Forecasting Earth's average temperature using Berkeley earth data

Dhanalakshmi Naik dn2952@rit.edu
College of Computing and Information Sciences
Rochester Institute of Technology
98 Lomb Memorial Drive
Rochester, NY 14623, USA

Abstract:
Accurate analysis and prediction of weather and climate is exceptionally challenging due to the higher order and often complex interactions between the many erratic variables that influence everyday climate. Daily and weekly weather forecasting is done using real-time observations combined with knowledge of spatial trends and patterns. Daily weather prediction algorithms yield short-term predictions with fairly accurate results. However, these become less accurate over a longer time horizon. The motivation for this research stems from this and attempts at providing a suitable forecast to predict long term trends.

To accurately predict spatial and temporal climate patterns over longer prediction windows, this research employs time-series analysis to define conditions and predict averaged temperature on the Earth's surface for the next 10 years. The forecasting techniques employed in this report are ARIMA, Holt Winter and Neural Networks. Results from each technique are presented and predictions between 2016 and 2026 are shown.

This study concludes that the average temperature is on an upward trend, 0.2^{O}/decade and resonates the leading opinion amongst the scientific community. Comparative studies show similar results (Hansen.J).

Keywords*: ARIMA model, Holt Winters Forecasting, Neural Network Forecasting, Ljung's Box test, BIC*

1. Introduction:

Weather forecasts are made usually a few days at a time using data collected from weather satellites, weather stations, and other land/sea based streams. The chaotic and highly complex interaction of the weather system makes weather forecasting inherently uncertain. Given the chaotic nature of the atmosphere, there is limit to accurately predicting weather within reasonable accuracy. The limit as identified from observation is two weeks (Lorenz).

One may then question the accuracy of climate prediction, given that weather is only predictable for about 2 weeks. The answer lies in how "Climate" is defined. It is defined as the prevailing weather conditions over a long period. In other words, it is an averaged statistical representation of weather conditions. The strongest characterizing parameters of climate are averaged temperature and precipitation (National Research Council. [NRC]). This study focuses on the analysis and forecasting of the former, i.e. averaged earth temperatures for the coming decade.

Along with forecasting yearly averaged temperature changes, the report also predicts the change in variance of these predictions. Such a presentation of the results would indicate the extremes of conditions that one could expect and would also indicate the prediction confidence intervals.

Accurate climate forecasting has a profound social impact and utility. Knowledge of accurate forecasting data helps plan key infrastructure, contingency and development activities to minimize human's negative impact on the climate. Developing countries can utilize this vital information in a myriad of ways, viz. drive key energy policies and better manage environmental resources, all of which are key in promoting socio-economic progress.

This paragraph depicts the outline for the rest of this report. **Section 2**(Data Set and Methodology) describes the data set and outlines the dataset preprocessing technique employed; **Section 3**(Forecasting using Time Series Analysis) employs the methods described in Sec2 and presents the time-series analysis of the data; **Section 4**(Computational Results) presents the results of the various analysis models used; **Section 5**(Conclusion) summarizes the results and presents the future work.

2. Data Set and Methodology:

The Berkeley Earth data (Standford Solar Centre) provides a tabulated dataset of the earth's weather observations from the year 1753A.D. till present. The included data is evenly disturbed, sparse and consists of many attributes (depicted in Table 1).

Attribute Name	Description
DateRange	Start: 01/01/1753 End: 12/1/2015
LandAverageTemperature	Global Avg. Max. land temperature in Celsius
LandAverageTemperatureUncertainty	95% confidence around the average
LandMaxTemperature	Global Average Max. land temperature in Celsius
LandMaxTemperatureUncertainty	95% confidence interval around Max. land temperature
LandMinTemperature	Global average minimum land temperature in Celsius
LandMinTemperatureUncertainity	95% confidence interval around the Min. land temperature
LandAndOceanAverageTemperature	Global average land and ocean temperature in celsius
LandAndOceanAverageTemperature Uncertainty	95% confidence interval around global average land and ocean temperature.

Table 1: List of Berkeley earth data (http://berkeleyearth.lbl.gov/auto/Global/Land_and_Ocean_complete.txt)] attributes and their description

While the historical data is present from 1753A.D., we limit our analysis for years following 1904A.D. owing to the better confidence margins in the collected and tabulated data. Decreased uncertainty and increased confidence intervals could be attributed to better measuring techniques, standardized processes and better sensing capabilities.

