Brandon Rhodes
12/10/2019
STAT 33600
Project

Introduction

The Central Englad Temperature (CET) series has been of paramount interest to many researchers over the past few decades, given some changes in the global temperature in recent history. Many people are concerned with the current trend in the global temperature, and they want to know mainly two things: (i) is there an increase in the global temperature, whether it be over the course of the Earth's history, and, more specifically, in the recent history; and (ii) if there is an increase in the temperature, people want to know what is the mechanism behind it, and people especially want to know whether or not it can be ascribed to human behavior. These are interesting questions, and the CET dataset is a good start for getting some insight to the questions.

CET dataset and its importance

The CET dataset is the one which has the most comprehensive instrumental observations of the temperature in areas of central England. Most other datasets which have information on the global temperature dating back as far as the CET are not instrumental, or there is reason to heavily question the precision of the instruments. In addition to measurments not being from instruments, a lot of the measurements were indirect – e.g. sometimes the amount of rainfall or amount of snowfall on a given day would be noted as a proxy for temperature. As Manley (1974), these measurements are highly variable and often times they do not have any useful correlation with the temperature. With this is mind, the CET dataset is valuable because it offers us the longest credible time series that we can work with.

Claims to be answered

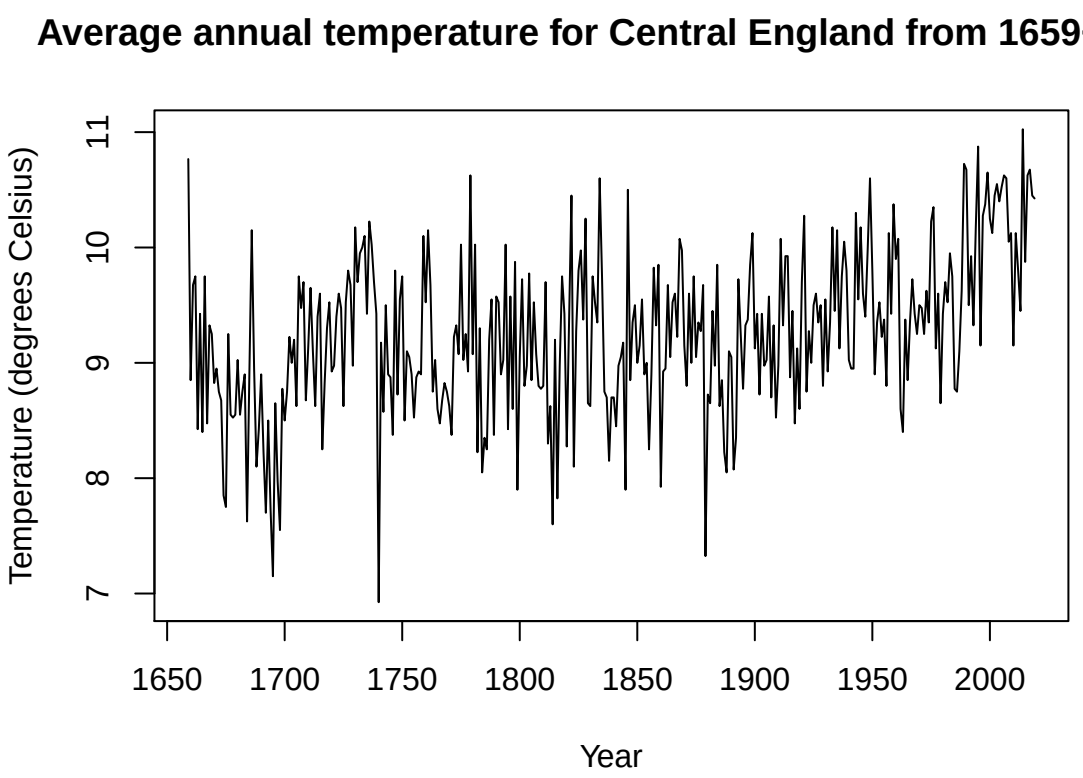
The claims we will be considering come from some of the literature on the topic. In this project, we want to investigate the following claims:

1. There is a warming trend at the macro level.
2. There has been an increasing trend in the global temperature recently (last ~ 20 years).
3. There is oscillatory behavior of the global temperature.

Outline of project

The project report here will follow a basic track to justify the claims above. First, to justify the claim of a macro warming trend, I will show that the series is not stationary and that we will need to difference or transform it to do so. Depending on the nature of this process will give insight as to what the trend is. Second, I will look more specifically at the most recent years and predict the next few years. These predictions will depend on the model, so I will take time and care to ensure we have a model which has a good fit with the data so that we will have some confidence in the conclusions we reach will. Last, to examine the oscillatory behavior of the series, I will look at the series from a different perspective: the frequency domain. I will again work with a stationary version of the series, and from there I will do some analysis of the spectra. The above should suffice in allowing us to justify the claims made in the literature. At the end, I will discus whether they are justified and state further conclusions.

