Model Diagnostics and Predictions for Monthly Rainfall in Perth, Australia

Stats 485 Project Team 9
Breanna Blackwell (301396525)
Jason Dang (301400032)
Zhiran Tong (301390137)
Seokhyun Yoon (301349313)

Executive Summary:

Our goal for this project was to determine an appropriate time series model for precipitation in Perth, Australia, and to use this model to predict future rainfall. Our data was downloaded from Kaggle and includes the daily rainfall in Perth from the 1st of July 2008 till the 20th of July 2017. After removing missing values and taking the average rainfall per month, we were left with a data set of the average monthly rainfall in Perth spanning 8 years from July 2008 (Month 1) until June 2017 (Month 108).

We began by making scatter plots of the data and saw that there was likely a 12-month seasonal trend. This was in accordance with our assumptions about the data, theorizing that there would be more rain throughout the winter months (Jun-Aug) and less throughout the summer months (Dec-Feb). Initially, we suspected that the average rainfall every 12 months (for example, from June year x to June x+1) would decrease over time due to factors such as climate change and earth warming. However, given our limited dataset of only 8 years, we could not detect any evidence to assume stochastic or deterministic trends that would lead us to believe that this process is non-stationary. Therefore, we conducted our analysis assuming that our data for average monthly rainfall was stationary with a constant mean.

After using the Box-Jenkens method and evaluating the autocorrelation function (ACF), the partial autocorrelation function (PACF), and the extended autocorrelation function (EACF), the plots supported our conclusion that the data contained seasonality. This was illustrated by the alternating correlations in the ACF model with negative correlations at lag=5-6 and positive correlations at lag=11-12. The Box-Jenkens method concluded that we should consider the ARMA(3,2) and the ARMA(2,1) models. These models were identified as suitable candidates for capturing the observed seasonality in the time series data.

After applying the dynamic method and analyzing the residuals, we found that all ARMA(2,1) parameters were significant, but its residuals appeared to be correlated. We then added a seasonal component to ARMA(2,1), and we discovered that the parameters of ARMA(2,1)x$(1,0)_{12}$ were all significant, and the predicted residuals were very close to white noise. As a result, we concluded that ARMA(2,1)x$(1,0)_{12}$ was an appropriate model to use for the Perth precipitation data.

Using our chosen seasonal ARMA model, we predicted the estimated monthly rainfall for Perth 12 months from now and found a resulting 95% prediction limit which is a crucial metric for forecasting reliability. Our model, ARMA(2,1)x$(1,0)_{12}$, accommodated the observed seasonality and demonstrated robustness, aligning with the white noise characteristics in its residuals. Additionally, from detecting seasonality with Box-Jenkins to refining with the dynamic method, converged on ARMA(2,1)x$(1,0)_{12}$ as the optimal model. This choice captured the nuanced patterns in Perth's rainfall data, and guided us to relatively reliable predictions for the future.

Introduction:

Our motivation for this analysis was stemmed by the increasing news headlines of wildfires in Australia over the years. As a response, we wanted to develop a robust precipitation forecasting model with a horizon of 12 months. This report will examine modeling the precipitation levels in Perth, Australia. Our goal was to compare the performances of an ARMA(p,q) model and a seasonal ARIMA model and determine which of these models would be better for forecasting the precipitation levels in Perth, Australia.

Dataset:

The source of the precipitation dataset (mm) used in this project was drawn from numerous weather stations available on the Australian Government of Meteorology website. However, for the purpose of this analysis, a more cleaned and preprocessed version of the data was used from Kaggle. The data collected from Kaggle spanned over 8 years, ranging from 2008 to 2017.

The Australian precipitation dataset contained approximately 2.4% missing values. Filtering our dataset to a more specific and well-known location, Perth, solved the majority of our issue of missing values. However, we noticed that April 2011, February 2012, and December 2012 were still missing. As a result, we manually entered the monthly averages for these months with the figures found on the Australian Government Bureau of Meteorology website. Considering the class's scope, we found this fix adequate as the additional data came from the same source as our original dataset.