Even given the reduced date range, the study found the data set requiring imputations for smoothing out uneven or missing temporal entries. The missing values were handled using central imputation methods that replace missing data with estimated values.
In order to further reduce risk of accidental introduction of biases during the imputation, the dataset was transformed to from a monthly to a yearly interval. This transformation was done through a weighted averaging of monthly global temperature.

A preliminary analysis of the relationships of these attributes was plotted (See Figure 1). It is important to note here that averaged temperature considers the land temperatures.

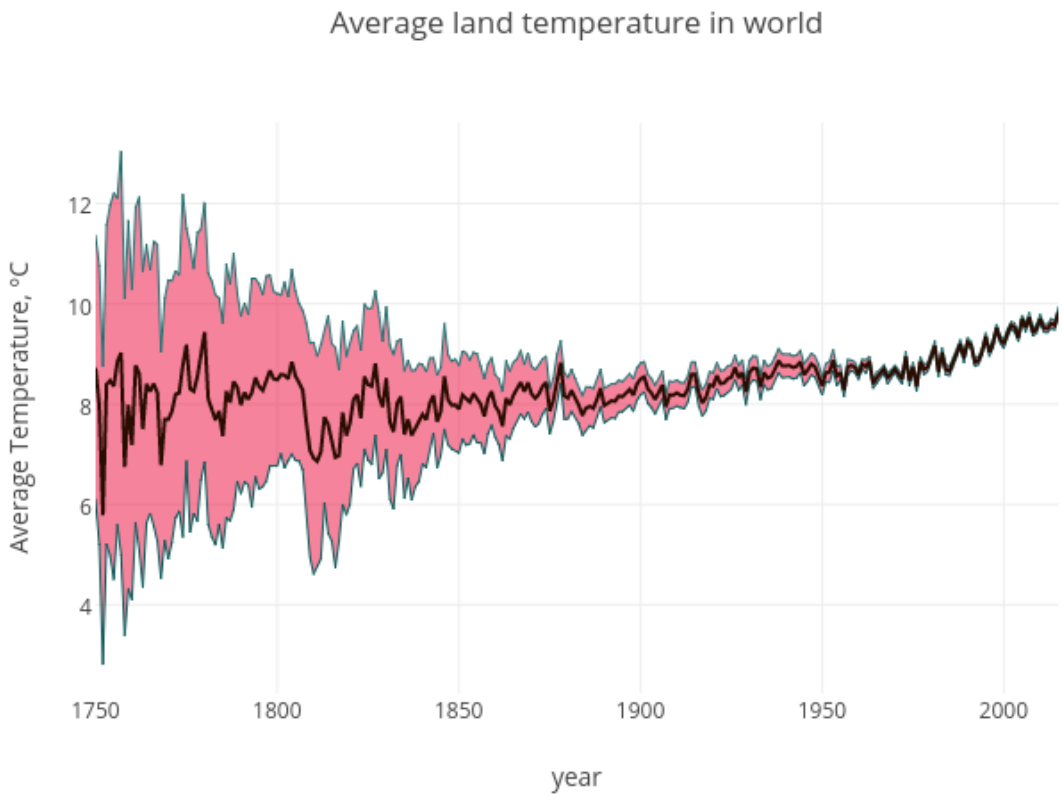


Figure 1: Yearly averaged global average mean temperatures using Berkeley Earth Data (http://berkeleyearth.lbl.gov/auto/Global/Land_and_Ocean_complete.txt) Plotted using plot.ly

It can be observed from Fig (1) that there is a steady increase in the average land temperature through the past century. Also, the uncertainty band decreases towards the later part of the data set. As stated before, this is due to the higher observation accuracies that resulted from better observation sources such as weather satellites and the like.

The data tabulation, pre-processing to avoid missing data-points, algorithm implementation and subsequent analysis is done using the 'R' v3.3.2 coding platform. The details of the time series analysis and forecasting models are discussed in the next section.

