Time Series Analysis of Climate Change from England Temperature Data

Introduction

The goal of this paper is to analyze and evaluate the claims made by several researchers (Jones, Harvey, et al) regarding climate change. Having read through all the papers, I list the major claims as follow:

1) Jones argues for a change in approach to the time series analysis of climate change from using a linear trend analysis to applying more sophisticated stochastic analysis that can account for stochastic nature of temperature more accurately as it is time series data.

2) One debate surrounding climate change is that one side denies the whole global warming as an aberrant behavior caused by human activity, and believes the a few centuries-long increment in temperature arises from the natural cyclic activity of earth while the opposite side argues, as well known, that the global warming is caused by human industrial activity. Harvey argues that the inference of and confidence interval in the temperature projection that the earth will only get warmer is uncertain and concedes that that the surge in temperature could be interpreted as random spike that can happen naturally from stochastic nature of climate.

3) The final major claim made in the papers is whether global warming has slowed down in recent decade. Some scientists have argued that there was a “hiatus in global warming” in the past 15 years as the actual temperatures were far below the projected temperature. According to the papers, two camps to the hiatus issue contend that: 1) global warming has completely stopped after 1998 2) it has slowed down but did not stop. The latter claims that the planet is warming but not as fast as it was projected for unknown reasons.

In this paper, I will incorporate the advanced statistical analysis mentioned in 1), and use various time series tools (namely, non-parametric fitting and time domain approach time series techniques) to evaluate the claims mentioned in 2) and 3).

Data

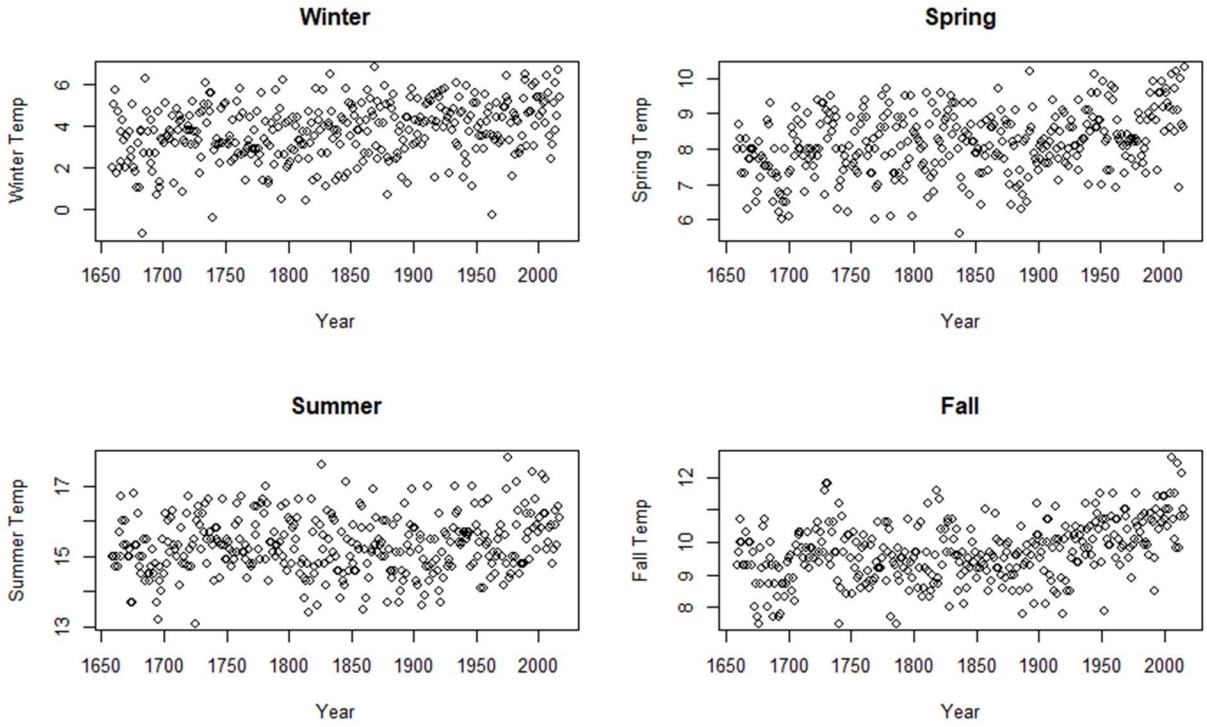
This paper draws data from central England which was originally put together by Manley (1974), and is now routinely updated by the Hadley Centre in England. The data are available from <www.cru.uea.ac.uk/¾mikeh/datasets/uk/cet.htm>, and consist of a series of monthly mean surface air temperatures for a region representative of central England. The potential problems with data measurement and the evolution of techniques and tools to measure it are detailed in Jones. Succinctly put, he claims that researchers should have faith in the ethics and standards adhered by contemporary scholars and technicians that measured them. In this paper, I will not contest this claim, and take the temperature data as a credible measurement.