Exploratory Phase

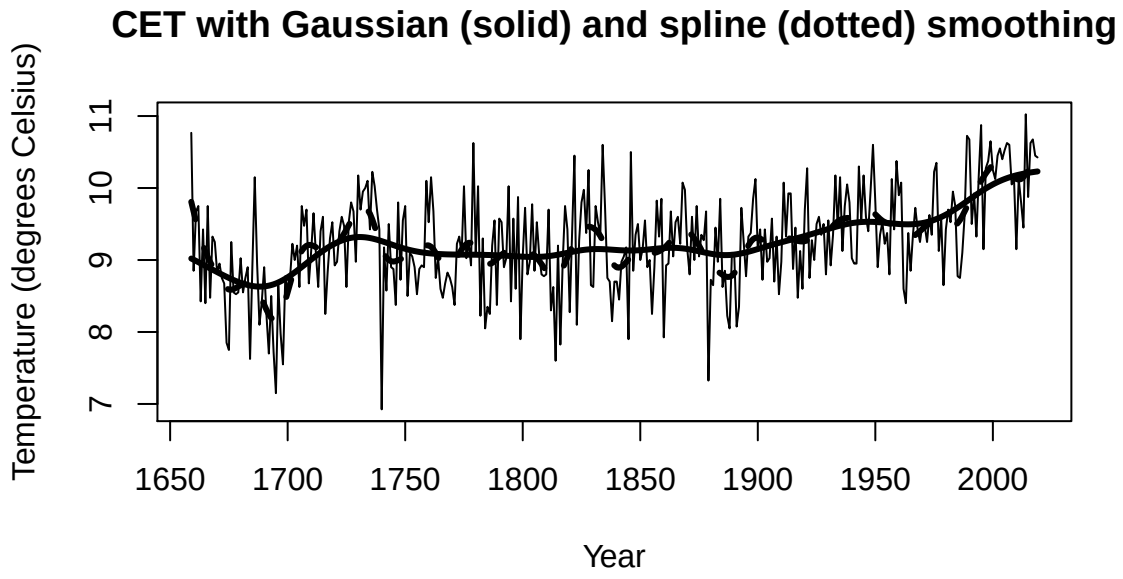


Notes on the data

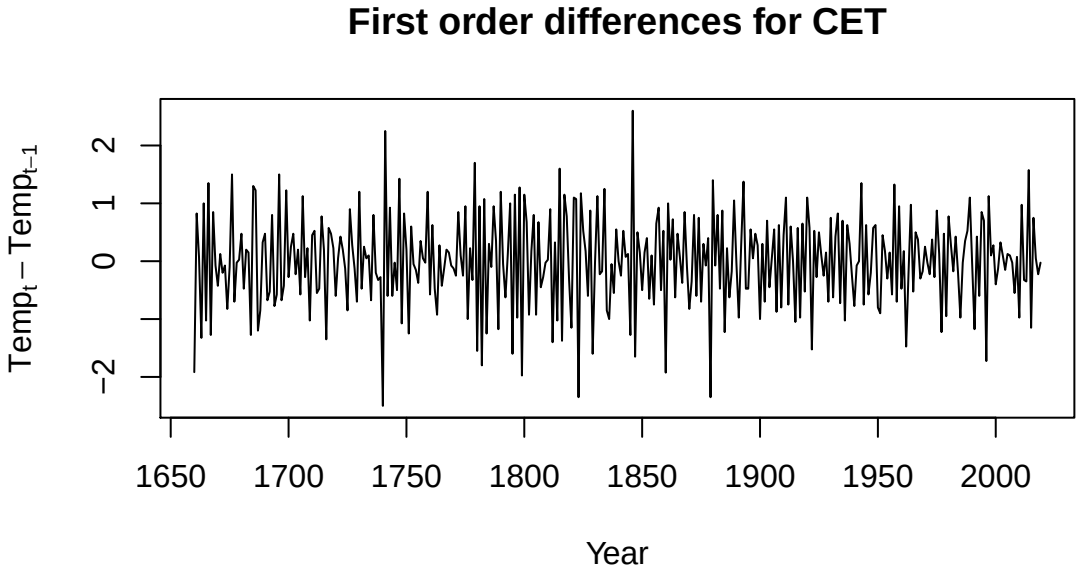
We see that the data are not stationary: the mean does not appear to be constant in time. The most deviation from this constant behavior is in the recent history, starting around the 1950s. We also so a large dip in the global temperature in the late 17th century, which has been referred to as the Little Ice Age. This cooling will not be of interest to us, except when we consider the macro trend in the global temperatures. Since the series is not stationary, we will need to make it stationary in some way to be able to apply any ARIMA models and for much of our inference to hold. Below, we will consider different ways to achieve this stationarity when selecting our model.

Selecting model class

In the plot above, we see the series is not stationary because there are definite trends in the data, so before we do any analysis it will be useful to try to remove these trends to make it stationary. In the reference papers, the authors had used non-parametric methods to examine the trend, such as Gaussian kernel smoothing, local polynomial smoothing, Butterworth filters, etc., and below I will show Gaussian kernel smoothing and cubic spline smoothing to emphasize the presence of the trend. I used the bandwidth of 35 for the Gaussian kernel, which is based on the value reported in Harvey et al. (2003); I used smoothing parameter $\lambda = 2.9 \times 10^{-6}$, which was based on a leave-one out cross-validation.



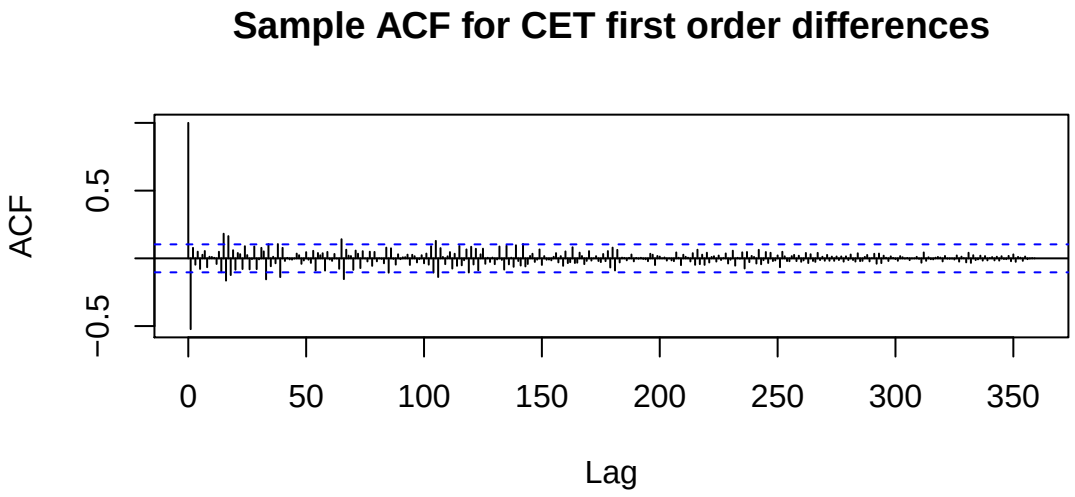
Although both of these methods are promising with regard to analyzing the data, I will follow a different track. I will consider making the series stationary by considering the first order differences of the series. We ultimately want to answer – and hopefully in a manner which is precise – the question of whether or not there is a general increase in temperature, and using a first order difference model could be a first attempt at giving a more precise answer to the question. As I will show below, considering a first order difference will effectively remove the trend making the resulting series appear relatively stationary, and then we can proceed to model it as a ARMA series. I will not consider any non-linear time series for a couple of reasons. First, I think it would be better to start with a more simple hypothesis. Scond, there is no apparent reason to believe that the nonstationarity of the original series is due to its covariance structure: the variance across the whole series appears to be constant, so heteroskedasticity should not be a problem.