3. Forecasting using Time Series Analysis:

The data obtained after data cleaning is converted to time series format for further analysis. The dimension of the data that is being analyzed is 112, 2, which means that there are112 rows and 2 columns considered out of which one being the date column. Like any other time-series, the first step of understanding the given time series is by plotting a time series graph which will give preliminary information about the underlying tread and seasonality. The obtained as a preliminary analysis is shown below (See Figure 2).

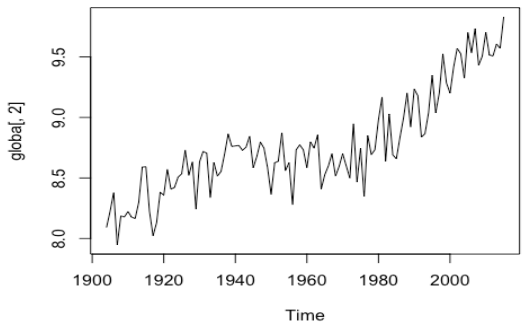


Figure 2: Time Series plot from 1904-2015

It is seen that there is an upward trend in the earth's average temperature for the past century. There is no seasonality in the given data as interpreted from Fig (2). To be confident about seasonal modes not being applied to the models being build a test for seasonality was carried out which resulted in 'FASLE' value. It was concluded that there was no seasonality in the data being analyzed and no seasonal models were applied for the analysis.

Next step in the analysis is analysis of auto correlation function(ACF). This is used mainly in time series analysis to find patterns in the data. Specifically, ACF tells the correlation between points separated by various time lags. The ACF and its sister function Partial Auto Covariance function are used in the Box-Jenkins/ARIMA modeling approach to determine how past and future data points are related in a time series.

From Fig (3) it can be interpreted that the ACF function is decaying slowing staying well above the significant line. That says that the time series is a non-stationary times series. The non-stationary time series is converted to stationary time series by differentiating for analysis of ARIMA model in the further steps. This is also a Moving Average of order infinity MA (∞). When it is presented with Moving Average of order infinity, Auto Regressive model is to be considered for analysis.

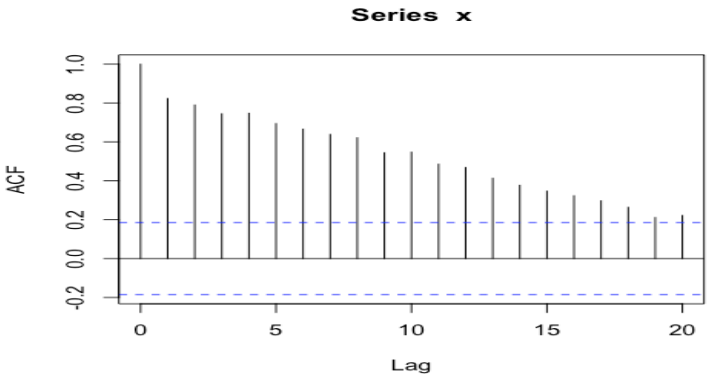


Figure 3: Auto correlation function

In time series analysis, the partial autocorrelation function (PACF) gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags. From Fig(4) it is interpreted that there is in an Auto Regression of order 4 i.e. AR(4). Further implementation of ARIMA models, Holt Winter and Neural networks forecasting used to predict and forecast the possible average temperature of the earth for the next is discussed in the next section.

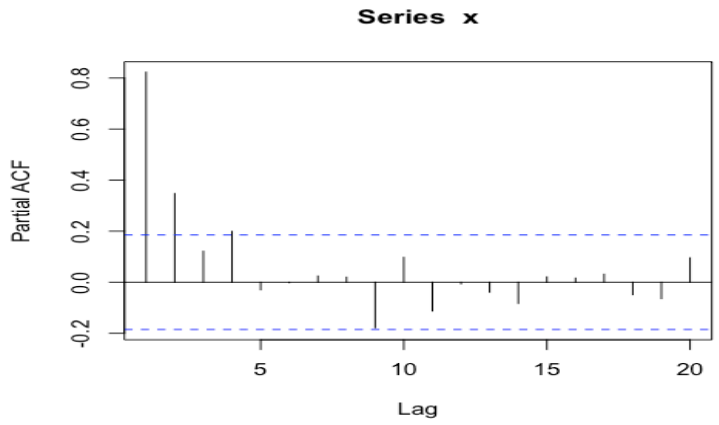


Figure 4: PACF for the time series

4. Computational Results:

All the computations were carried out in R 3.3.2 with various time series packages available. Some of the packages extensively used are 'forecast', 'tSeries'. Package 'DMwR' was used for data cleaning and to perform central imputations to make the data ready for analysis.

