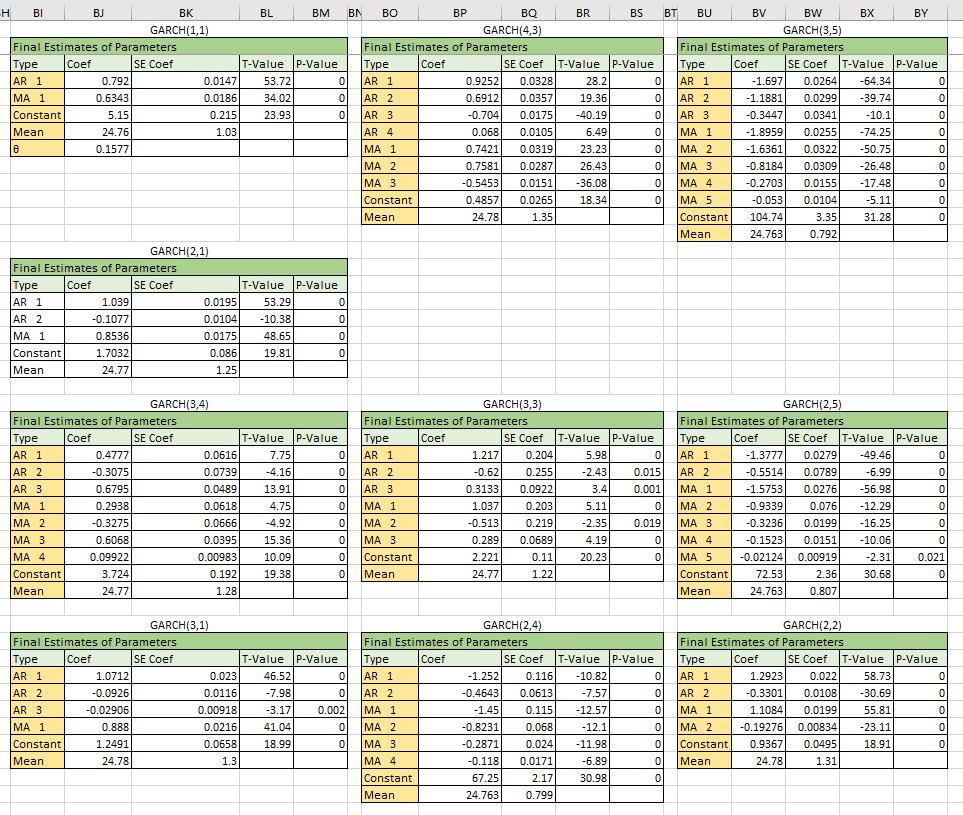
According to the parameters above, the coverage rate could be calculated like the picture below.



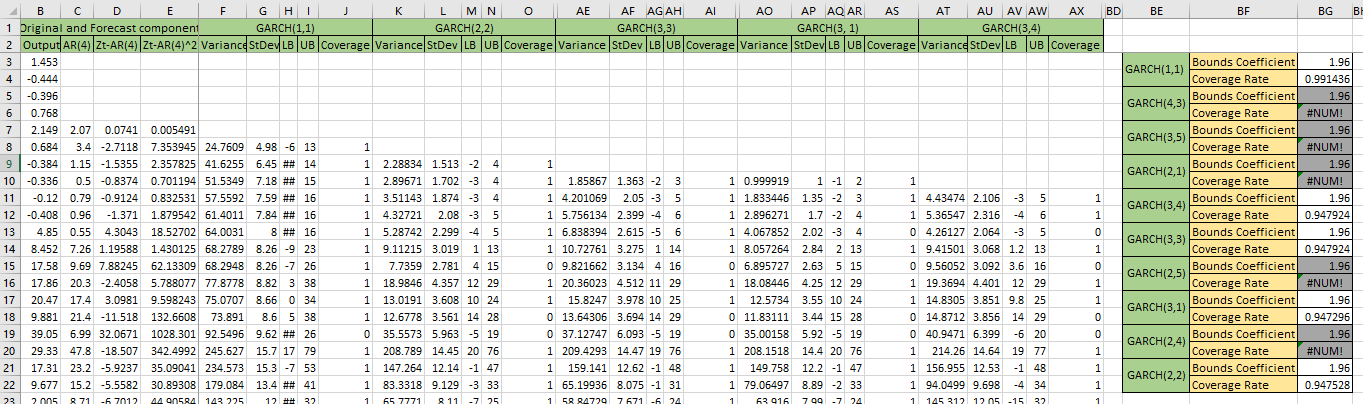
According to the result, all the coverage rate for the five model are similar, and they all approach 89.6% for score 1.96.

* + 1. **For GARCH model**

According to the squared residuals for GARCH model, 10 **ARMA** model for the dataset could be found, the parameters like the picture below.