As we analyzed the daily precipitation graph, we noticed our graph did not appear like the example of the annual rainfall in LA on our first day of class. Therefore, we considered aggregating our data. However, due to our limited data of 8 years, we chose to aggregate our data monthly rather than yearly. As a result, this provided us with 108 observations.

Analysis:

Figure 1.1 appeared to exhibit strong seasonality. This was characterized by its recurring negative correlation every 6 months and positive correlation every 12 months. However, after analyzing the 12-month periods more carefully, there seemed to be a slight downward trend where the precipitation level was suggested to decrease over time. This seemed to align with the recent news headlines concerning Australian wildfires. However, we did not believe this trend would continue forever. We thought that this downward trend may be due to the limited 8-year time frame this dataset was sampled. Therefore, given the inclusion of a more extensive dataset sampled over 100 years, we believed the precipitation patterns would be stationary with a constant mean. As a result, we continued our analysis under this assumption.

Box Jenkins Method:

The ACF plot of our data shown in Figure 1.2 showed a dampened sine-wave pattern, a common property of an AR(2) model. Additionally, it's possible there was an ARMA(2,1) model with complex roots. We could see a recurring pattern every 12 months, suggesting we could also have a seasonal model with a period of s=12. The ACF plot with a lag of 108 (Figure 1.1) also looked like it was exponentially decaying every 12 lags, which could indicate that the model is an AR(1). However, we did see a slight increase from lag=1 to around lag=10/lag=12 before the model decreases, indicating that it could be an AR(2) or higher-order model.

If we consider the MA(q) component, we could see that there was still a high correlation up to lag=70, which could indicate seasonality. Seasonality was also corroborated by the depiction of alternating positive and negative correlations in the ACF plot.
The PCAF plot (Figure 1.3) showed that there were significant terms at lag=1, lag=3, lag=5 and lag=7 and possibly lag=11. The spikes at lag=3, lag=5, and lag=7 indicated that some randomness was happening, which was not inconsistent with a seasonal model. Additionally, we could see a negative correlation between every lag=5-6 and a positive correlation and lag=11-12, which supported our earlier claim for the ACF model. Both the ACF plot (Figure 1.2) and the PACF plot (Figure 1.3) suggested that our model cannot be explained with either an AR(p) model or a MA(q) model. Thus, we considered an ARMA(p,q).

Looking for our leading 0 in Figure 1.4, we could see that ARMA(3,2) would be adequate for explaining our model, considering there are only 3 significant values (depicted by x's) in the triangle. However, we could also consider a simpler model, such as ARMA(2,1). Upon evaluating these models, we were confronted with the decision between ARMA(3,2) and ARMA(2,1) structures. The ARMA(3,2) model had two additional parameters compared to the ARMA(2,1) model. However, the increase in log-likelihood and the decrease in AIC did not justify the complexity introduced by the extra parameters based on the result above. Additionally, adhering to the principle of parsimony and considering a simpler model, ARMA(2,1) contained 6 violations, possibly due to randomness, or their value may be just outside of the confidence interval. Therefore, we pursued the ARMA(2,1) model.

Dynamics Method:

Considering our data's nature, we opted to include a seasonal component for testing in the dynamics method. Because we have a short duration of the rainfall data, 8 years from 2008 to 2017, we could not find any evidence to prove our data to be non-stationary. Therefore, we have set the order of difference, d, to be zero for both the non-seasonal and seasonal parts.

We started testing the simplest ARMA model, ARMA(1,0). After fitting ARMA(1,0) to our data, the AR(1) term appeared to be significant. Additionally, the residuals were shown to be correlated, with high values hanging together and low values hanging together (Figure 2.1). The 2th, 5th, 7th, 11th, 12th, and 19th lag residuals were all significant. This has shown us

evidence to test a higher-order model. We therefore tested ARMA(2,1). All parameters were significant for ARMA(2,1), but the residuals still appeared to be correlated, with the 5th, 6th, 7th, 11th, 12th, 13th, and 19th lags being significant (Figure 2.2). We then tested ARMA(3,2). After fitting ARMA(3,2), the AR(3) term was insignificant, with an s.e. of 0.10 and an estimate of 0.05. Therefore, ARMA(2,1) were the appropriate parameters for the non-seasonal part. We hoped that after adding a seasonal component, the residuals would become close to un-correlated.