Using ARIMA model:

```
arima(x = x, order = c(4, 1, 0), method

Coefficients:
          ar1      ar2      ar3      ar
      -0.6463  -0.4244  -0.3999  -0.062
s.e.   0.0958   0.1070   0.1073   0.096

sigma^2 estimated as 0.03252:  log like
```

Figure 5: Results after differentiating the ARIMA (4,0,0)

Various ARIMA models were analyzed before choosing the best ARIMA model for this problem that is being analyzed. ARIMA (4,1,0) was chosen as the model since this model presented a better result. The p-values of Ljung Box Statistic is high and the residuals resemble white noise compared to the other models. Hence the order of the model present is ARIMA (4,1,0). Figure 5 shows the values of the coefficients of the chosen model.

In order to arrive at the best model, the time series was differenced but it did not provide a satisfactory result as the one obtained by integration the ARMA model once i.e. d=1. Figure 6 shows the result obtained by the best ARIMA model with ACF residuals and Ljung's Box test.

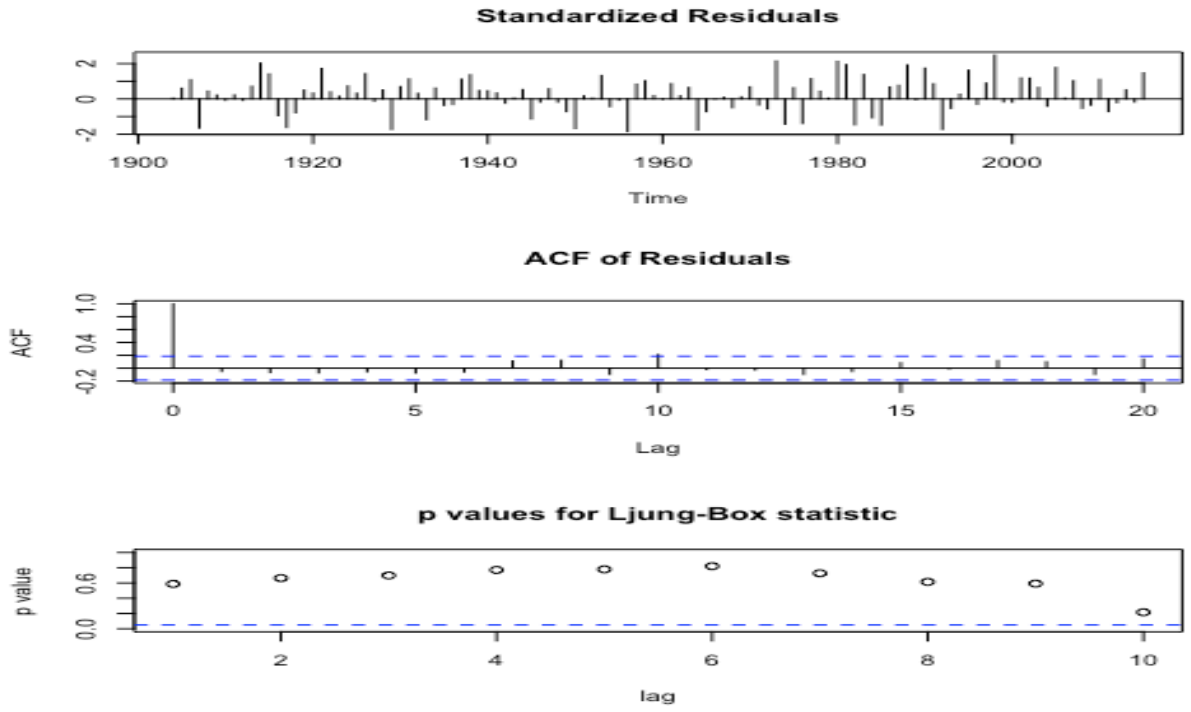


Figure 6: Residuals and Ljung's box test for ARIMA (4,1,0)

Final ARIMA model is chosen by selecting the best BIC from all the models built. From the results obtained ARIMA (1,1,2) has the lowest BIC value of -51.140932. Although ARIMA (0,1,1) had the BIC value of -50.23 which is almost close to the selected model, for this analysis ARIMA (1,1,2) is selected. Based on the model selected from the best BIC further predictions and forecasts will be done using ARIMA (1,1,2) model.