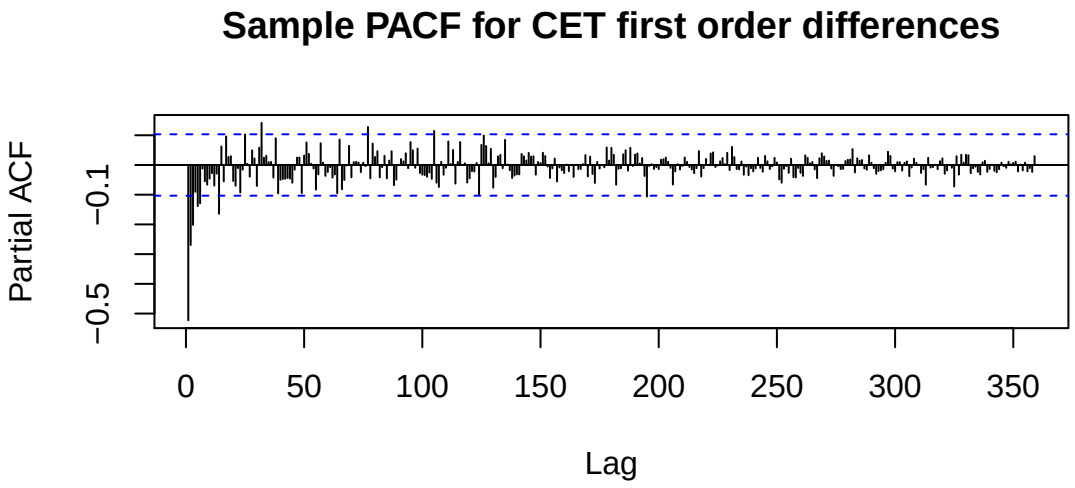


Considering the first order difference time series, in addition to making the series stationary, offers us a more intuitive interpretation and easier forecasting. For the interpretation, if we observe the mean of the time series as positive, this would suggest that there is an increase in the temperature and vice versa for negative mean; moreover, we will know by large sample theory (or bootstrap) the sampling distribution of this mean, and we will be able to make a more precise statement about our conclusion. As for forecasting, we will not face the same problems of extrapolating from the data as a nonparametric smoothing method would face. Now, we will consider which (F)ARIMA model we will choose.

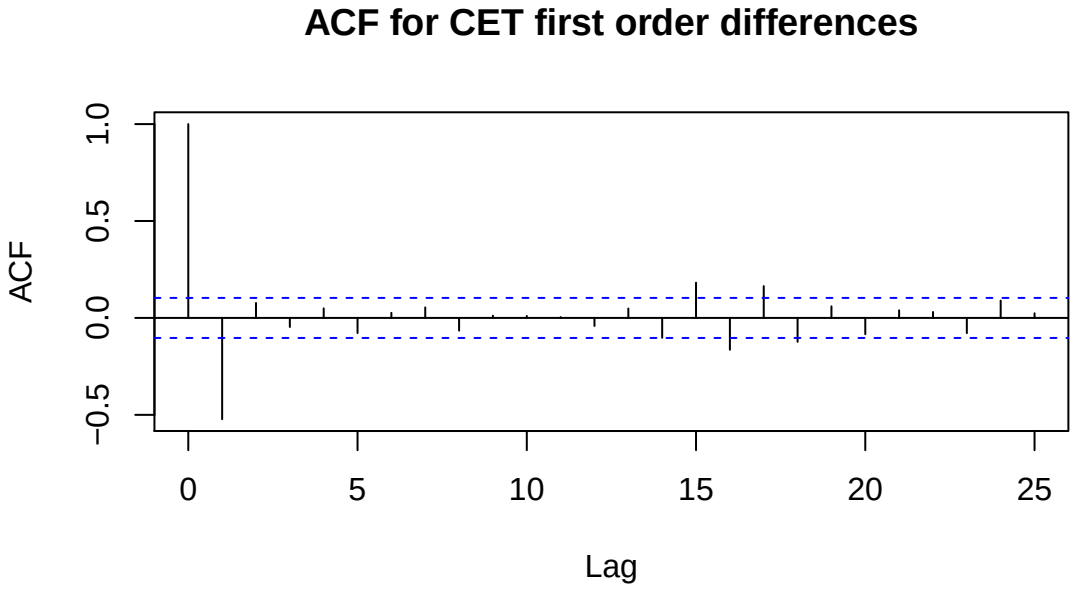
Choosing model within (F)ARIMA

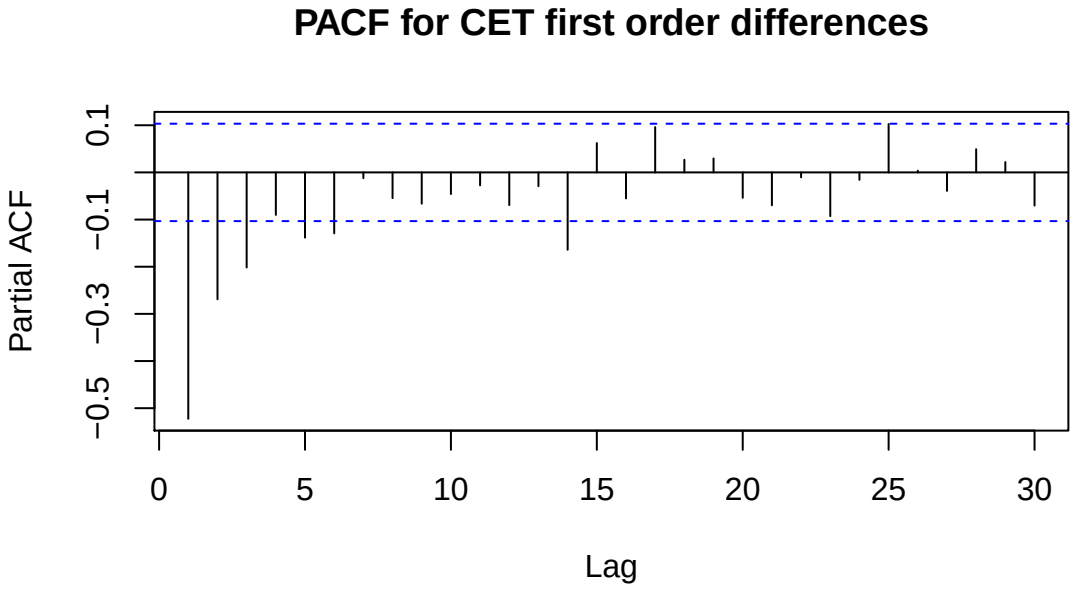
As a first step in trying to determine the orders for the FARIMA model, we will look at the sample autocorrelation and partial autocorrelation. The first plot below is the sample ACF and PACF when looking at large values of the lag. We see that there does not appear to be any notable long range behavior in the series, so I am not inclined to explore any fractional model; however, a more careful analysis of the power spectrum may be worth looking at to make sure. For now, we will just consider those more immediate lags and fit an ARIMA model.





Below, we see that there is a significant negative correlation at lag 1 year in the autocorrelation, with a few peaks which may be significant around lags 15-17 years. The peak at lag 1 is readily apparent in the series above: the series tends to oscillate about its mean. The potential peaks around lag 15-17 years could indicate some two decade cycle of weather behavior, but right now it is not clear whether or not those peaks are significant. More generally speaking, there is definitely some behavior at lag 1 year, but the sample ACF appears to taper off after it. The behavior of the PACF shows a few significant peaks at lags of 2 and 3 years, but it tapers off afterwards. The behavior of the ACF and PACF suggest that the series of first order differences is either MA(1) or ARMA; it does not suggest that we are dealing with a purely autoregressive process. We will proceed to fit both MA(1) and ARMA models, evaluating the fits as we go.