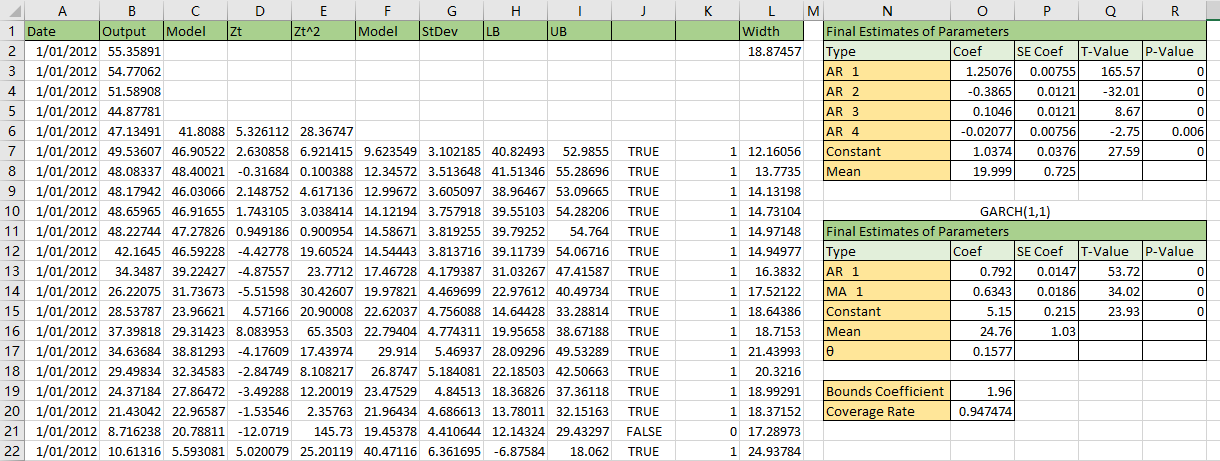
For some models will occurs negative values, which will lead to the specified model unavaliable. The coverage rate for each model like the picture below.



According to the result the best model for the residuals is **GARCH(1,1)**, the coverage could be 99.15%.

## Apply all models to the 2012 output data.

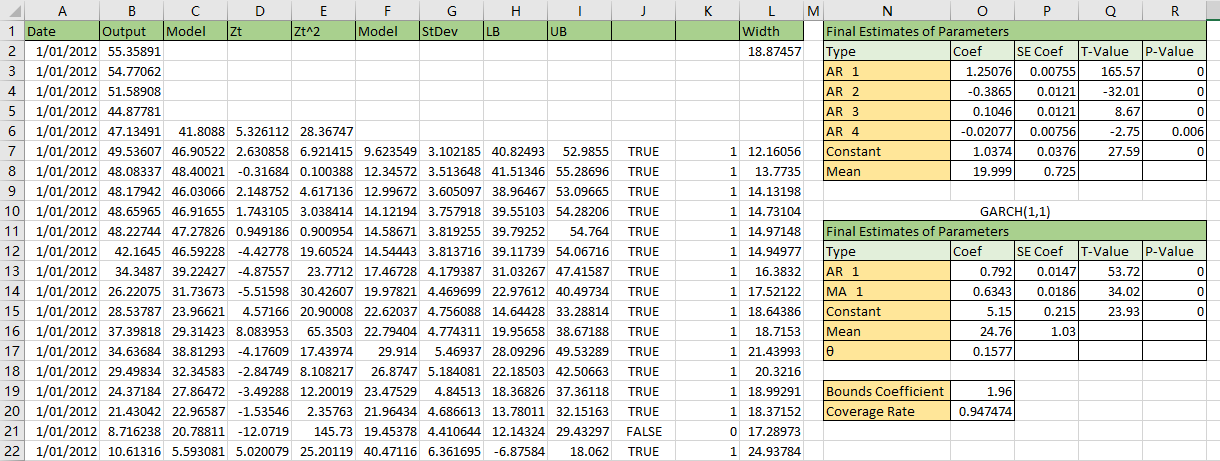
According to previous steps, we got two models for the dataset. One is AR(4), another one is GARCH(1,1). All the models will apply to 2012 dataset. The result likes the picture below.



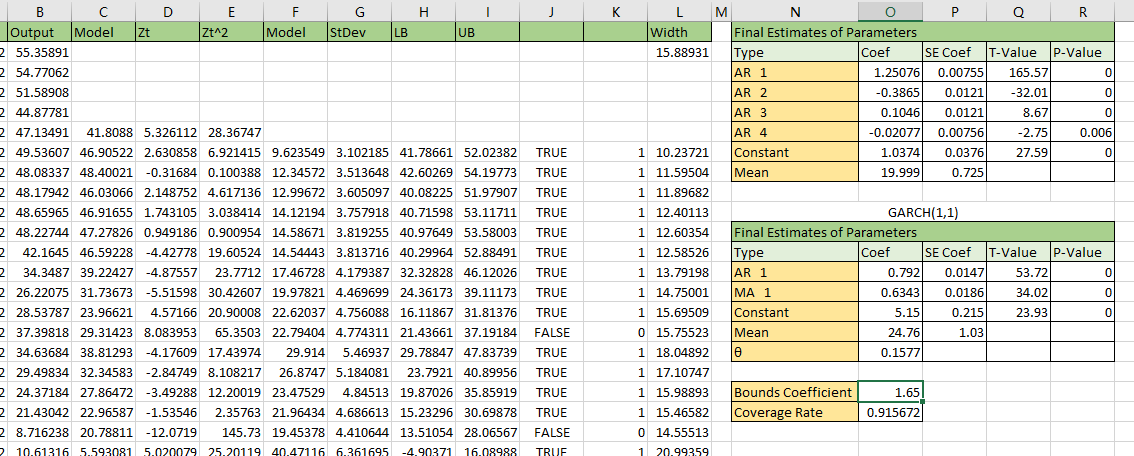
## Evaluate the performance

The score of 90% is about 1.65 and the score of 95% is about 1.96. The 95% result like the picture below.

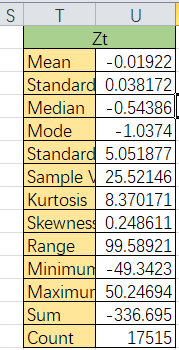
The 95% coverage result like the picture below.



The 90% coverage result like the picture below.



The statistical summary for the residuals of 2012 data likes the picture below.