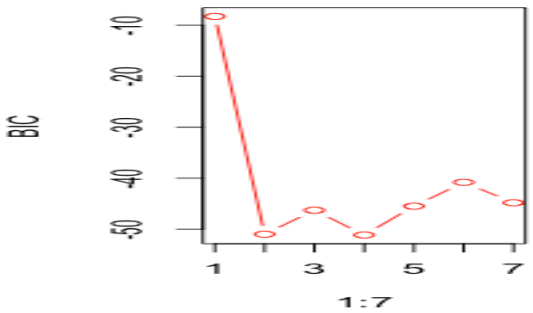


Figure 7:Choosing the best BIC

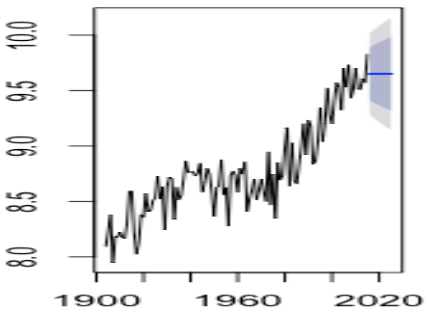


Figure 8:Forecast for ARIMA (1,1,2)

Forecast results built for ARIMA (1,1,2) is shown in the fig.8(See figure 8)
This forecast shown above from the ARIMA fitted model shows that there is slight increase in the earth's average temperature in the next decade. The increase is going be an average about 0.2˚ C. But when you consider 2015 which was one of the hottest years, the temperature is going to decrease by 0.2°C.

Using Holt Winter's Exponential Smoothing and Forecasting:

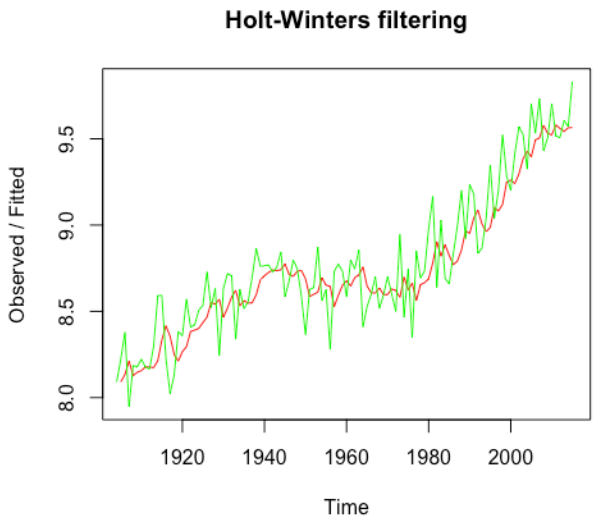


Figure 9: Exponential Smoothing Using Holt Winter

It can be observed that the time series of the forecast by holt Winter is much smoother than the given time series. Accuracy of the forecasted time series can be measured by sum of squared errors which is 3.78 in this case which means that the forecasted times series is almost accurate and is close to the given time series. Alpha value for this Holt Winter is alpha: 0. 3189865.Aplha value tells us that the forecasts are based on both recent and less recent observations.

Forecasting using Holt winters:

Holt Winter Forecast gives you a forecast value with 80% prediction interval and 95% prediction confidence interval as shown in the fig.10(See figure 10). The forecast obtained from the Holt winter shows that there is going be a slight increase in the earth's average temperature in the coming 10 years. But when compared to the average temperature of 2015, the earth's average temperature is going to decrease by 0.2 degree Celsius.