We then continued with our tests for the seasonal component. We started with ARMA(2,1)x$(1,0)_{12}$. It was shown that all parameters were significant, and the residuals appeared to be very close to white noise, with lag 7 and lag 11 being slightly over the 2 standard deviations; however, this could be due to randomness (Figure 2.3). The residuals vs time plot (Figure 2.3) showed no systematic trend, and the histogram of the residuals is acceptable (Figure 2.3), with approximately symmetric and mean centered roughly around zero. We then proceeded to test ARMA(2,1)x$(2,1)_{12}$. The residuals after fitting ARMA(2,1)x$(2,1)_{12}$ appeared to be white noise (Figure 2.4); however, the AR(2) term of the seasonal component was not significant, with an estimate of 0.12 and an s.e. of 0.15. Therefore, ARMA(2,1)x$(1,0)_{12}$ was the parsimonious model that satisfied the dynamics method criteria.

Forecasting:

Building on the comprehensive analysis provided by the Box Jenkins and Dynamics Method, we used the ARMA(2,1)x$(1,0)_{12}$ to forecast the next 12-month period. This was shown in figure 3.0. Figure 3.0 extended our historical data trend into 2018, with the shaded confidence intervals providing an estimate of our forecast's accuracy. These intervals were important as the error margins appear quite large. This suggested that the model was actually not too effective at estimating future values. However, it was important to note that the margin of errors included negative values, which was not possible. Therefore, the forecast's margins of error should be smaller than they actually are. However, the margins of error were still quite large.

Unsurprisingly, the downward trend in precipitation levels continued into our 12-month forecast. This compelling narrative resonated with the arising concerns of wildfire risks. However, we again approached this trend with a degree of caution, mindful of the limited 8-year dataset that may not capture the full scope of the variability in precipitation levels in Perth. Furthermore, we could see that the average monthly precipitation reverted back to its mean.

Validation of Forecast with Actual Data:

To ensure the validity of our forecasting model, we sourced actual rainfall measurements from the Perth Airport weather station. However, it is important to note that our initial dataset

sampled its observations from numerous stations across Perth. Therefore, while this station may provide a reasonable benchmark, the data from Perth Airport was only an approximate gauge of accuracy and may not fully capture everything in our monthly aggregated dataset. This could be seen in figure 3.2 as although the red line, the Perth Airport weather station dataset, closely followed our actual dataset prior to our forecast there are still some notable deviations.

As we compared the forecast coming from our ARMA(2,1)x$(1,0)_{12}$ model to the actual measurements from the Perth Airport station, figure 3.2 showed that our forecast made a relatively decent prediction for rainfall in 2017/2018. Although our margins of error were quite large, our prediction remained to be prominently inside this margin of error and followed the seasonality of our data. In addition, it appeared that only a few of the months were over and underestimated in contrast to the remaining months that our model forecasts. However, this was again unsurprising to us due to the large margins of error we obtained when we first forecasted the model.

Conclusion:

Going through a vigorous procedure following the Box-Jenkins and Dynamics method we concluded that ARMA(2,1)x$(1,0)_{12}$ was the appropriate model to use for the Perth precipitation data. After we compared our prediction with the validating data, we discovered that the validating data predominantly lied within the 95% confidence interval of our prediction (Figure 3.2). Our prediction also successfully modeled the seasonality that the validating data presented (Figure 3.2). This validation enhanced our confidence in the model's ability to predict future rainfall levels. As a result, we would recommend using an ARMA(2,1)x$(1,0)_{12}$ for modeling and forecasting.