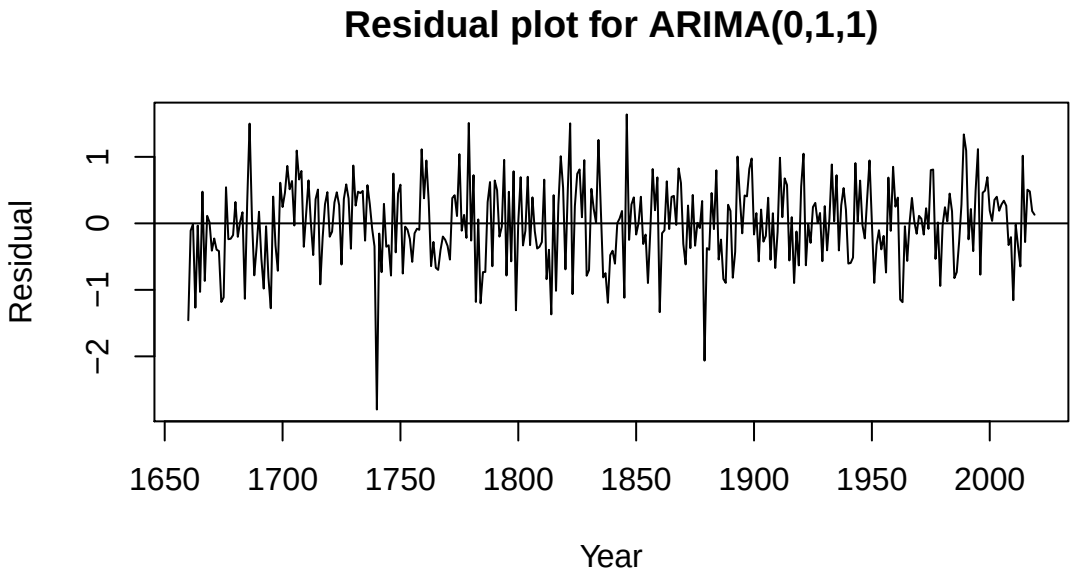


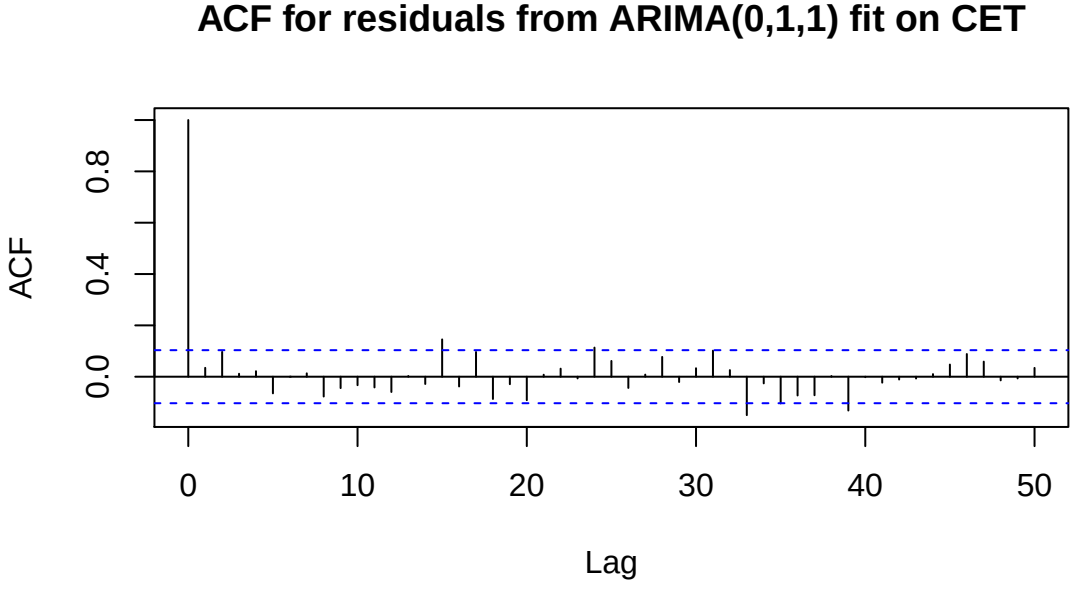


Fitting ARIMA(0,1,1) model

First, we will fit the simpler first order moving average model for the differenced series, with mean zero Gaussian noise. This model means that every observation in the series $y_t = x_t - x_{t-1}$ is dependent upon past values only through the noise terms at the immediately preceding year and the current year. In terms of mathematical notation, this means the following:

$$x_t - x_{t-1} = y_t = \theta_1 w_{t-1} + w_t \qquad \text{where } w_i \overset{\text{iid}}{\sim} N(0, \sigma_w^2)$$





After fitting an MA(1) model to the differenced series, we have the fitted model as well as the estimated parameters and standard errors. We see that the residual plot does not look terrible, but there does still appear to be some cyclic trend in the residuals, indicating potential lack of fit. The fitted model has an estimated mean which is positive, but small, at 0.0029 degrees Fahrenheit. Given the nature of our model, this would indicate a slight overall warming trend in the data. The moving average parameter is quite large at -0.86. This would indicate the tendency for fluctations in the process to come back toward the mean; in other words, although, there is a slight overall trend, it is not pronounced. The estimated standard deviation is 0.62 degrees Fahrenheit. To further examine the fit, I used the Ljung-Box test statistic for lag 30 years, and the p-value was 0.06, which suggests that we should explore some other fits.

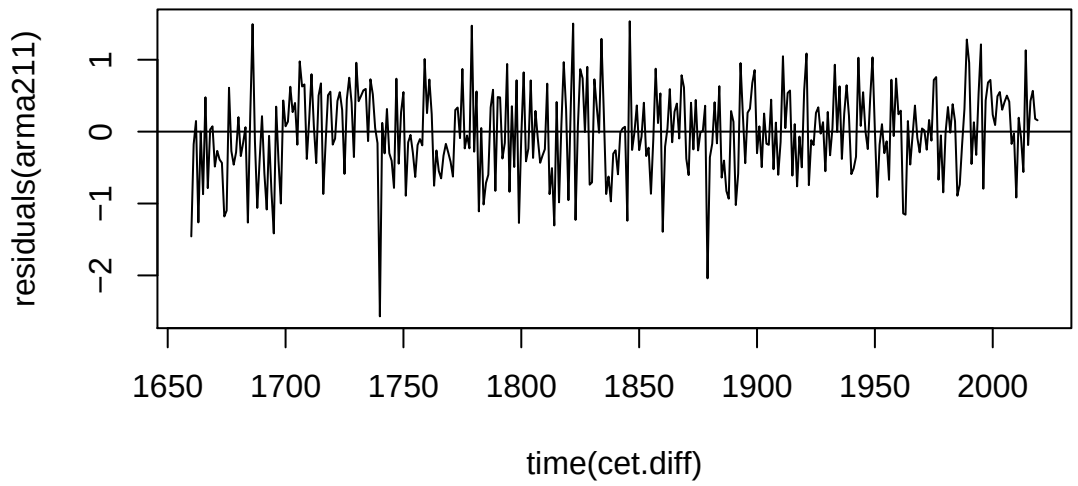
Parameter	Estimate	Standard error
μ	0.0029	2.2^{-5}
θ_1	-0.86	1.5×10^{-3}
σ_w^2	0.38	
Ljung-Box statistic	$\chi^2_{28} = 5.6489$	$p = 0.06$