According to the statistical summary, the standard deviation is about 5.05. The mean predictiion interval width of 95% coverage is about 18.87, for 90% coverage is about 15.89. The real coverage with 1.96 score is 94.75%, and 91.57% for 1.65 score. So we could get the formulas like below:

94.75% ≈ 95%

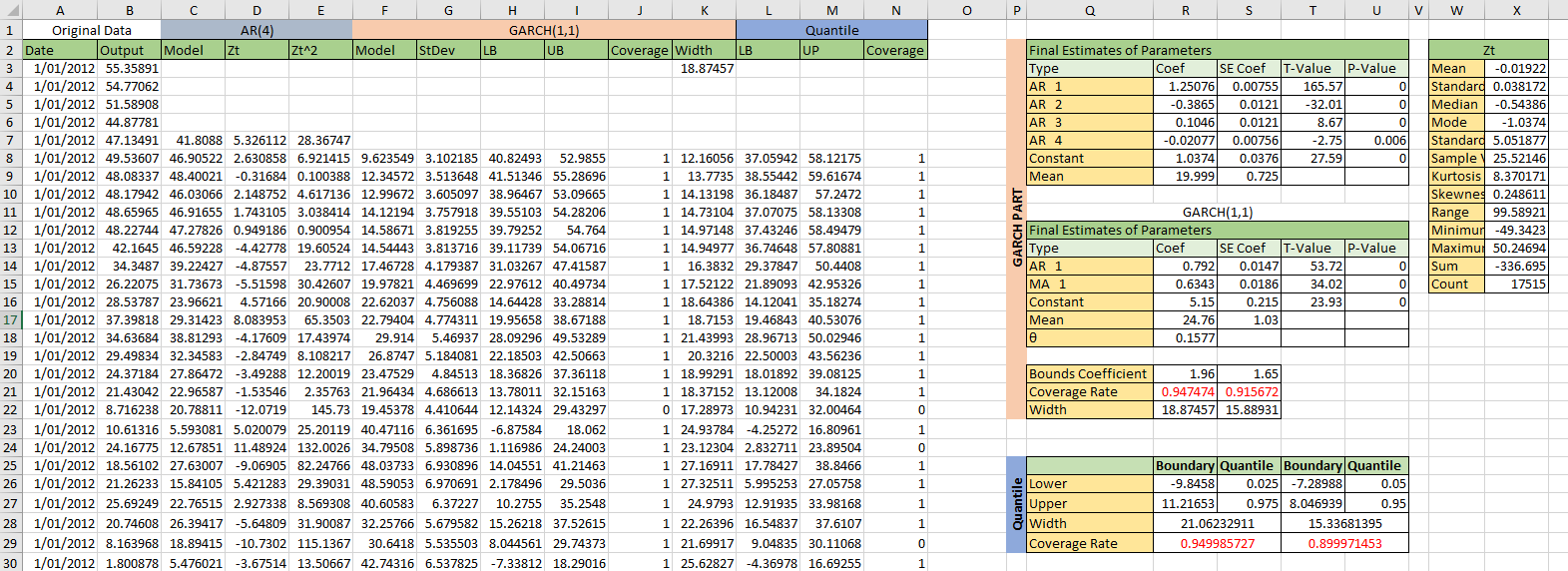
91.57% > 90%

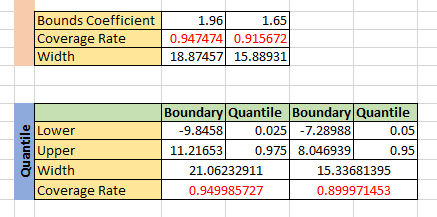
5.05\*1.96\*2=19.8>18.87

5.05\*1.65\*2=16.665>15.89

The results suggest that the model is quiet well for the dataset.

## Compare the results



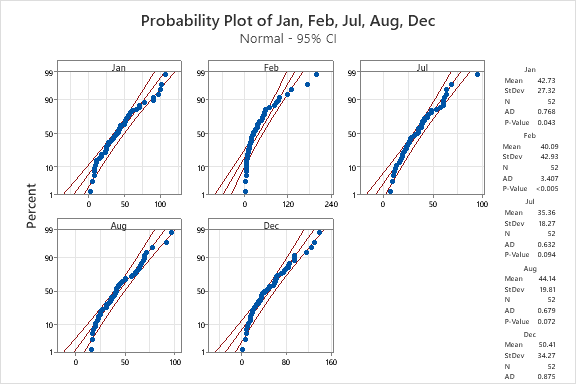
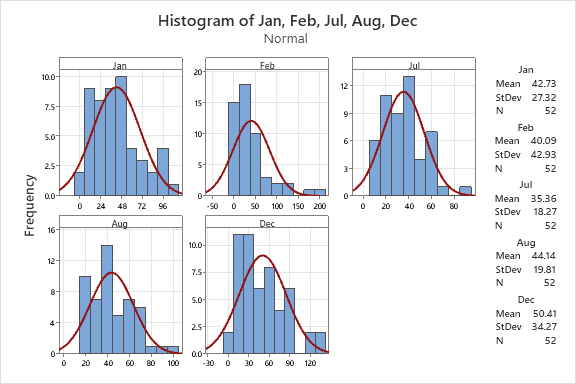


# Dataset of Melbourne Airport Rain



## Test normality for December, January, February, July, August

For normality test, the ppplot and histogram could be used for testing.