Appendix :

Figure 1.1

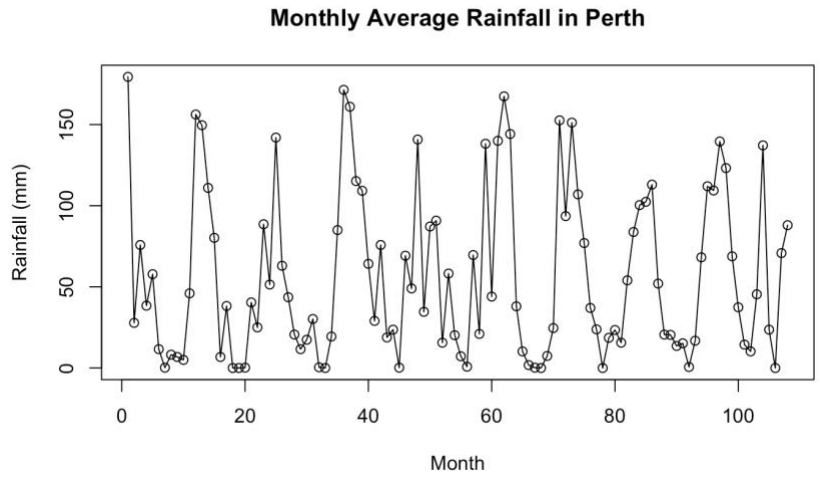


Figure 1.2

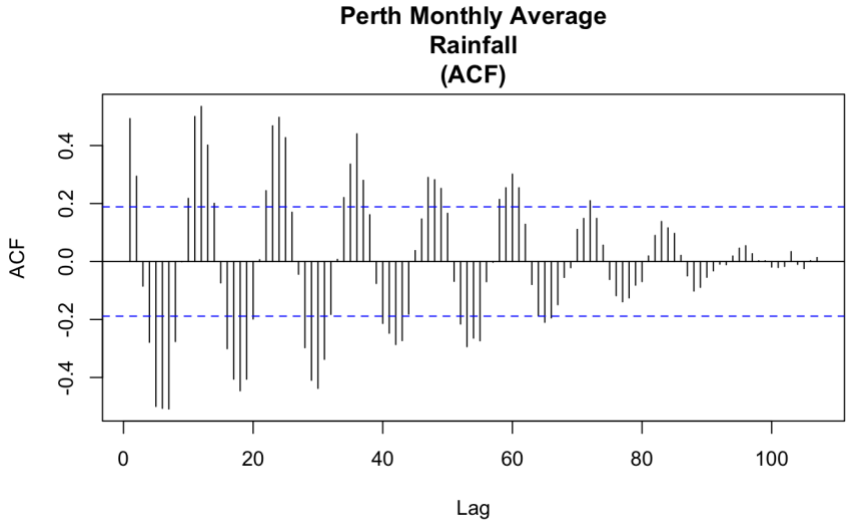


Figure 1.3

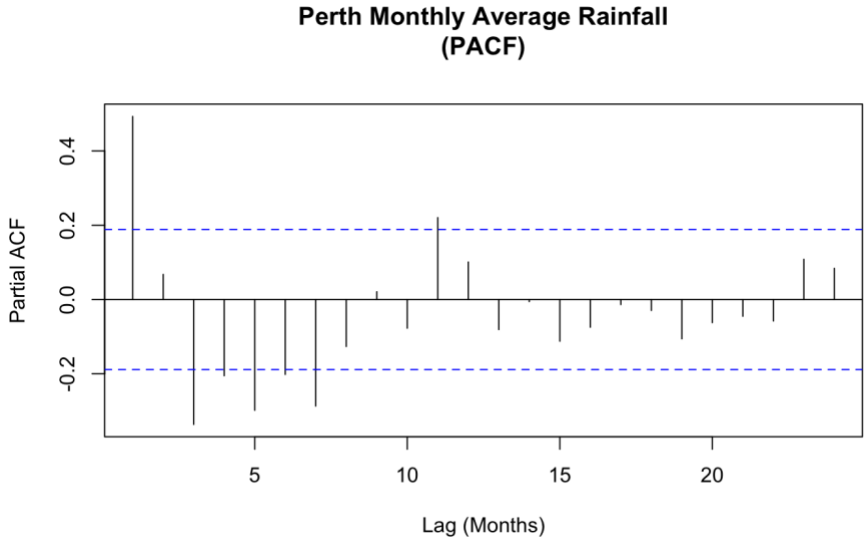


Figure 1.4