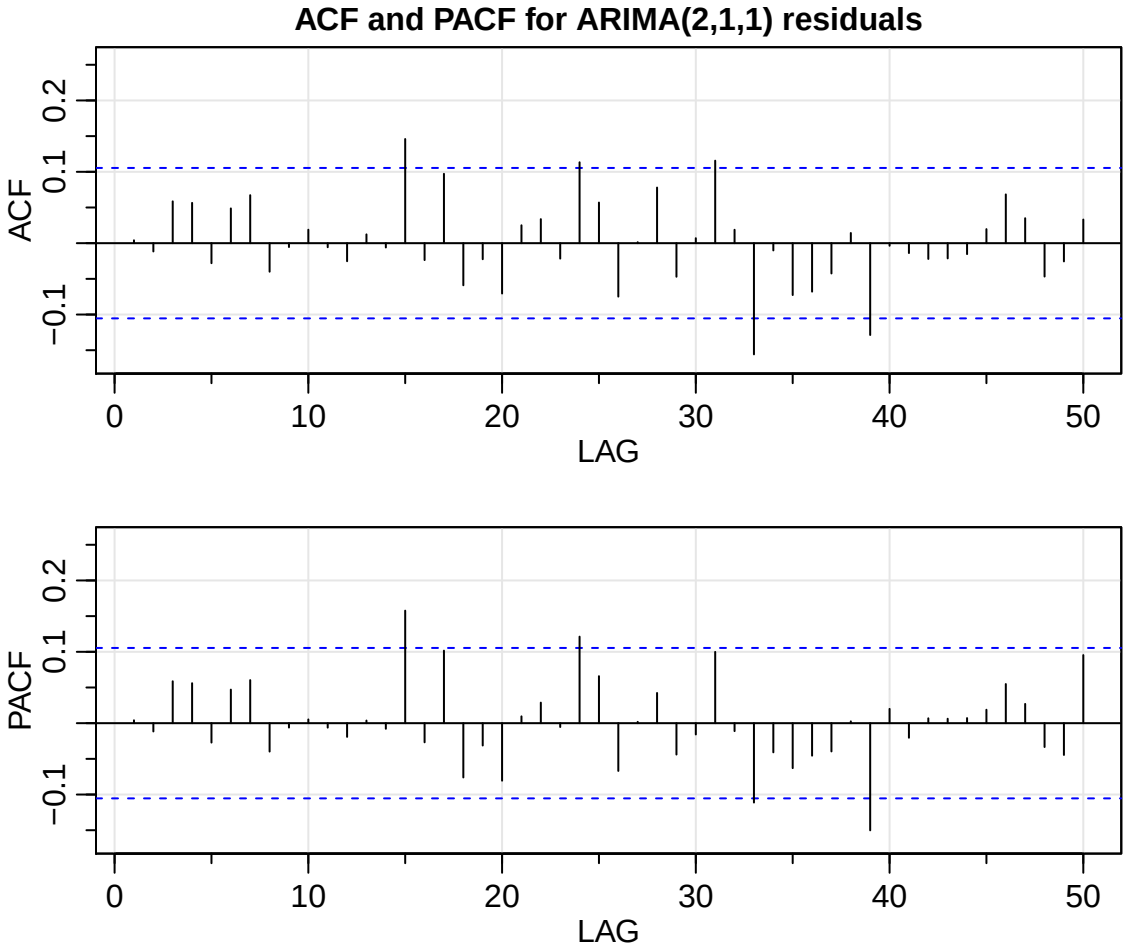
Fitting ARIMA(p,d,q) model

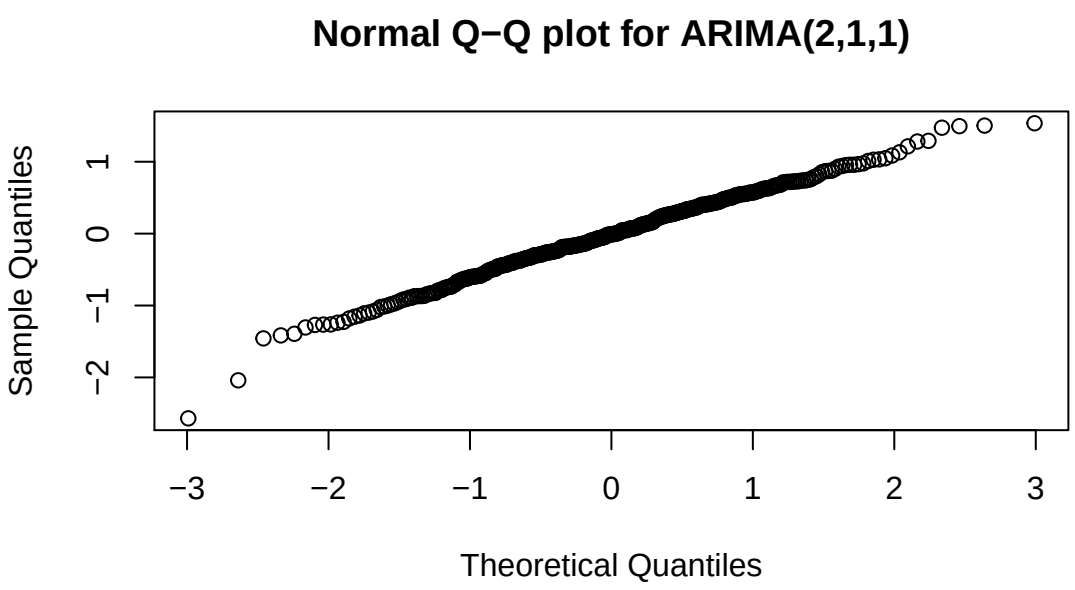
After testing several models, I will chose the ARIMA(2,1,1). After doing a grid search over candidate models for $p = 1, \ldots, 10$ and $q = 1, \ldots, 10$ and $d = 1, \ldots, 5$, I found that the models ARIMA(1,1,2) and ARIMA(2,1,1) had better performance when maximizing the loglikelihood and minimizing the Bayesian Information Criterion and AIC. The ARIMA(1,1,2) model was the maximizer and minimizer; however, the margin by which it did so over ARIMA(2,1,1) was small. In terms of log-likelihood, the likelihood ratio is 1.09 in favor of ARIMA(2,1,1), so there is not strong evidence that this model is markedly better. This small difference made it a more comfortable decision to choose the model with more autoregressive terms for its clearer interpretation. I did not consider fractional differences because in the ACF and PACF for the differenced series, it did not appear that there was any long range behavior.