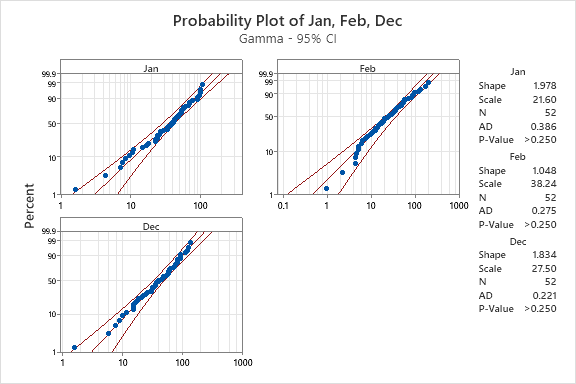


According to the histogram all the distribution of the months are right skewed. According to the ppplot result, we reject January, February and December dataset follow normal distribution because of p-value is less than 0.05.

## Test for Gamma fit.

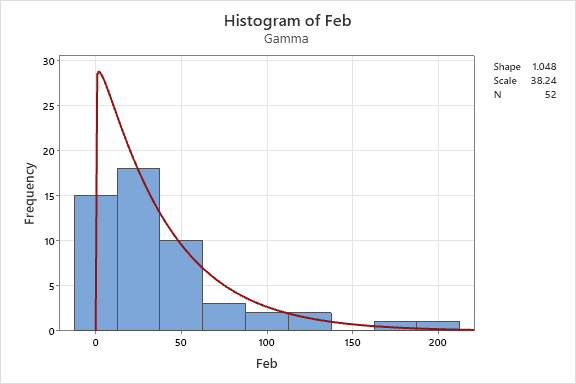
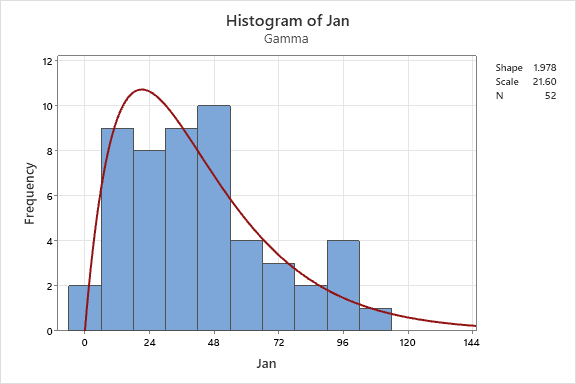
There are two steps for this question. The first step is to test the gamma distribution, and then to calculate the α and β parameters. Another step is to get the distribution and visualize them. According to previous step, the datasets of January, February and December will be processed.

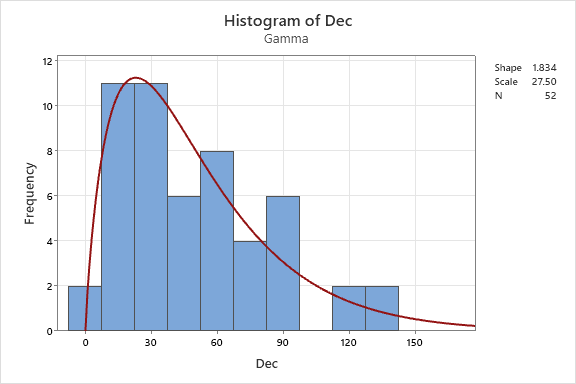
* + 1. Gamma distribution test

According to the right graph, the p-values are all greater than 0.05, it means we cannot reject the datasets follow gamma distribution. And the parameters of gamma distribution for each dataset are also listed on the graph.

* + 1. Visualize the distribution.

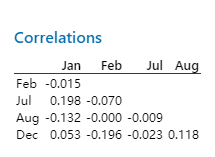
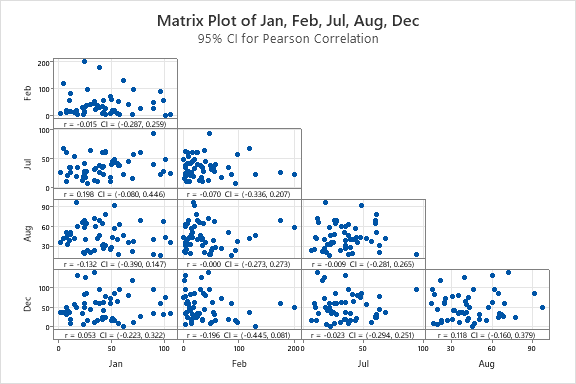
This part will show the gamma distribution for each month.





## Test the correlation.

For the correlation test, the Pearson correlation is used here. The Matrix Plot and correlation matrix like the pictures below.

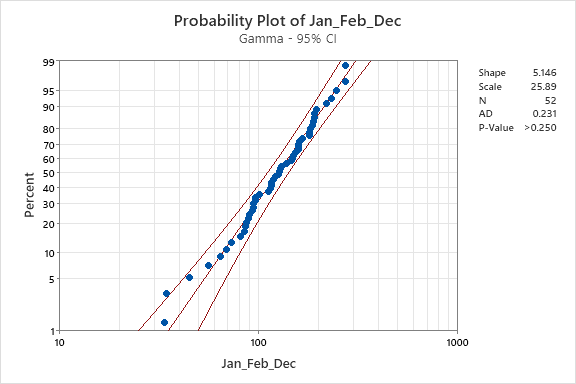


According to the correlation matrix, the data of Jan with July have the highest correlation, even the highest correlation is very week, so we regard there is no correlation among the datasets.