However, the inference from three centuries of data may be extremely limited because they may not be sufficient data points in terms of time coverage and geographic locality for extrapolating whether the change in global warming is anthropogenic or not. In fact, this comes to play an important role in my final analysis. In addition, one should be cautious in generalizing the climate behavior to globe from CET because the data are pertinent only to this locality

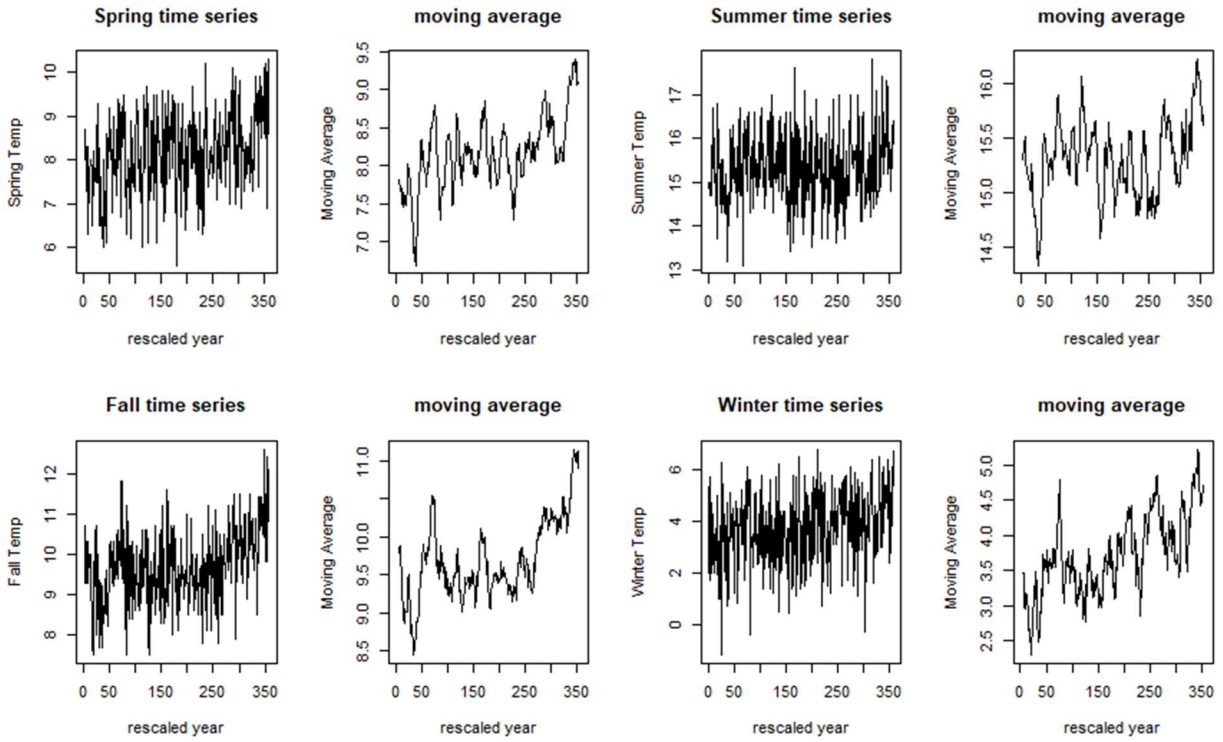
Analysis

I start the analysis by confirming the presence of global warming in the data. We can dissect the data by seasons in order to arrive at more specific and nuanced inference of data. Once we do, we detect that seasons have different warming patterns, so we can have more specific patterns that arise from the data. In addition, we can also control for seasonality effect

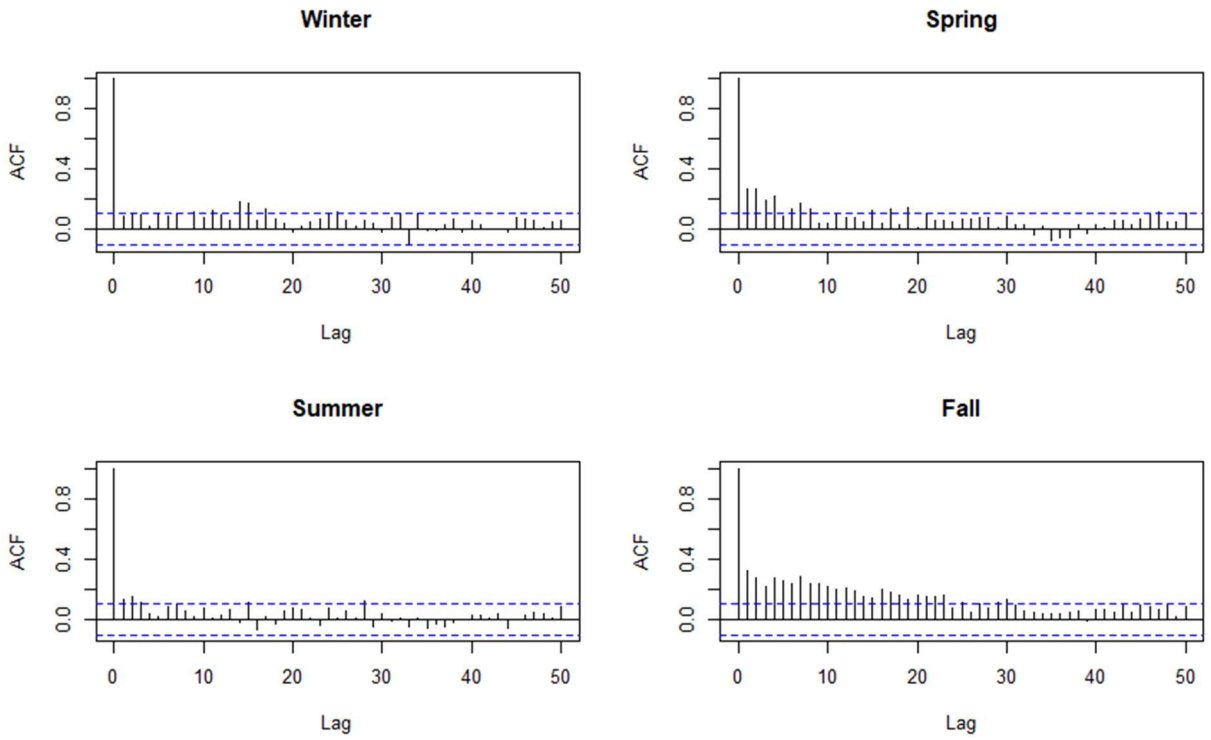
that can hide other important patterns of the changes in temperature. From the plot of data by seasons, we see divergent patterns among them. Fall temperature has a clear increasing trend, dating from 1900. Other seasons do not have the increasing trend until 1950s.