Model	Log-likelihood	AIC	BIC
ARIMA(2,1,1)	-332.43	674.86	694.29
ARIMA(1,1,2)	-332.34	674.69	694.12
LR	1.09	in favor of ARIMA(1,1,2)	

Below, we observe that the residual plot looks to behave nicely. It resembles that of white noise, which indicates that there are not any flagrant violations of our model assumptions. Using the Ljung-Box test statistic, we see that for a lag 30 years, we have a p-value of 0.18, which also suggests that the fit is good. In the ACF and PACF for the residuals, we see that at lags 1 and 2 years, the most important for our purposes, the estimated correlations do not reach anywhere near significance, which is also a positive sign for the model. Last, we see that the normal quantile-quantile plot shows that our error distribution has a slightly lighter tail than that of the normal distribution on the right side, but slightly heavier tail on the left. This is appears to be a light violation of our normality assumption, so if we were evaluating more detailed claims this would be something to investigate more; here, we will not be worried about this mild violation.


residuals(arma211)
1
0
−1
−2
1650
1700
1750
1800
1850
1900
1950
2000
time(cet.diff)


ACF and PACF for ARIMA(2,1,1) residuals
ACF
0.2
0.1
−0.1
0
10
20
30
40
50
LAG
PACF
0.2
0.1
−0.1
0
10
20
30
40
50
LAG



The model fit is described by the equation below, where the current observation is model as a linear combination of the two previous values and the immediately preceding noise term. The assumption for the standard errors from this model is assuming mean zero Gaussian noise.

$$x_t - x_{t-1} = y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 w_{t-1} + w_t \qquad \text{where } w_i \overset{\text{iid}}{\sim} N(0, \sigma_w^2)$$

The estimates $\hat{\phi}_1, \hat{\phi}_2$ we have for the autoregressive parameters are 0.11 ± 0.06 and 0.17 ± 0.06, indicating that there is some dependence between the current difference and previous differences. The estimate for the moving average parameter is -0.95 ± 0.04, much larger than the autoregressive parameters. This captures the oscillating behavior of the differences, suggesting the tendency for the undifferenced series is to go back to a mean state. The intercept is estimated at 0.0034 ± 0.0023. This intercept is small, which may indicate a small drift of the temperature upwards. This is consistent with the ARIMA(0,1,1) model above; however, we must be cautious about drawing this conclusion because the standard error for this estimate is much larger in terms of its ratio than the other parameters we have estimated. Either way, the effect is small.

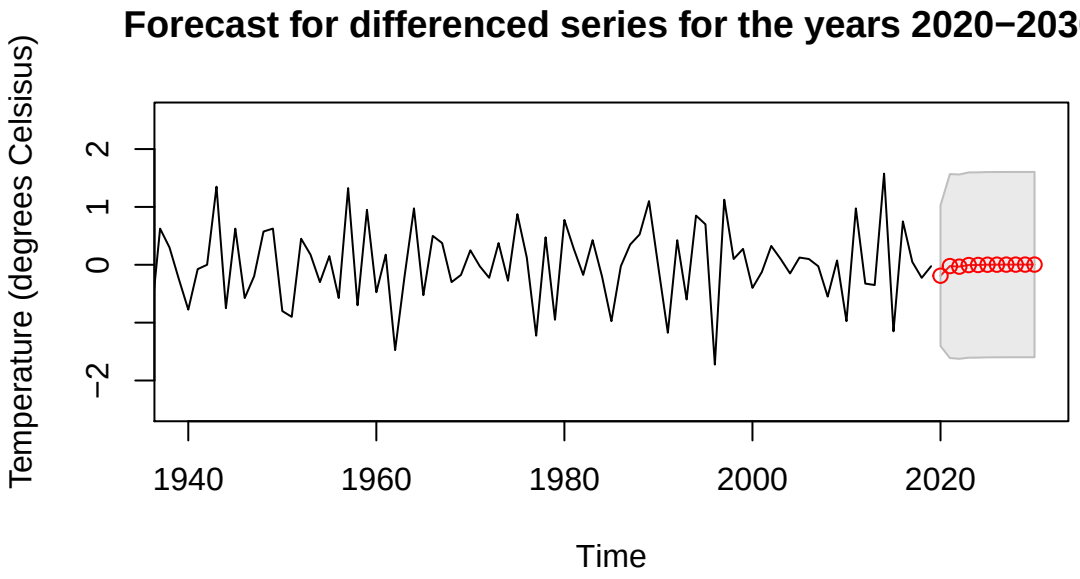
Parameter	Estimate	Standard error
μ	0.0034	0.0023
θ_1	-0.95	0.036
ϕ_1	0.108	0.065
ϕ_2	0.1701	0.063
σ_w^2	0.369	
Ljung-Box statistic	$\chi^2_{26} = 32.3$	$p = 0.18$

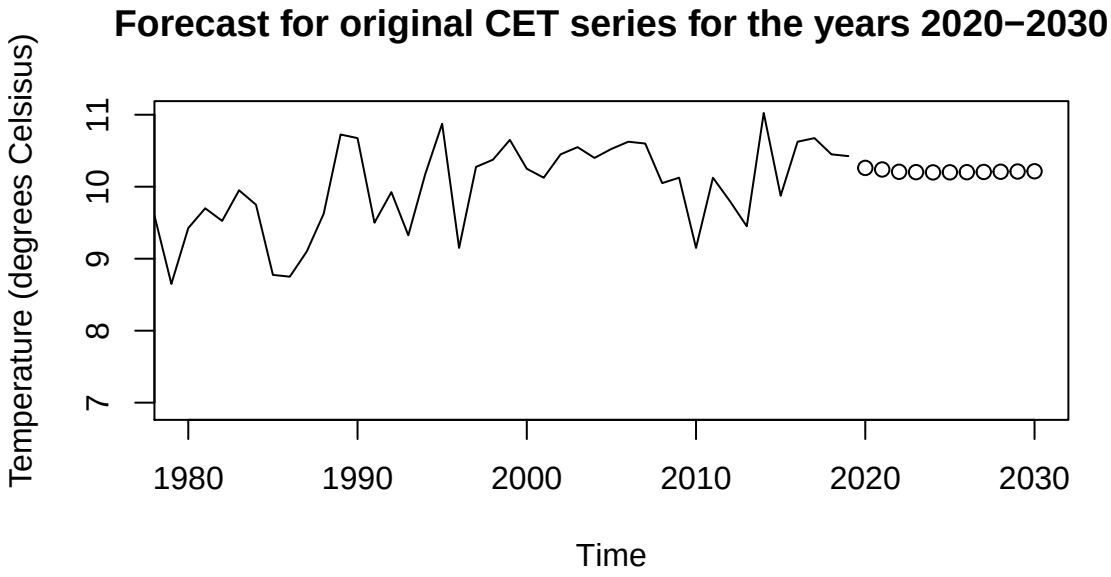
Model predictions on data for future years

To help answer the question of whether or not the temperature of the Earth is currently trending upwards, we can examine our model's predictions on the new few years. This means we will do one step aheady prediction to get $\hat{y}_{2020}, \ldots, \hat{y}_{2030}$, and we can get the estimated temperature $\hat{x}_{2019+k}$ by taking $\hat{x}_{2019+k} + \hat{y}_{2019+k+1}$. Below, we will see a plot of the 10 year forecast[1] for the differences $\hat{y}$ along with their approximate

[1] This is a pretty good pun ...