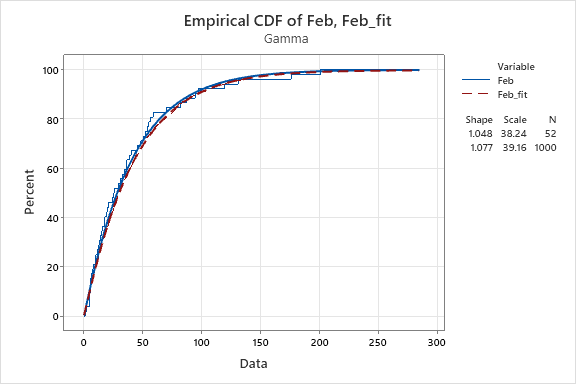
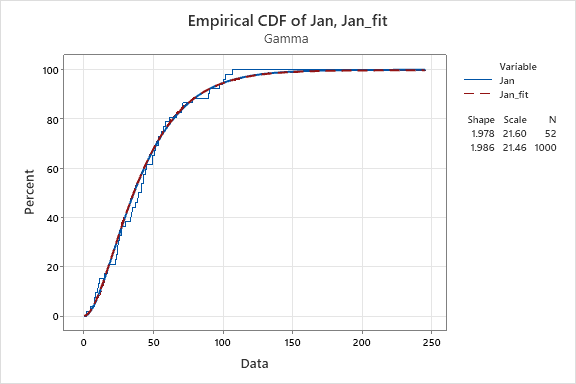
## Synthetic and empirical CDFs for January, February, December and total.

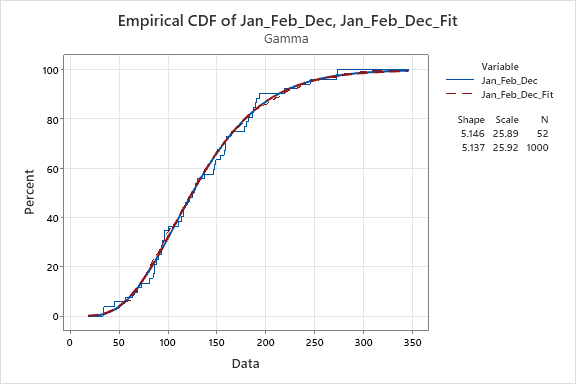
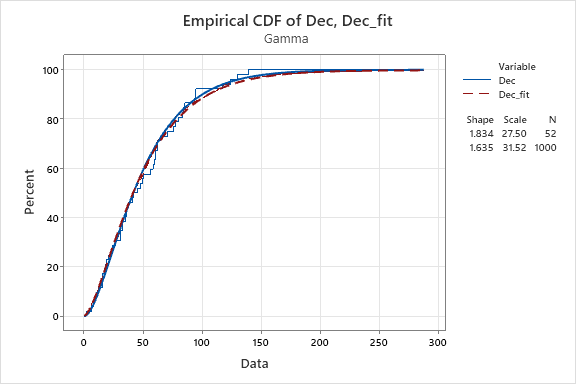
The first step is to get the total data then to test gamma distribution for it. The second step is to generate synthetic data for the datasets that follow gamma distribution, then to do empirical CDFs for them.

Make sum of January, February, and December to form the total data, then to do gamma test like the picture below.



According to this graph above, it can not to reject the total data follow gamma distribution. So, next step is to do empirical CDFs for the four datasets, details like the pictures below.

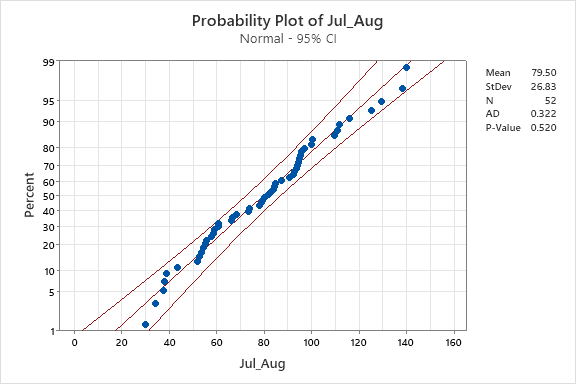




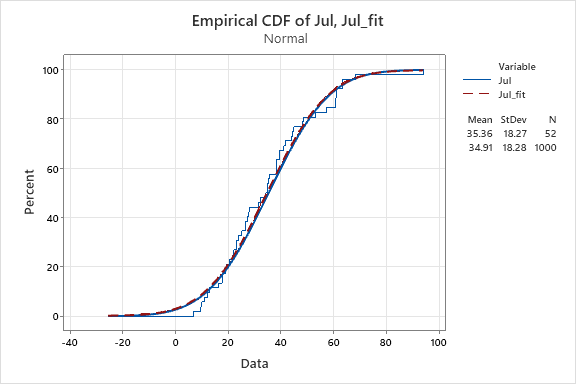
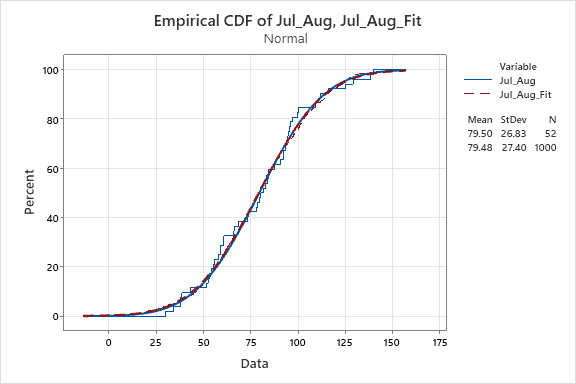
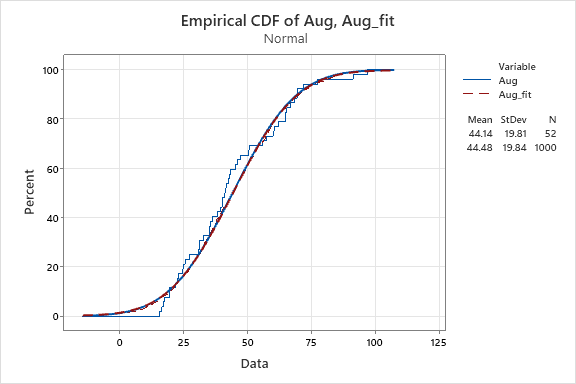
## Synthetic and CDFs for July, August.