Since we are interested in fluctuations of decadal temperature in order look at change in long term trend and long-term dependence between successive years of data. We take a moving average of decade-long lagged variables, i.e. we employ a moving average filter by taking an average of 10-samples and this helps remove noise.

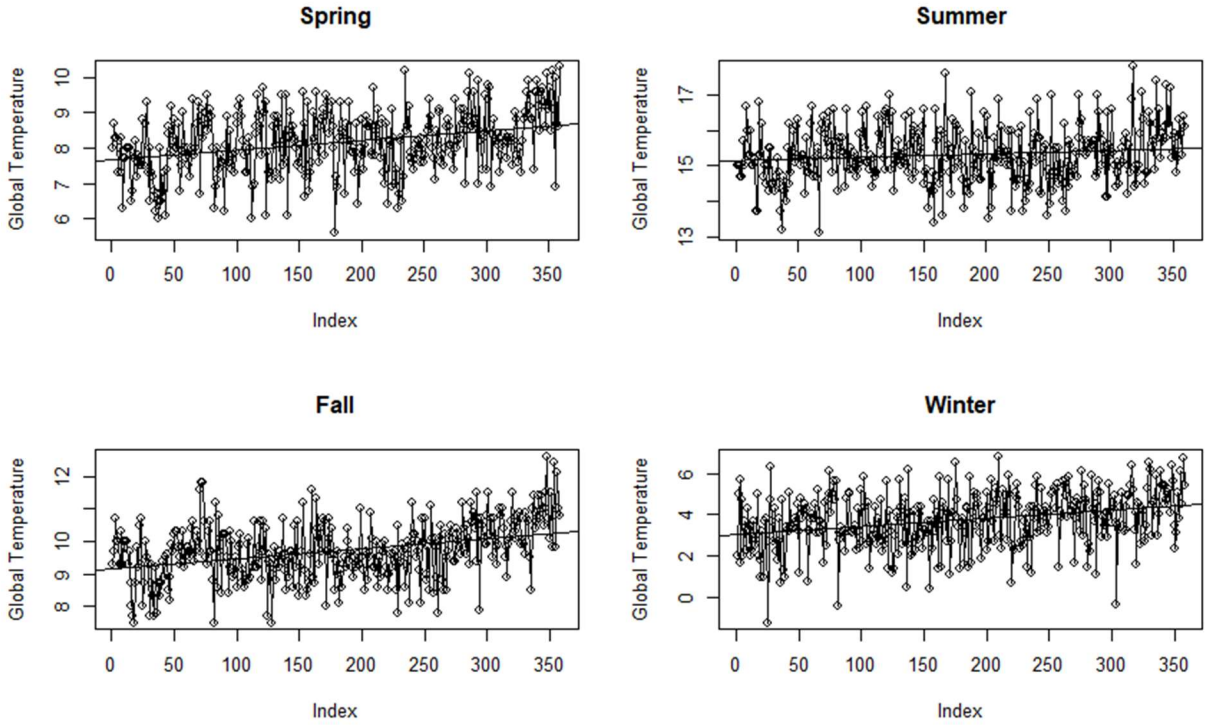


The moving average plot confirms our initial look at the data that the temperature has been increasing in the past few centuries, and we can put to rest to any claim that warming does not exist as a function of time.



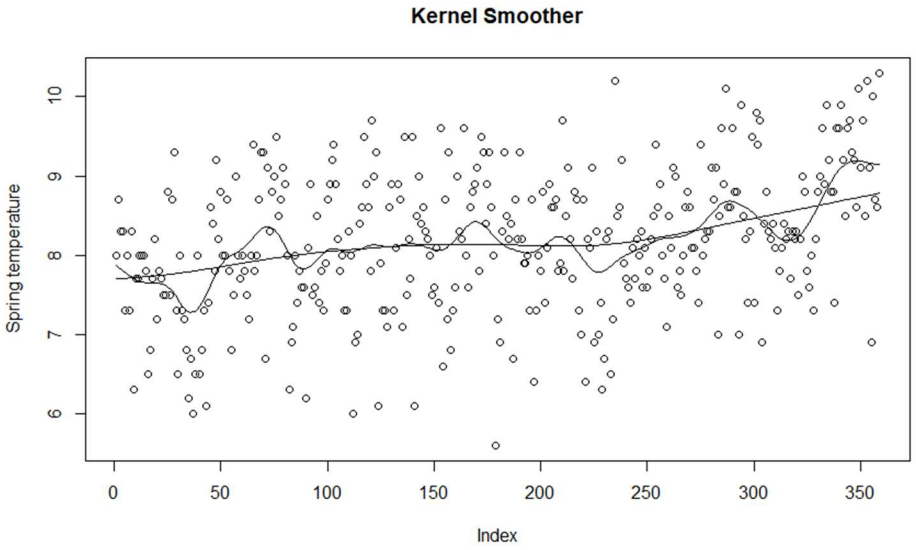
Next, I use time-series analysis to see if data contains any autoregressive property which would indicate time-dependent relation. Again, we use seasonal data to segment information as much as possible because seasonality effect may exist. Based on ACF values of the original data of each season, we see that time dependence is nearly zero for winter and summer. In contrast, spring and fall have steep time dependence particularly during the fall season, as we see several lags still significant at up to 20 years. This tells us two seasons are essentially random variables without any time dependence and whose covariance for each time lag is nearly zero, while others have long time lag up to two decades.