95% confidence intervals, assuming Gaussian noise which is mean 0. As we established above, this noise assumption is not quite right, but it is sufficient for the purposes here because we are just looking for a general upward trend.



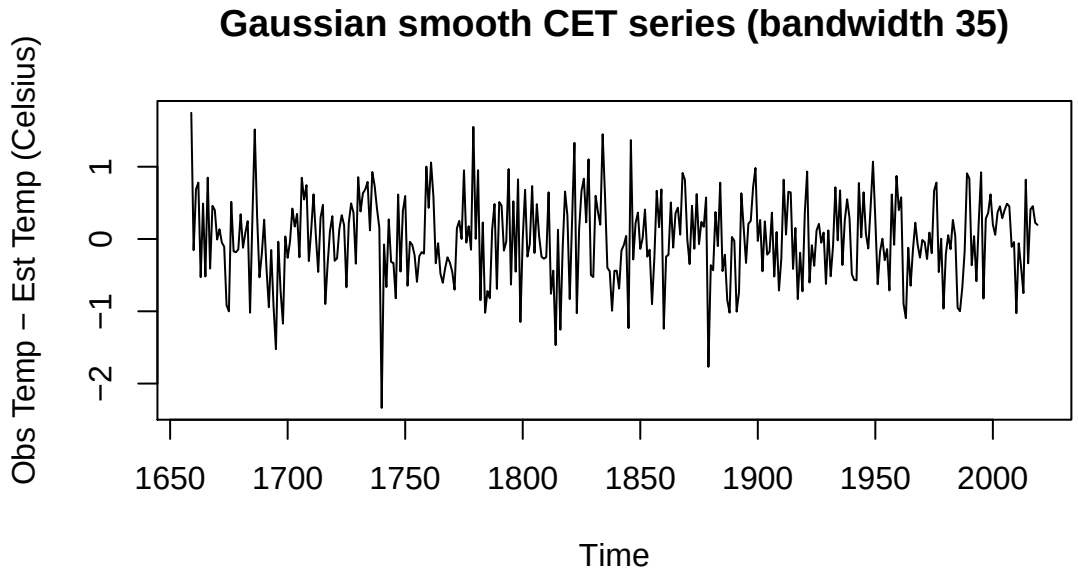


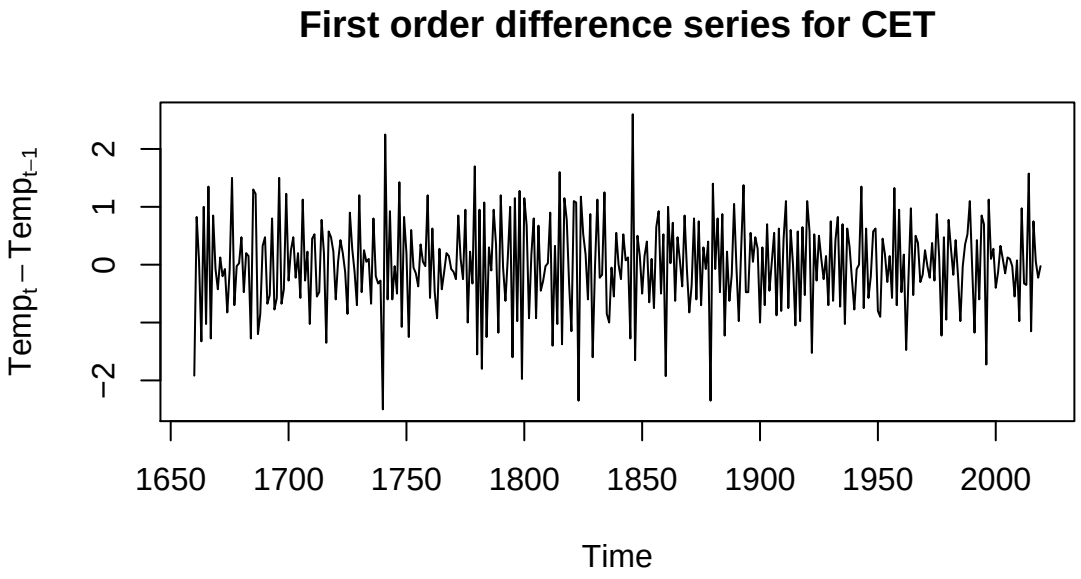
Predicted Difference	Interval for differnce	Predicted temperature (degrees Celsius)
-0.189	(-1.40, 1.03)	10.3
-0.022	(-1.61, 1.57)	10.2
-0.032	(-1.625, 1.56)	10.2
-0.005	(-1.61, 1.60)	10.2
-0.004	(-1.61, 1.60)	10.2
0.001	(-1.60, 1.60)	10.2
0.002	(-1.60, 1.60)	10.2
0.003	(-1.60, 1.60)	10.2
0.003	(-1.60, 1.60)	10.2
0.003	(-1.60, 1.60)	10.2
0.003	(-1.60, 1.61)	10.2

We see that the predictions for the next ten years indicate some upward trend in the global temperature. However, these predicions have large variance, meaning that in reality it may be a different, or more subtle, story. Notably, the confidence intervals for the predictions of the differences contains both positive and negative values, which means that if this model were close to being correct, then it would also be likely to observe some decrease in the global temperature in the next few years. This is in line with current work, which states that the global temperature has continued to rise; however, we can not conclude here that this tend is due to external factors, such as human behavior. To further look into other factors around the global temperature trend, I will look into periodic behavior of the series.

Analysis of periodic trends

This section will be concerned with the analysis of periodic trends in the series. Looking for periodic trends in this series could give us an idea of whether or not the observed increase in global temperature is a fact which would be consistent with the behavior of the whole series. First, I will use the stationary series obtained when removing the trend by Gaussian kernel smooth, shown below. This series is not as rough as the differenced series, and for this reason it will be better to use when trying to identify larger periodic trends.