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o x x x x x o x x  x  x  o
1 o x o o o o x x o o x  x  o  x
2 o x o o x o x o o o o  o  o  o
3 x x o o o o x o o o o  o  o  o
4 x x o o o o o o o o o  o  o  o
5 x x o o o x o o o o o  o  o  o
6 x x o o o x o o o o o  o  o  o
7 x x x x o o o o o o o  o  o  o
```

Figure 2.1:

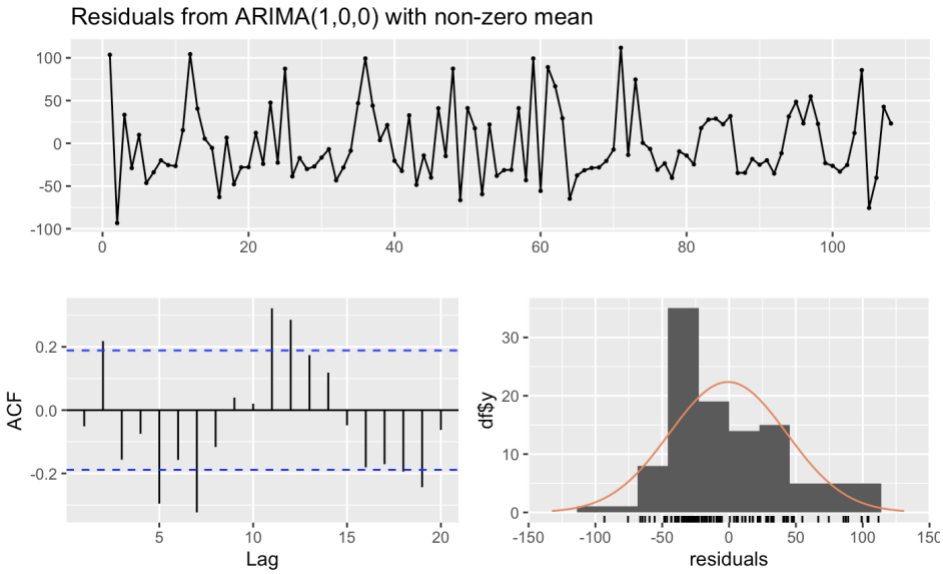


Figure 2.2:

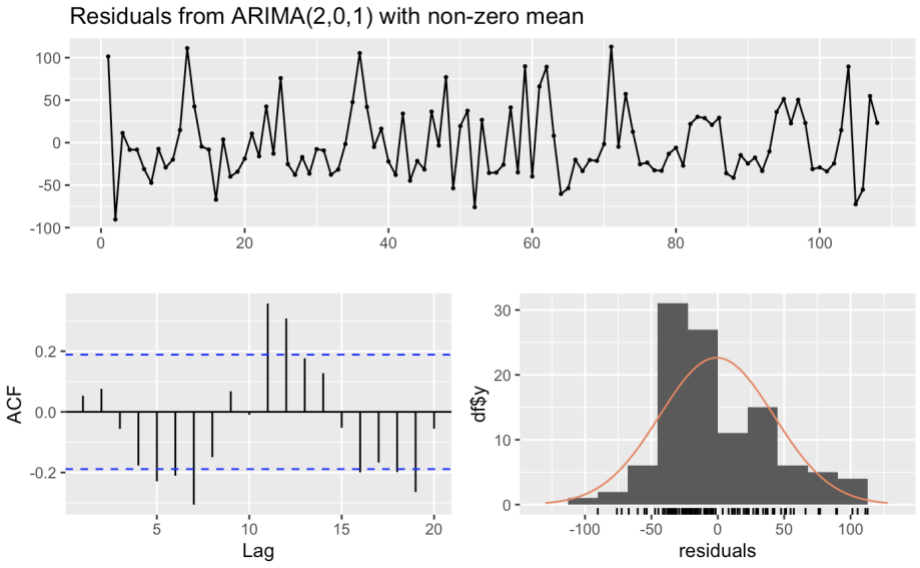