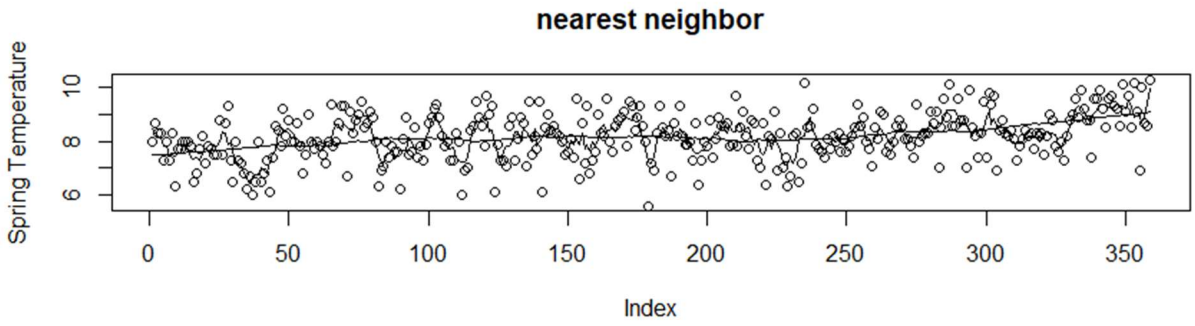
Next we use linear regression on the xyplot to test whether any exists by linear regression and also to fit a line that serves as a visual aid, as lines often do in myriad data points. We see that for each season the trend is almost flat for the past three centuries, but gently increasing trend in for spring and fall.

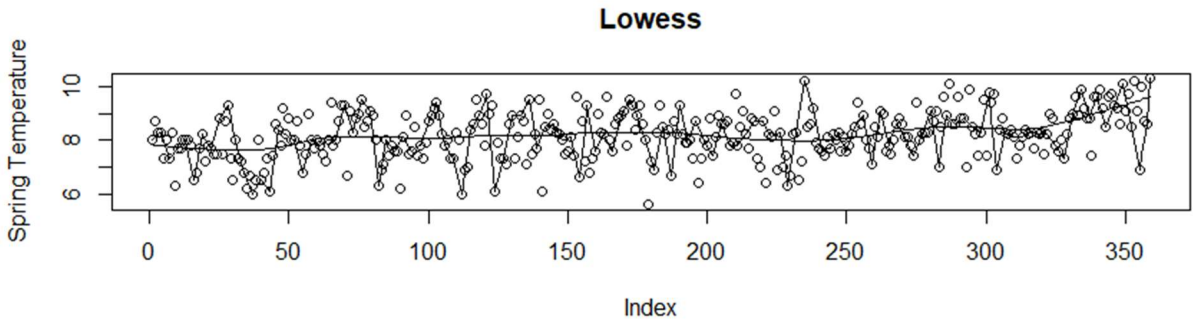


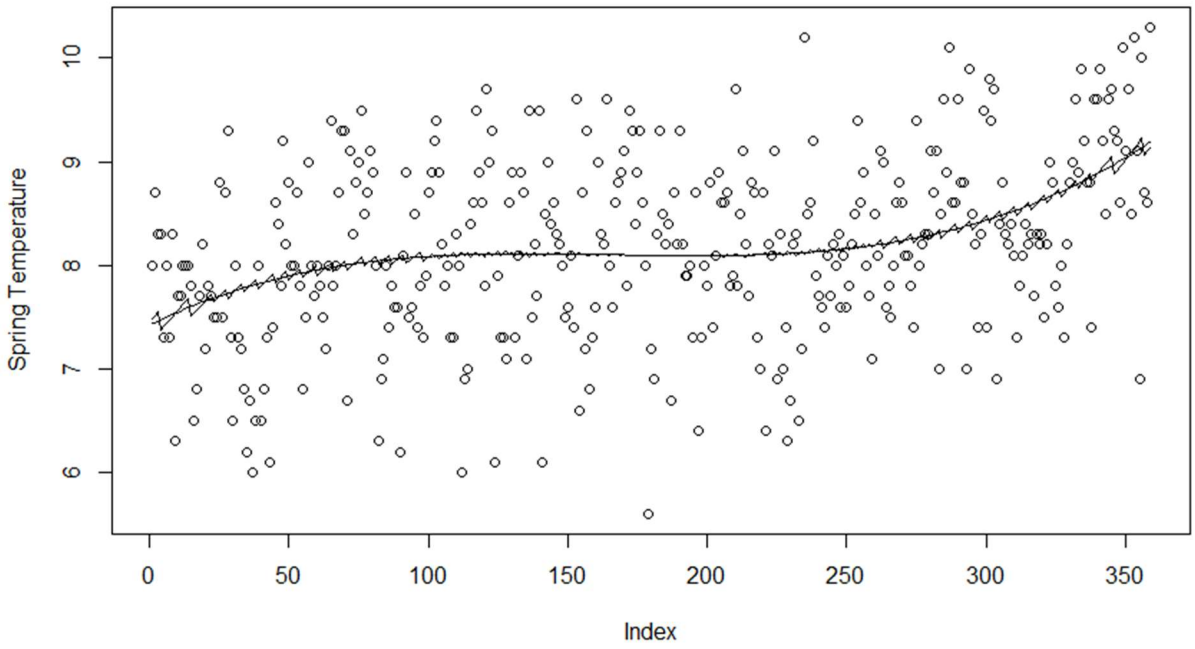
Now that we have finished a rudimentary analysis of the data, we approach the data via more sophisticated tools to create a more localized fit to each neighborhood of data. Localized fitting will help us answer the question whether warming is taking place at different rates.

I employ several non-parametric fitting with cross-validated bandwidth, which yields the model with the smallest mean squared error, to fit the seasonal data.