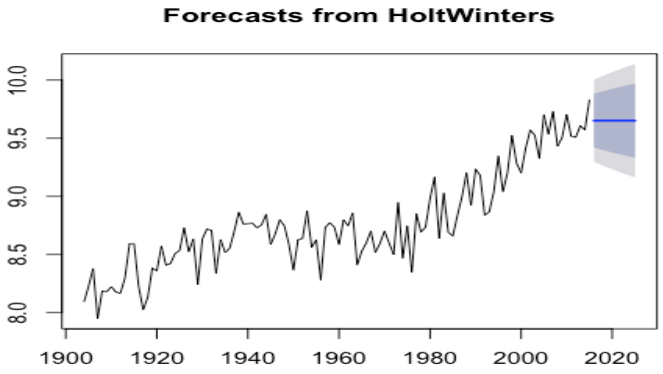


Figure 10:Holt Winter Forecast

To check the accuracy of the forecast, forecast errors are calculated. For this forecast the error is checked by checking the residuals value of the fitted model. If there are correlations between forecast errors for successive predictions, it is likely that the simple exponential smoothing forecasts could be improved upon by another forecasting.

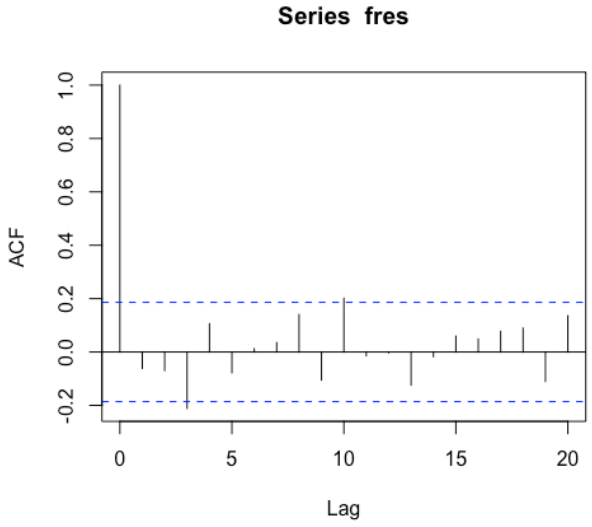


Figure 11; ACF of residuals

By plotting the ACF of residuals obtained from the fitted model, it is seen that auto correlation at is touching the significance line at Lag 4 and at Lag 10. Ljung's Box Test is conducted to determine if there is any significant non-zero autocorrelation between lag 1-20. The result of the test has a p-value of 0.1628. This means that there is no evidence of non- auto correlation function. The predictive model cannot be further improved upon, but to be sure normal distribution of forecast errors is check as shown in the figure below (See figure 12)

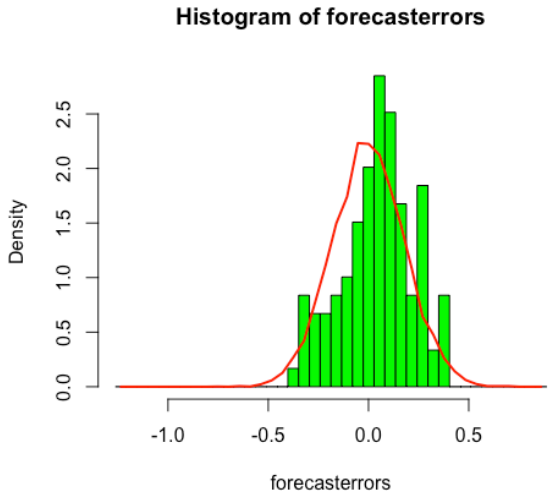
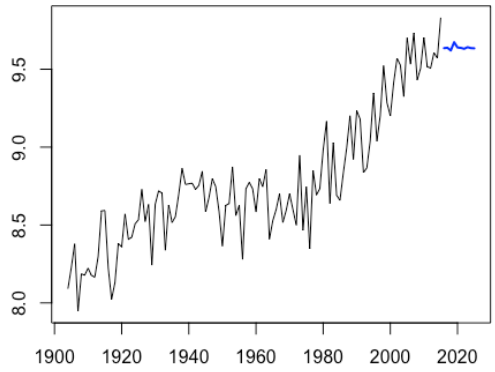


Figure 12: Error distribution

From figure 12 it can be inferred that the error is roughly centered around zero and normally distributed. This means that the error is normally distributed around zero and the predictive model cannot be further improved upon.

Forecasting using Neural Nets:



Forecasting model was built using nnetar from the forecast library in R. The results obtained from the neural network model did not have any significant difference from the other model.
Neural Net model like other two model discussed above forecasted that there will be a slight increase the earth's average temperature in the next 10 years. The result is shown in figure 13(See figure 13).