Figure 2.3:

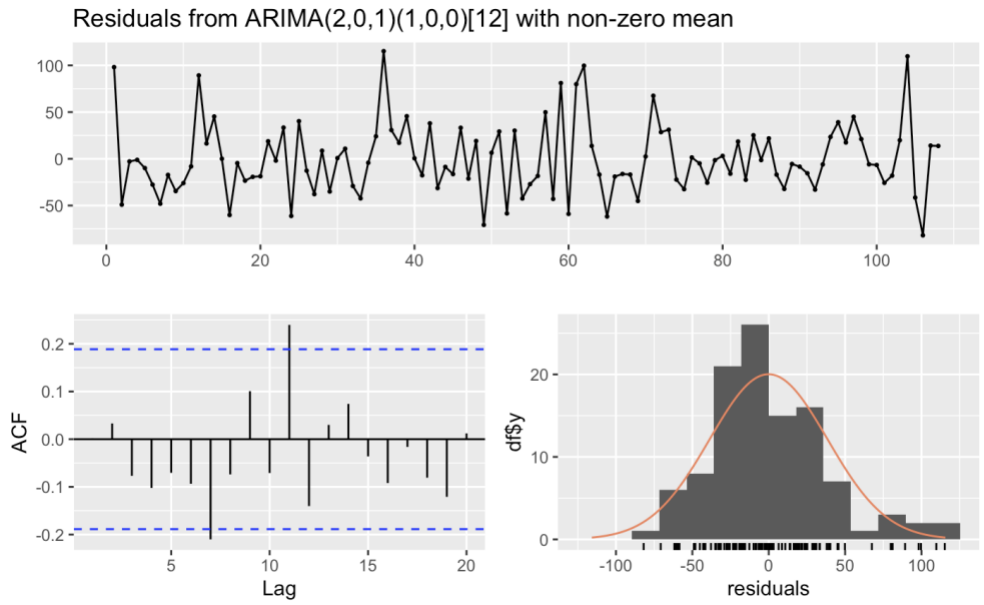


Figure 2.4:

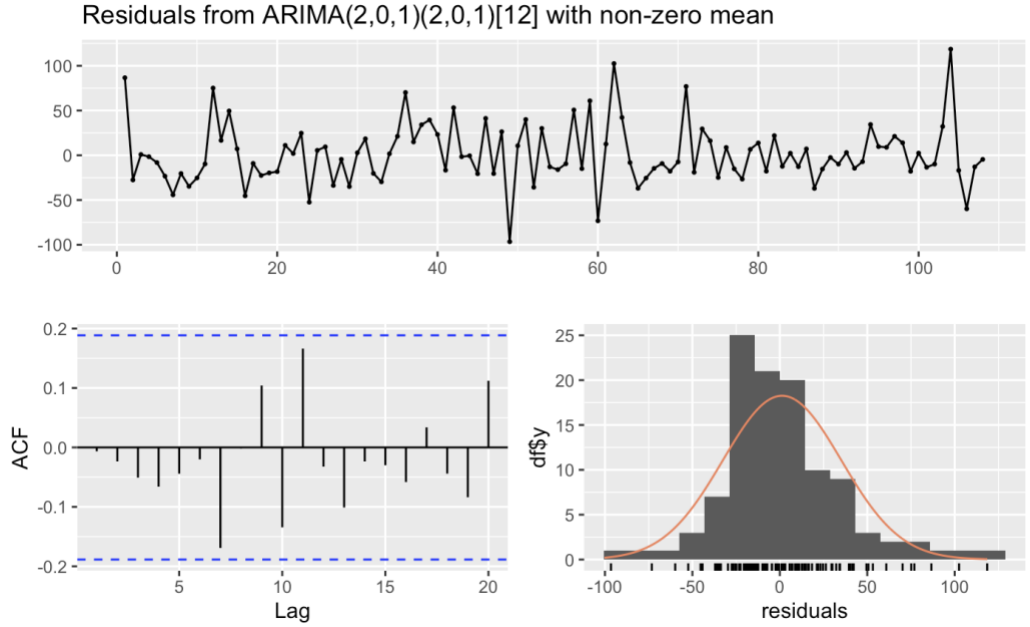


Figure 3.0:

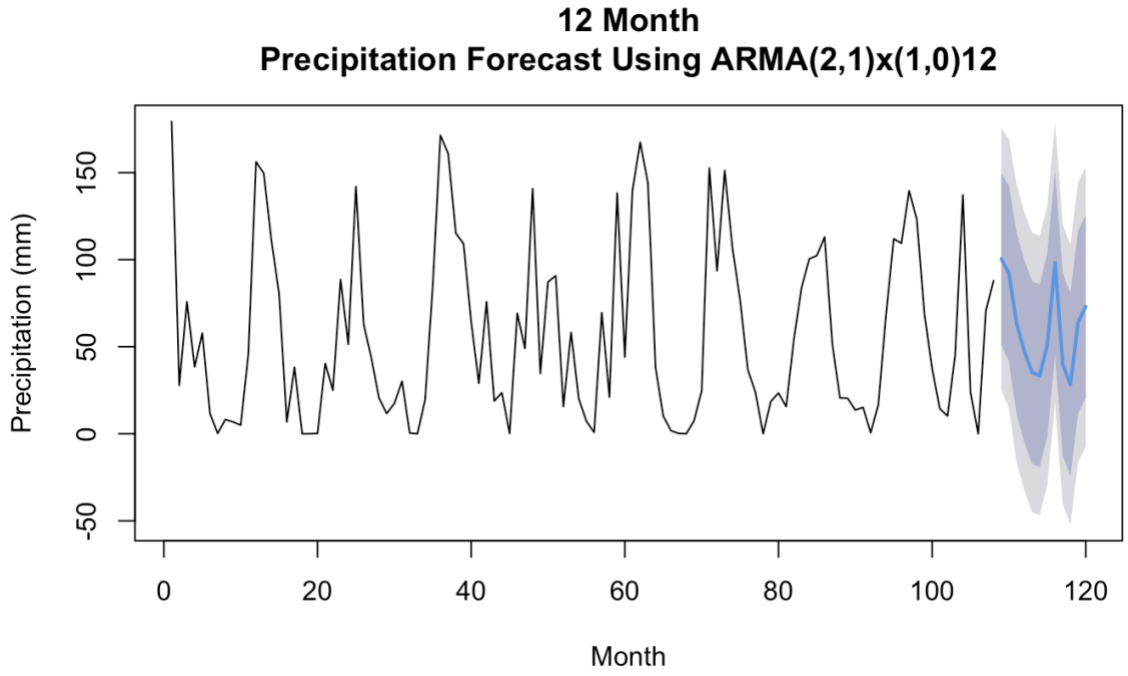


Figure 3.1:

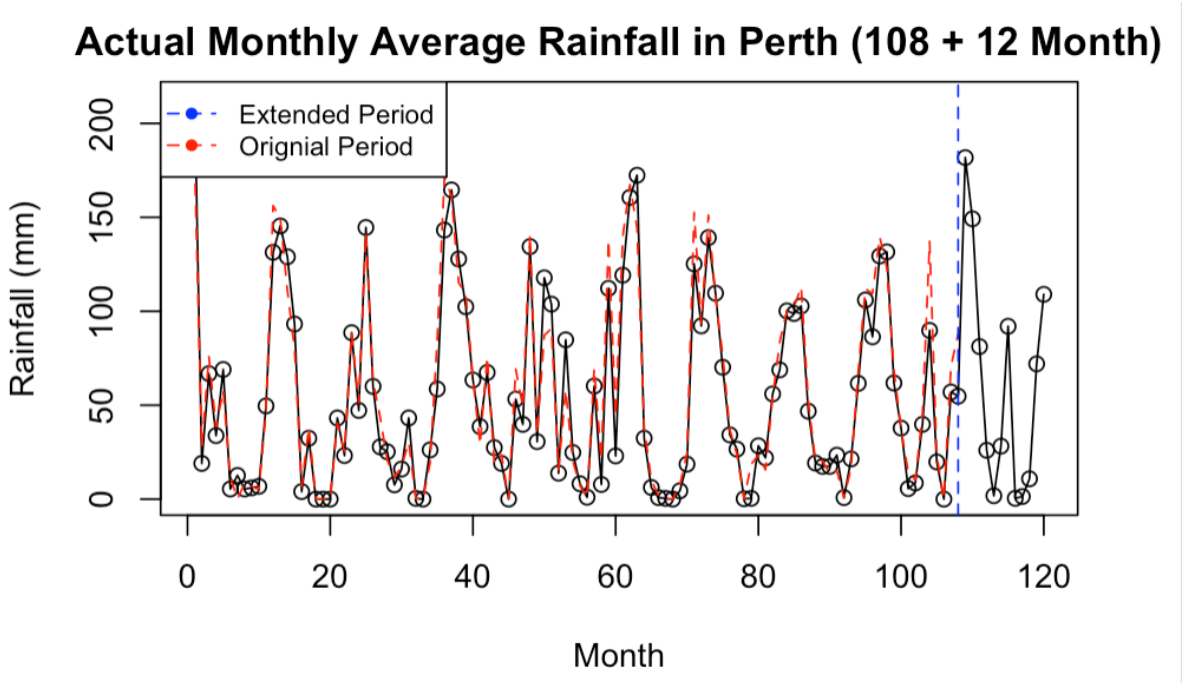