Non-parametric techniques show different smoothing patterns depending on the season and show very oscillating behavior. Since non-parametric techniques are in essence localized means, the vicissitudes imply frequent cyclic behavior of climate, which is consistent with the climate science observation about vacillating temperature in temperate regions. The graphs exhibit somewhat extreme behavior particularly at tails but this is inherent to localized regression techniques because they have fewer data points. This also entails larger confidence bands around the edges - called boundary bias. If we do use polynomial regression, we obtain a different result where the next temperature seems projected to increase in power terms and this might be why the projection made by some was so high because it forecasted extremely higher. On the other hand, Lowess/KNN algorithms yield different conclusions for the predictions for the spring data.

We see that the regression result depends on which smoothing parameter we use, and this is the important result because depending on how we choose a parameter and which technique we

implement, the fitting is either negatively sloped or positively sloped for the most recent 15 years.It is also true that depending on how we tolerate cross-validation score, the fit may also change. In the end, the choice of model is up to subjective judgment about which fit is better and how significant cross-validation score is.

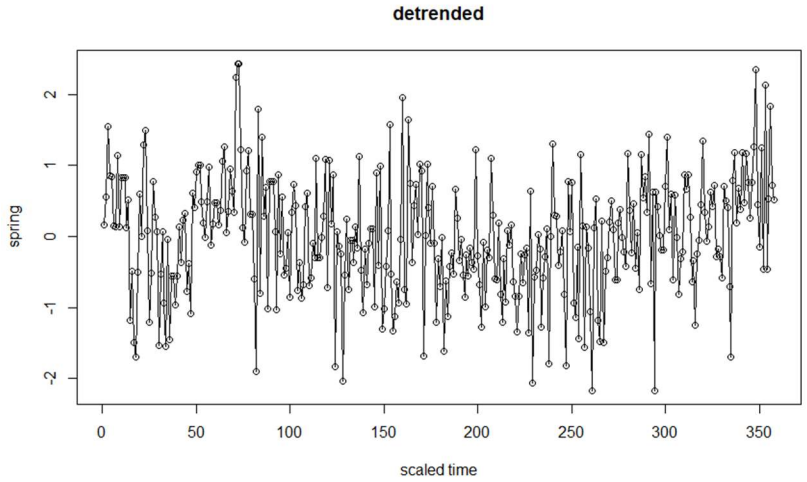
Using non-parametric fitting, I evaluated Harvey's claim that the region experienced three periods of cooling, stability, and warming throughout the data, that can be roughly associated with the beginning and the end of the Industrial Revolution, but it is difficult to gauge if this is true based on various non-parametric models. There appears to be untrended variations and not necessarily cooling in the early data.

Next, I also used broken stick linear regression on the data and see whether the hypothesis that the warming has decreased over the recent 15 years is true.