The process for this question is the same with 2.4. The difference is the datasets of July and August follow normal distribution. When generate Synthetic data using normal distribution.

Test the total value of July and August, then to test the normality for it, details like the picture below.



According to the result above, we cannot reject the total data of July and August follows normal distribution because of the p-value is greater than 0.05. The next step is to do empirical CDFs for the three datasets, details like the pictures below.

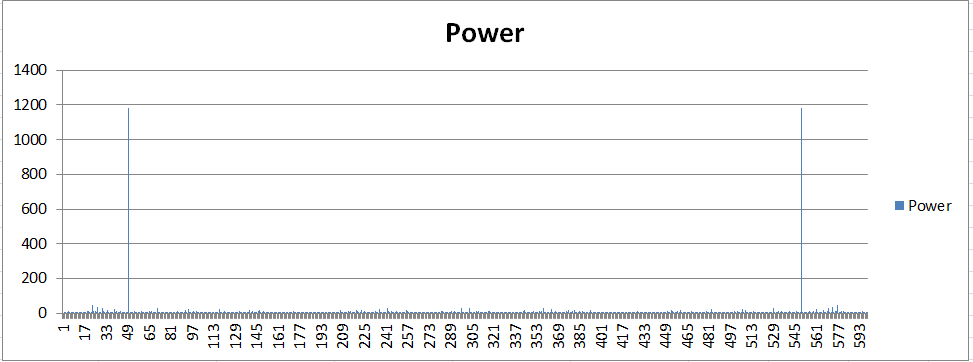
# Datasets of Mt Gambier By Months Temperature and Mt Gambier Rainfall



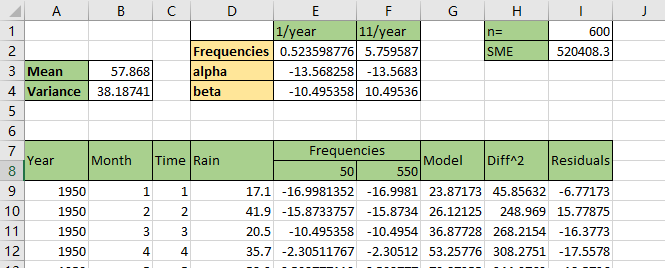
## Model the seasonality and then get residuals.

There are three steps for this question. The first step is to find the best frequencies, the second step is to find the proper parameters for seasonality, and the last step is to visualize the seasonality result.

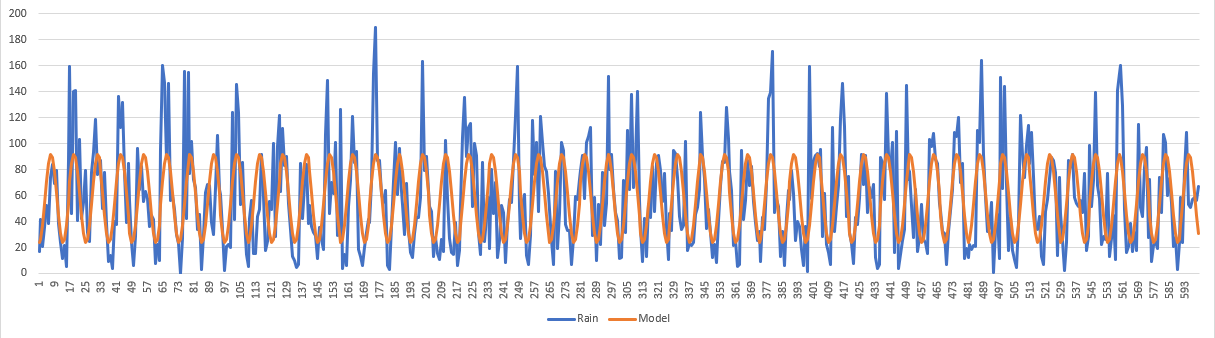
The picture below is the frequencies for the dataset. The 50 and 550 is the best for the dataset.



The picture below is the seasonality parameters using the frequencies got from step1. And also we got the final model and the residuals.



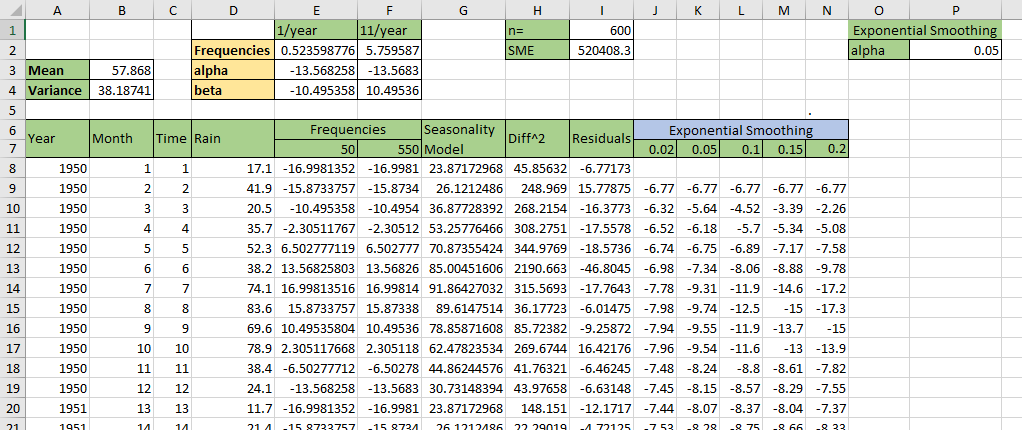
The picture below is the visualization for the final model of seasonality.