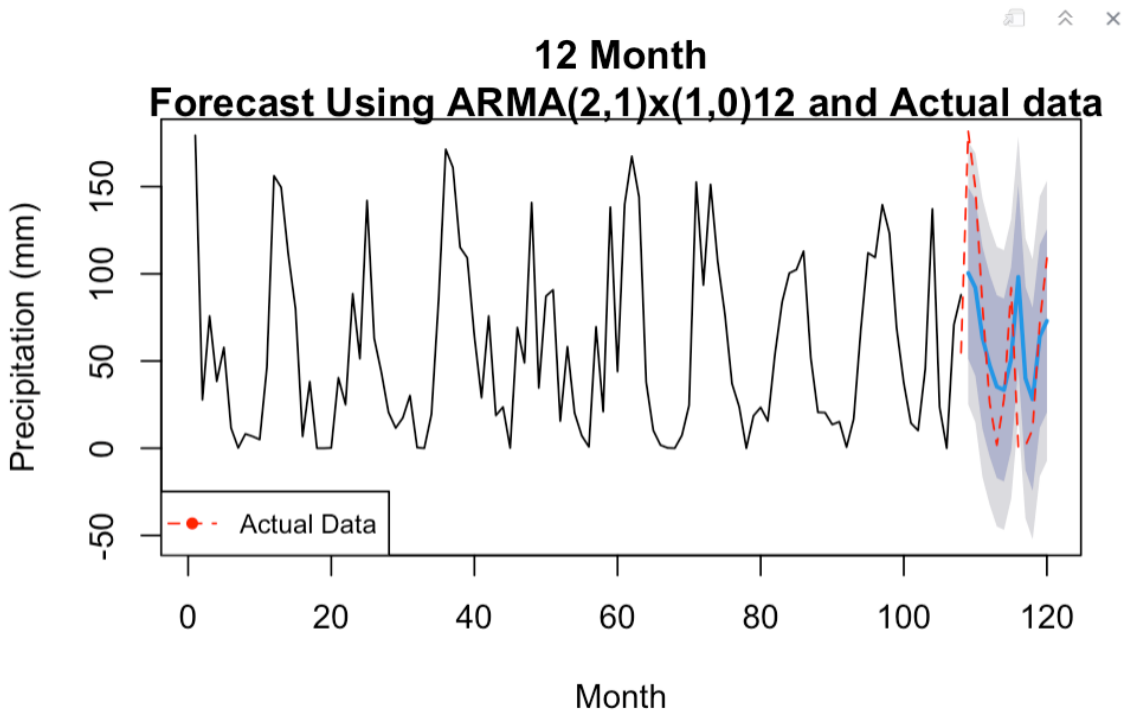


Figure 3.2:


12 Month
Forecast Using ARMA(2,1)x(1,0)12 and Actual data
Precipitation (mm)
-50
0
50
100
150
Actual Data
0
20
40
60
80
100
120
Month