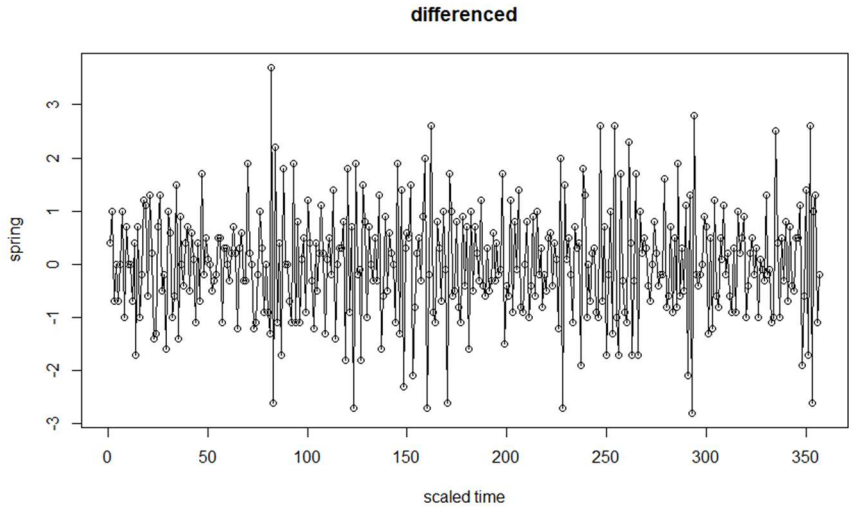
```
Analysis of Variance Table

Model 1: MAM ~ Year
Model 2: MAM ~ Year + Year:I
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    357 254.54
2    356 248.69  1    5.8486 8.3724 0.004044 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is statistically significant result that there is a distinct change in slope in the past 15 years. So the F-test of single linear fit versus broken stick regression tells us that the region experienced a decrease of growth of temperature in the past 15 years.

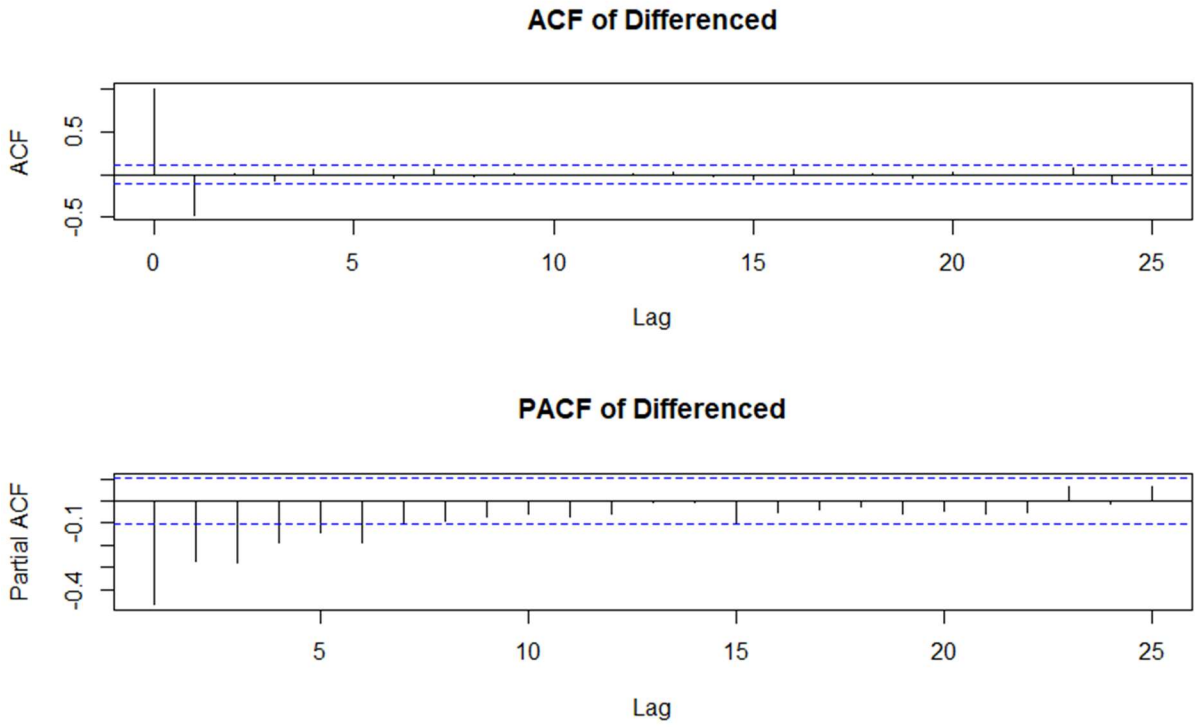
Next we analyze the stochastic behavior around the trend lines because I am curious what time series analysis might indicate about the trend of the past fifteen years. Detrending the spring data, we see some non-stationary time series in the data.



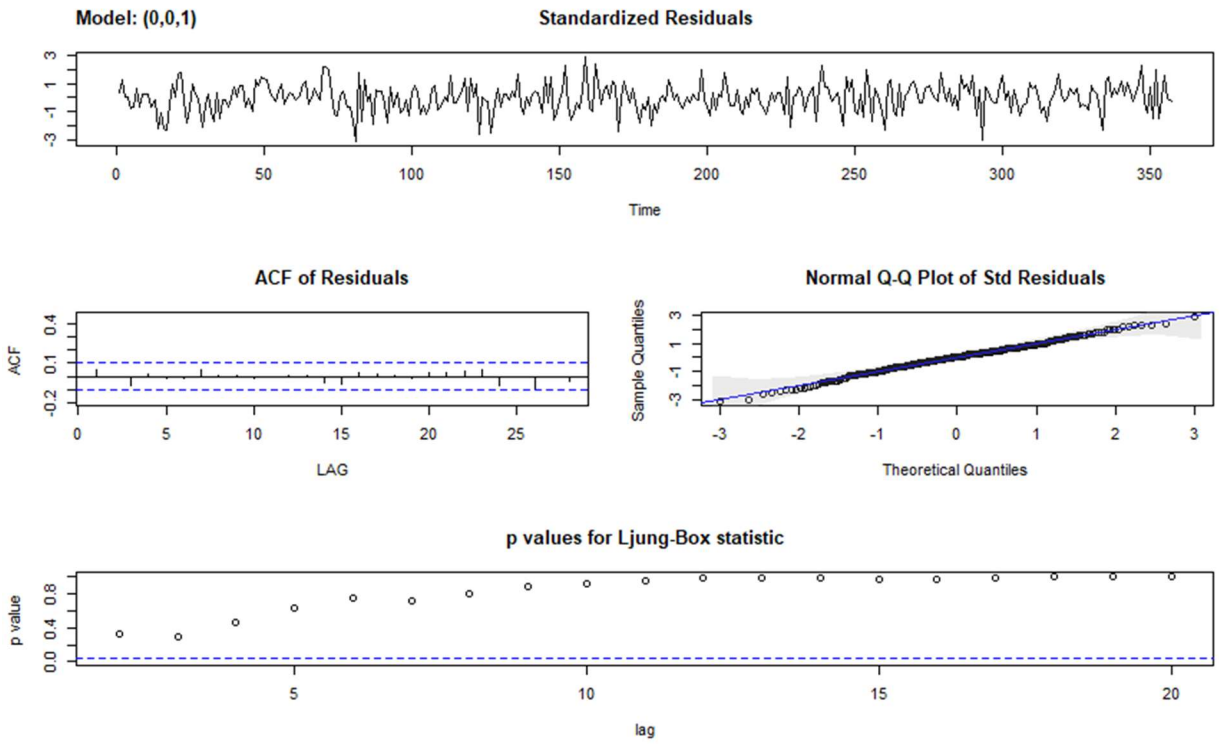
In addition, we take a first-order difference of the data and then try to model this differenced data.



It looks like a first differencing yields a stationary process. The analysis has no need for second differencing because I do not detect any quadratic trend. The model now appears more stationary, so we conduct an ARIMA analysis to model the differenced series.



The MA(1) model seems appropriate for the differenced series as we see a slowly decaying PACF and a single spike for lag 1 in ACF plot.



The model diagnostic analysis suggests that we have residuals that are not autocorrelated and are roughly normally distributed, and whose p-values for each time lag is small only for lag 1. Model diagnostics are okay, and the prediction of ARIMA shows correctly that the temperature rise from 2001 to 2016 will be slow, but the forecasting with ARIMA shows there is a considerable uncertainty with the prediction as the standard errors are enormous relative to the estimates

```
> predict(arima(df[-342:-358], order = c(0,0,1)), n.ahead = 17)
$pred
Time Series:
Start = 342
End = 358
Frequency = 1
 [1] -0.153202157  0.002928748  0.002928748  0.002928748
0.002928748  0.002928748  0.002928748  0.002928748  0.002928748
[10]  0.002928748  0.002928748  0.002928748  0.002928748
0.002928748  0.002928748  0.002928748  0.002928748