Figure 13:Neural Net Forecast

Conclusion:

Overall the results from the analysis seems satisfactory indicating that there is going to be a significant increase in the average temperature of the earth in the next 10 years. Holt Winter model provided an elaborate result of the analysis whose error rate was validated as well.

The analysis using ARIMA model forecasting, Holt Winter and Neural Network demonstrated that there has been an increase in the average global temperature on Earth. The data exhibits a rapidly decreasing auto covariance function thereby effectively fitting the AIRMA model to the time series data.

The error residual of the fitted models resembles white noise and hence one could infer that that the models have successfully extracted most information out of the data-set. Alongside the forecasts indicate that there would be a rise in the earth's average temperature by 0.2° / decade which is alarming considering the rate at which the average temperature increased over the past century.

In the future, a combined study of average land and ocean temperature can be carried out to get a broader understanding on key issues like climate change, global warming, erratic weather patterns can be determined. Further study max/min temperatures. Delving deep into this problem can help understand as to why the climate change is a concern and how one could do their bit to save the environment from warming up at this exponential rate.

Acknowledgement:
Special thanks to Dr. Ernest Fokoue for his guidance in Time-Series Analysis and Forecasting Theories, and his generous R code

Reference

Hansen.J, Sato.M, Ruedy.R,Lo. K, Lea.D, and Elizade.M. "Global Temperature Change." (n.d.).

Lorenz, E. N. Atmospheric predictability experiments with large numerical model. Tellus. 34, 505513. (1982).

National Research Council. [NRC]. "Climate Forecasting and Its Uses, doi: 10.17226/6370." (1999).

<solar-center.stanford.edu/sun-on-earth/glob-warm.html>.

(http://berkeleyearth.lbl.gov/auto/Global/Land_and_Ocean_complete.txt)