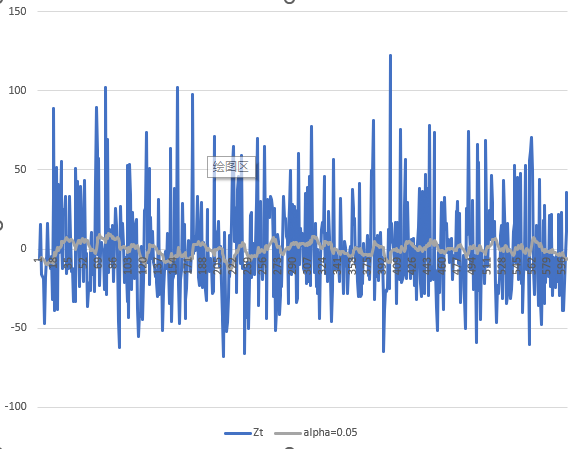
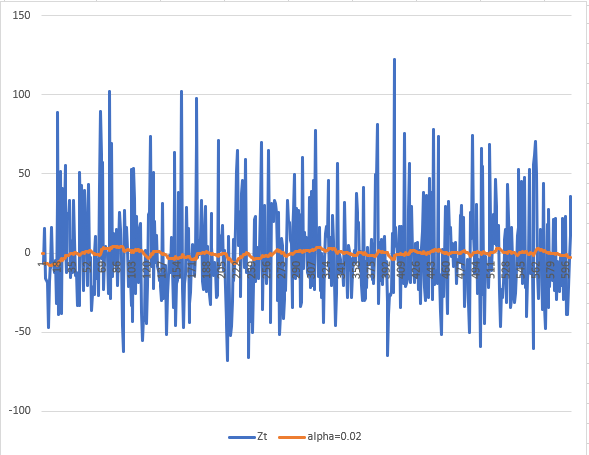
## Use exponential smoothing to see the overall trend

For this question, I will the dataset calculated using exponential smoothing with various values, then to visualize them.

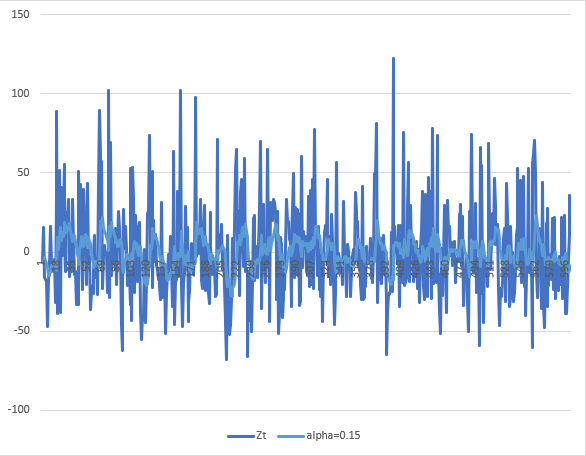
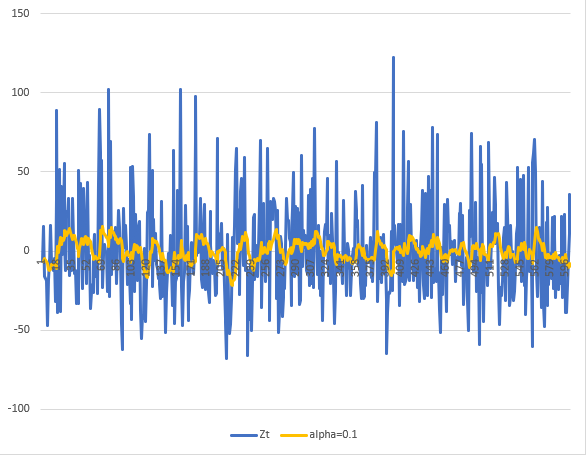
The datasets get from exponential smoothing with different values like the pictures below.



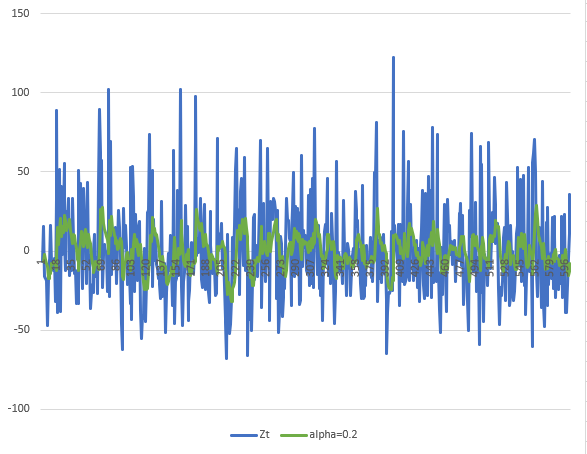
According to the picture above, five different values of alpha are used to calculate. Next step will show the overall trend based on the five parameters.