$se
```

```
Time Series:
Start = 342
End = 358
Frequency = 1
 [1] 0.7963374 1.0702853 1.0702853 1.0702853 1.0702853 1.0702853
1.0702853 1.0702853 1.0702853 1.0702853 1.0702853 1.0702853
[12] 1.0702853 1.0702853 1.0702853 1.0702853 1.0702853 1.0702853
```

This uncertainty regarding the prediction supports Harvey's claim that inference of the hypothesis that the aberrant rise in global temperature in the past 100 years is human-forced and will continue to rise is uncertain because the prediction is too uncertain.

Conclusion

In this paper, I implemented various techniques to assess the claims made in the readings. In particular, I placed a heavy emphasis on testing whether the warming has slowed down in the past 15 years, and from broken-stick regression's z-test on the time coefficient reveals the warming has indeed slowed down compared to the past trend. The test of hypothesis that the slope of trend is zero failed to be rejected, which means the warming has not stopped but the slope has decreased in the past 15 years. However, it should be noted that some non-parametric fitting can indicate upward or downward trend, depending on the choice of smoothing parameters. In addition, summer and winter months have had insignificant changes in temperature, so the warming should really refer to spring and autumn months only. At least for these months, the claim that global warming has stopped seems statistically void.

However, having computed the confidence band for regression and ARIMA analysis, I concur with Harvey that one should not be so assured to claim the warming will continue to occur. It is well within reason to believe that the global temperature anomaly record is the result of a random cyclic phenomenon - still unpredictable due to lack of deep understanding in climate

- rather than a forced physical phenomenon; that is. that a random activity of nature can produce the temperature record as we've observed it. However, we should keep in mind some kind of systemic forcing, be it anthropogenic or natural, can produce this record as well. Data prove one hypothesis is right but does not necessarily disprove all the other hypotheses. Also, the extent of my analysis is only using two data features - time and temperature, so the best we can do is make conclusion about association between the two variables, which leaves out a cornucopia of variables that affect the temperature. Hence the next research should focus on including as many relevant variables as we can and back up the result of statistical analysis via climate science theories.

Reference

Gayathri Vaidyanathan (February 25, 2016) Did Global Warming Slow Down in the 2000s, or Not? Scientists clarify the recent confusion. http://www.scientificamerican.com/article/did-global-warming-slow-down-in-the-2000s-or-not/

JONES, P. D. & BRADLEY, R. S. (1992a). *Climatic variations in the longest instrumental records. In Climate Since A.D. 1500, Ed*. R. S. Bradley and P. D. Jones, London: Routledge. pp. 246ā68.

JONES, P. D. & BRADLEY, R. S. (1992b). *Climatic variations over the last 500 years. In Climate Since A.D. 1500, Ed.* R. S. Bradley and P. D. Jones, London: Routledge. pp. 649ā65.

BENNER, T. C. (1999). Central England temperatures: Long-term variability and teleconnections. Int. J. Climatol. 19, 391ā403.

HARVEY, D. I. & MILLS, T. C. (2003). Modelling trends in central England temperatures. J. Forecasting 22, 35ā47.

```
# R Code
```

```
data <- read.table("C:/Users/jihun/Desktop/stat 361 data
file.txt", quote="\"", comment.char="")

data.soi <- read.table("C:/Users/jihun/Desktop/soi index.txt",
quote="\"", comment.char="")

# winter, spring, summer, autumn

Year = data$V1

DJF = data$V2

MAM = data$V3

JJA = data$V4

SON = data$V5


# exclude the first point

par(mfrow=c(2,2))

plot(Year[-1],DJF[-1], xlab="Year", ylab="Winter Temp",
main="Winter")

plot(Year, MAM, xlab="Year", ylab="Spring Temp", main="Spring")

plot(Year, JJA, xlab="Year", ylab="Summer Temp", main="Summer")

plot(Year[-359], SON[-359], xlab="Year", ylab="Fall Temp",
main="Fall")


# filtered by moving average

par(mfrow=c(2,4))

v1 = filter(MAM, sides=2, rep(1/10,10))

plot.ts(MAM, xlab="rescaled year", ylab= "Spring
Temp",main="Spring time series")

plot.ts(v1, xlab="rescaled year", ylab="Moving Average",
main="moving average")
```

```
v2 = filter(JJA, sides=2, rep(1/10,10))

plot.ts(JJA, xlab="rescaled year", ylab= "Summer
Temp",main="Summer time series")

plot.ts(v2, xlab="rescaled year", ylab="Moving Average",
main="moving average")


v3 = filter(SON[-359], sides=2, rep(1/10,10))

plot.ts(SON[-359], xlab="rescaled year", ylab= "Fall
Temp",main="Fall time series")

plot.ts(v3, xlab="rescaled year", ylab="Moving Average",
main="moving average")




v4 = filter(DJF[-1], sides=2, rep(1/10,10))

plot.ts(DJF[-1], xlab="rescaled year", ylab= "Winter
Temp",main="Winter time series")

plot.ts(v4, xlab="rescaled year", ylab="Moving Average",
main="moving average")


# ACF

par(mfrow=c(2,2))

acf(DJF[-1],lag.max=50,main="Winter")

acf(MAM,lag.max=50,main="Spring")

acf(JJA,lag.max=50,main="Summer")

acf(SON[-359],lag.max=50,main="Fall")