Appendix A.1: Time Series Analysis R Code

```
glob<-read.csv("/Users/Dhan/Desktop/GlobalTemperatures1.csv")
dim(glob)
globa<-ts(glob, start = c(1904), frequency =1)
dim(globa)
plot(globa[,2])
names(glob)
x<-globa[,2]
#n<-length(x)
#x<-x[n:1]
ts.plot(x)
#par(mfrow=c(1,2))
acf(x)
pacf(x)
arima.x<-arima(x,order=c(4,0,0), method = "ML")
print(summary(arima.x))
arima.x
arima.x1<-arima(x,order=c(4,1,0), method = "ML")
arima.x1
tsdiag(arima.x1)
tsdiag(arima.x)
xd<-diff(x)
arima.xd<-arima(xd,order = c(4,0,0))
arima.xd
tsdiag(arima.xd)
order.arima <- matrix(c(1, 0, 0,
                        1, 1, 0,
                        0, 0, 1,
                        0, 1, 1,
                        2, 0, 0,
                        2, 1, 0,
                        0, 0, 2,
                        0, 1, 2,
                        1, 0, 1,
                        1, 1, 1,
                        2, 0, 1,
                        2, 1, 1,
                        1, 0, 2,
                        1, 1, 2,
```

```
                  3, 0, 0,
                  3, 1, 0,
                  0, 0, 3,
                  0, 1, 3,
                  4, 0, 0,
                  4, 1, 0,
                  0, 0, 4,
                  0, 1, 4,
                  2, 0, 2,
                  2, 1, 2,
                  3, 0, 1,
                  3, 1, 1,
                  1 ,0, 3,
                  1 ,1, 3,
                  5, 0, 0,
                  5, 1, 0,
                  0, 0, 5,
                  0, 1, 5,
                  1, 0, 4,
                  1, 1, 4,
                  4, 0, 1,
                  4, 1, 1,
                  2, 0, 3,
                  2, 1, 3,
                  3, 0, 2,
                  3, 1, 2,
                  6, 0, 0,
                  6, 1, 0,
                  0, 0, 6,
                  0, 1, 6,
                  5, 0, 1,
                  5, 1, 1,
                  1, 0, 5,
                  1, 1, 5,
                  4, 0, 2,
                  4, 1, 2,
                  2, 0, 4,
                  2, 1, 4,
                  3, 0, 3,
                  3, 1, 3),
                ncol=3,
                byrow=T)

complexity <- numeric(nrow(order.arima))
#x <- ((x[length(x):1])^(1/1))

for(j in 1:nrow(order.arima))
{
  p <- order.arima[j,1]
  d <- order.arima[j,2]
  q <- order.arima[j,3]
  arima.x <- arima(x, order=c(p,d,q))
  complexity[j] <- BIC(arima.x)
}

results          <- data.frame(order.arima, complexity, rowSums(order.arima))
```

```
colnames(results) <- c('p','d','q', 'BIC', 'size')
print(results)
mod.code <- NULL
for(j in 1:nrow(order.arima))
{
  mod.code <- c(mod.code, paste(order.arima[j,1],order.arima[j,2], order.arima[j,3], sep=''))
}

#x11()
#par(mfrow=c(1,2))

plot(results[,5], results[,4])
text(results[,5], results[,4], labels=mod.code, pos=4)


BIC <- numeric(7)

for(l in 1:7)
{
  BIC[l]<-min(complexity[which((order.arima[,1]+order.arima[,2]+order.arima[,3])==l)])
}
plot(1:7, BIC, type='b', col='red')
best <- min(which(results[,4]==min(results[,4])))
best
library(forecast)
arima.x <- Arima(x, order=order.arima[best, ])
#x11()
ts.plot(x, col='green', lty=1)
lines(fitted(arima.x), col='red', lty=2)
legend('top', c('original series','estimated series'), lty=c(1,2), col=c('green','red'), inset=0.02)
plot(forecast(Arima(x, order=order.arima[best, ])))
plot(forecast(Arima(x, order=c(1,1,2))))
pred <- predict(arima.x, n.ahead = 10)
pred
plot(pred)
plot(x,type='l',xlim=c(2015,2025),ylim=c(8,10),xlab = 'Year',ylab = 'Average Temperature')
lines(10^(pred$pred),col='blue')
lines(10^(pred$pred+2*pred$se),col='orange')
lines(10^(pred$pred-2*pred$se),col='orange')
ts.plot(x, pred$pred, lty = c(1,2), col=c('green','red'))
legend('bottom',c('Given','Predicted'),lty = c(1,2),col=c('green','red'))
tsdiag(Arima(x, order=order.arima[best, ]))
fit<-nnetar(x)
fcast<-forecast(fit)
plot(fcast)
predi<-predict(fit)
predi
ts.plot(x,predi$mean,lty=c(1,2))
hw<-HoltWinters(x, beta = FALSE, gamma=FALSE)
hw
hw$SSE
plot(hw, lty=c(1,2),col=c('green','red'), inset=0.03)
fcast<-forecast.HoltWinters(hw, h=10)
fpred<-predict(hw,n.ahead = 10)
plot(hw,predicted.values=fpred)
plot.forecast(fcast)
```

```
fcast$residuals
plot.ts(fcast$residuals)
fres<-na.omit(fcast$residuals)
acf(fres)
Box.test(fres, lag=20, type="Ljung-Box")

hw$SSE
plotForecastErrors <- function(forecasterrors)
{
 # make a histogram of the forecast errors:
 mybinsize <- IQR(forecasterrors)/4
 mysd   <- sd(forecasterrors)
 mymin  <- min(forecasterrors) - mysd*5
 mymax  <- max(forecasterrors) + mysd*3
 # generate normally distributed data with mean 0 and standard deviation mysd
 mynorm <- rnorm(10000, mean=0, sd=mysd)
 mymin2 <- min(mynorm)
 mymax2 <- max(mynorm)
 if (mymin2 < mymin) { mymin <- mymin2 }
 if (mymax2 > mymax) { mymax <- mymax2 }
 # make a red histogram of the forecast errors, with the normally distributed data overlaid:
 mybins <- seq(mymin, mymax, mybinsize)
 hist(forecasterrors, col="green", freq=FALSE, breaks=mybins)
 # freq=FALSE ensures the area under the histogram = 1
 # generate normally distributed data with mean 0 and standard deviation mysd
 myhist <- hist(mynorm, plot=FALSE, breaks=mybins)
 # plot the normal curve as a blue line on top of the histogram of forecast errors:
 points(myhist$mids, myhist$density, type="l", col="red", lwd=2)
}
plotForecastErrors(fres)
fit <- tbats(x)
seasonal <- !is.null(fit$seasonal)
seasonal
```