*α=0.02 α=0.05*



*α=0.1 α=0.15*



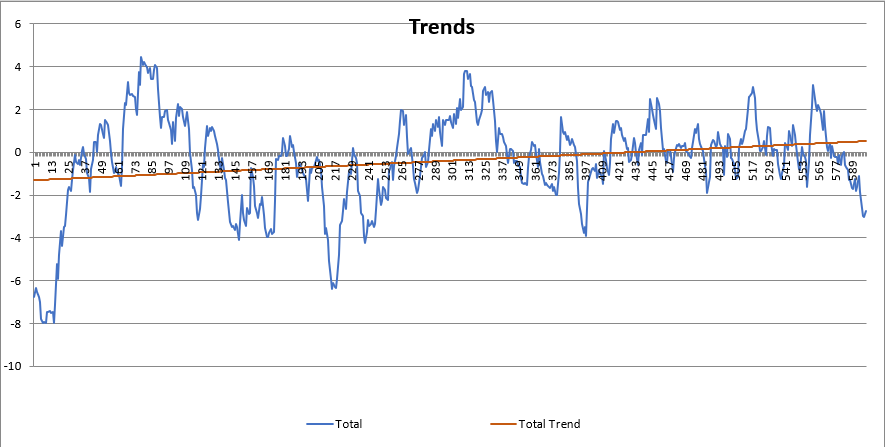
*α=0.2*

According to the graphs, it is much easier to extract the trend when the α=0.1.

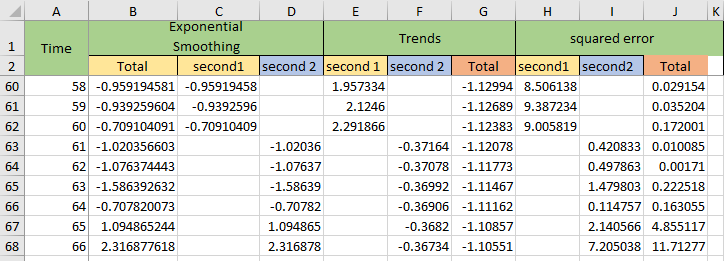
## Find the trends for the smoothed data of whole series and various sections

I will set the parameter α equals 0.02 of smoothed data, and then to process the smoothed data. There are 3 steps to do. The first step is to find the trend of the whole dataset. The second step is to split the dataset into multiple sections, and the last step is to find the trends for each section.

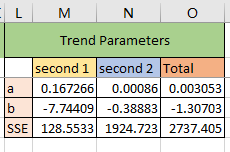
* + 1. Find the trend of whole dataset.  
       I use the univariate linear regression to model the trend. The result like the picture below.



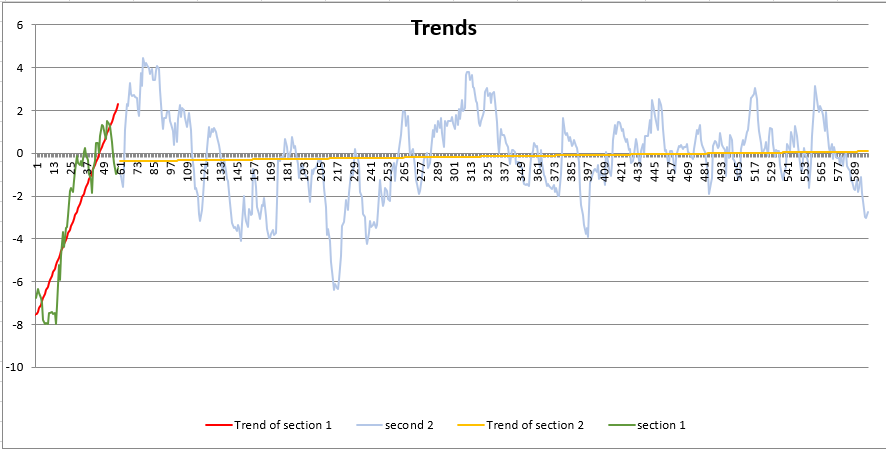
* + 1. Split whole dataset into multiple sections.  
       According the visualization of the dataset, the dataset could be split into two sections, the first section (from the beginning to 60) rise rapidly, and the second section oscillate around a variable. So the dataset could be split into two sections like the picture below.



* + 1. find the trends for the two sections  
       The trends of the two sections will use univariate linear regression to model, details like the picture below.



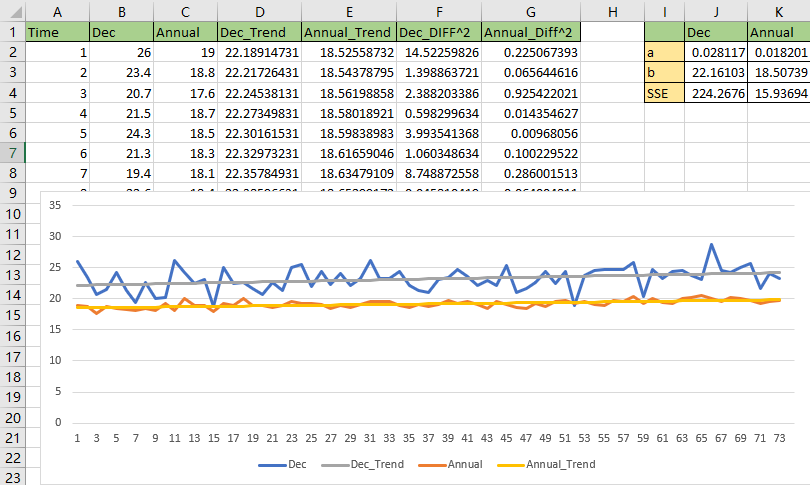
***The trends parameters for the two sections***



***The trends for the two sections***

## Take the data for the month of December and the Annual mean temperature from MtGambierByMonthsTemperature.xlsx and find the trend over time.

The univariate linear regression to model trends for the two datasets. The results like the picture below.



## How much has the mean temperature changed over time in each case?

According to the result from the last step, the mean temperature changed over time should be calculated via the linear regression parameters with the whole 73 years on this dataset.

The temperature changed over time for December should be,

a\*year = 0.028117 \* 73 = 2.05

The temperature changed over time for December should be,

a\*year = 0.018201 \* 73 = 1.32