# plot with fitted linear trend line
```

```
par(mfrow=c(2,2))


plot(MAM, type="o", ylab="Global Temperature", main="Spring")

abline(lm(MAM ~time(MAM)))


plot(JJA, type="o", ylab="Global Temperature", main="Summer")

abline(lm(JJA ~time(JJA)))


plot(SON[-359], type="o", ylab="Global Temperature",
main="Fall")

abline(lm(SON[-359] ~time(SON[-359])))


plot(DJF[-1], type="o", ylab="Global Temperature",
main="Winter")

abline(lm(DJF[-1] ~time(DJF[-1])))


par(mfrow=c(1,1))

plot(time(SON[-359]), resid(lm(SON[-359] ~ time(SON[-359]))),
type="o", main="detrended", xlab ="scaled time", ylab="spring")

plot(time(SON[-359:-358]), diff(SON[-359]), type="o",
main="differenced", xlab ="scaled time", ylab="autumn")


df <- diff(SON[-359])

par(mfrow=c(2,1))

acf(df, main="ACF of Differenced")

pacf(df, main="PACF of Differenced")
```

```
library(astsa)

sarima(df, 0,0,1)

sarima(SON[-359], 0, 1, 1)


# broken stick linear regression

I = ifelse(Year > 2000, 1, 0)

fit1 = lm(MAM ~ Year)

fit2 = lm(MAM ~ Year + Year:I)

anova(fit1,fit2)


summary(lm(JJA ~ time(JJA)))



# Periodogram

n=length(MAM)

I = abs(fft(MAM))^2/n

P = (4/n)*I[1:(n/2)]

f = 0:((n/2)-1)/n

par(mfrow=c(1,1))

plot(f, P, type="l", xlab="Frequency", ylab="Scaled
Periodogram")


par(mfrow=c(2,2))

plot(predict(arima(df[-342:-358], order = c(0,0,1)), n.ahead =
17)$pred[-1],c(2001:2016))

plot(c(2001:2016),df[343:358])
```

```
pre = predict(arima(df[-342:-358], order = c(0,0,1)), n.ahead =
17)$pred[-1]

length(pre)

plot(c(2001:2016),pre)

plot(c(2001:2016),df[343:358])



# Trend plus periodic (global)

par(mfrow=c(1,1))

t = Year - mean(Year)

t2 = t^2; t3 = t^3

cs=cos(2*pi*t); sn=sin(2*pi*t)

reg1 = lm(MAM ~ t + t2 + t3)

reg2 = lm(MAM ~ t + t2 + t3 + cs + sn)

plot(MAM, type="p", ylab="Spring Temperature")

lines(fitted(reg1)); lines(fitted(reg2))


# Kernel smoother

plot(MAM, type="p",ylab="Spring temperature", main="Kernel
Smoother")

lines(ksmooth(time(MAM), MAM, "normal", bandwidth=20))

lines(ksmooth(time(MAM), MAM, "normal", bandwidth=100))


# Nearest Neighbor and Lowess

par(mfrow=c(2,1))

plot(MAM, type="p", ylab="Spring Temperature", main="nearest
neighbor")
```

```
lines(supsmu(time(MAM), MAM, span=.2))
lines(supsmu(time(MAM), MAM, span=.005))
plot(MAM, type="p", ylab="Spring Temperature", main="Lowess")
lines(lowess(MAM, f=0.01)); lines(lowess(MAM, f=0.2))

# Smoothing spline
par(mfrow=c(2,1))
plot(time(MAM),MAM, main="Smoothing spline")
lines(smooth.spline(MAM, spar=1))
plot(time(MAM),MAM, main="Smoothing spline")
lines(smooth.spline(MAM, spar=1/2))

# periodogram of detrended
par(mfrow=c(1,1))
fit <- lm(MAM ~ time(MAM))
res <- resid(fit)
n = length(res)
ord = 1:((n-1)/2)
freq = (ord)/n
per = as.vector(abs(fft(res))^2/n)
# P = Mod(2*fft(res)/100)^2 ; Fr = 0:99/100
plot(freq,per[ord],type="l",main="scaled periodogram")

# PACf vs acf
par(mfrow=c(2,1))
```

```
pacf(MAM)

acf(MAM)


# AR(2) regression

regr = ar.ols(MAM[-(350:359)], order=2, demean=FALSE,
intercept=TRUE)

fore = predict(regr, n.ahead=10)

par(mfrow=c(1,1))

ts.plot(fore$pred)

ts.plot(MAM[350:359], fore$pred, ylab="MAM temp")

U=fore$pred + fore$se

L = fore$pred - fore$se

xx=c(time(U), rev(time(U))); yy = c(L, rev(U))

polyhon(xx, yy, border=8, col=gray(0.6, alpha= 0.2))

lines(fore$pred)


# Yule Walker Estimation

MAM.yw = ar.yw(MAM, order=2)

MAM.yw$x.mean # 8.169 mean estimate

MAM.yw$ar # (0.2145, 0.2081) coefficient estimates

sqrt(diag(MAM.yw$asy.var.coef)) # 0.05183, 0.05183 stabdard
errors

MAM.yw$var.pred # 0.7096 error variance estimate


MAM.pr = predict(MAM.yw, n.ahead=24)

plot(c(2017:2040),MAM.pr$pre)
```

```
# ts.plot(MAM, MAM.pr$pred)

# log-difference like GNP plot
ldMAM = log(diff(MAM)) # diff(MAM) has zeroes'
acf(ldMAM, na.action = na.pass)
pacf(ldMAM)
# AR (1)
sarima(ldMAM, 1, 0, 0)
# MA (2)
sarima(ldMAM, 0,0, 2)
ARMAtoMA(ar = 0.35, ma=0, 10)


# diagnostics of ARIMA model
sarima(gnpgr, 1, 0, 0)
# (xt - one-step-ahead prediction) /
# normality of residuals - must be standard normal
# zero correlation
# correlations shouldn't be too large - Ljung-Box test
# if there is autocorrelation, run WLS

# Regression with Lagged Variables

# spectrum
arma.spec(asdlog="no", main="MAM")
```