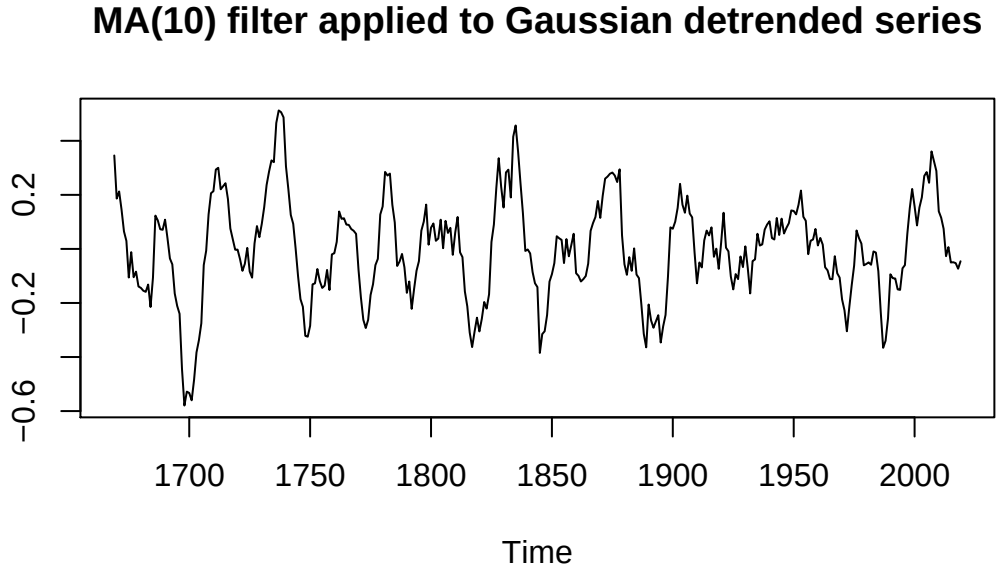




Now, to make emphasize the longer term trends even more, I will apply a moving average filter, as this is a low-pass filter. I will use a Daniell filter with $m = 5$, which means it will be smoothing over decade long windows. This notation for this is shown below. We see that this brings out the longer periodic behavior of the series, and now we will examine that behavior to look for any notable periodic trends.

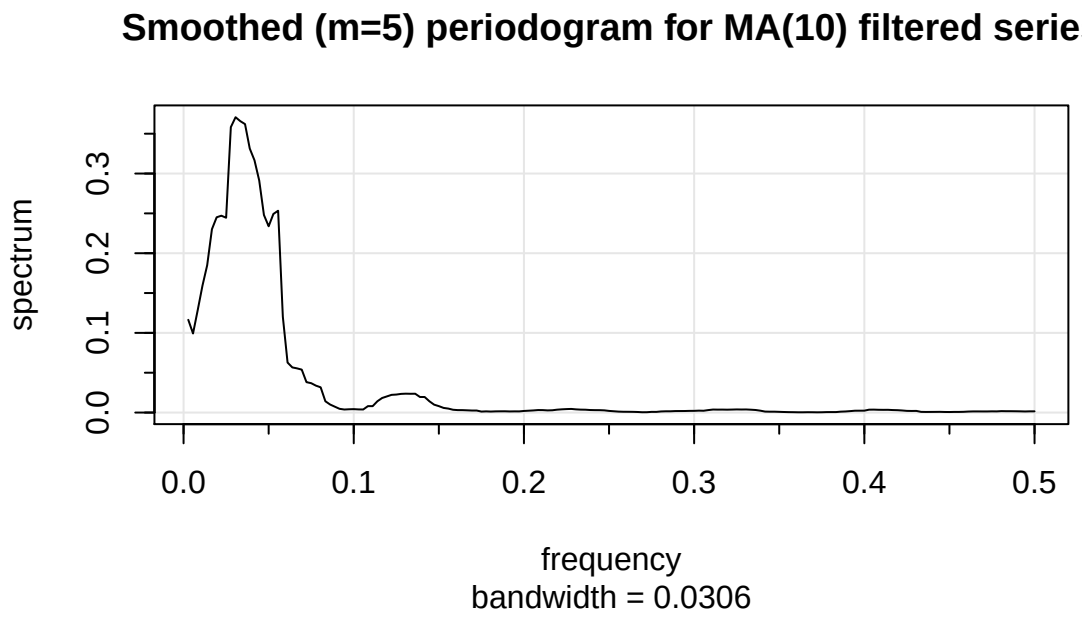
$$y_t = \frac{1}{10} \sum_{k=-5}^{5} x_{t-k}$$

Series filtered by 10 year moving average



The periodogram for the series passed through the linear filter is below. I smoothed the periodogram with

a Daniell filter using $m = 5$. We see that there is a peak at frequency 0.04, which corresponds to a period of 24 years. The estimate of the peak is 0.3167, and the confidence interval for this peak is $(0.19, 0.66)$, using the χ^2_{2L} approximation for $L = 2m$ here. This indicates this peak is important as the lower limit is markely larger than the power for any of the other frequencies.



This result is certainly indicative of the trend in the filtered series: we see a cycle of around every 25 years. This means that for our series detrended by a Gaussian filter, it fluctuates about the trend with a cycle of 25 years.

Discussion and conclusions

We've observed the behavior of the Central England Temperature (CET) data set, and we have seen that the analyses above are consistent with the global temperature rising. The series was not stationary, so in the first part of the analysis, we used the differenced series. Differencing the series made it stationary, and the fact that the series became stationary after differencing suggests that the increase which seems to appear in the CET data is legitimate. I found that fitting an ARIMA(2,1,1) model to the CET data was best in terms of interpretation and it was essentially equivalent in being the minimizer of both AIC and BIC; the model which actually minimized these two criterion was the ARIMA(1,1,2) model. The estimated coefficient for the mean for the differenced series was 0.0034 degrees Celsius, which again indicates the increase in global temperature. The model fit to the data is shown below, where the standard errors are shown in the subscripts.

$$x_t - x_{t-1} = y_t = 0.11_{(0.07)} y_{t-1} + 0.17_{(0.07)} y_{t-2} + -0.95_{(0.04)} w_{t-1} + w_t \qquad \hat{\sigma}^2_w = 0.37$$

I proceeded to do a 10 year forecast of the differenced series, and it led to predictions which did not continue to exhibit an increase in temperature, which may be seen as light evidence towards the claim that the warming of the global temperature is slowing down. The confidence intervals on these predictions, however, were wide enough to admit the possibility of a continual increase in the temperature or even a potential decrease in the coming years.

To further examine the behavior of the series, I considered the periodic behavior by doing a spectral analysis on a detrended and filtered version of the series. I detrended the series with a Gaussian kernel of bandwidth 35, which is consistent with previous literature, and I considered the moving average filter of window size 10 years to examine the low frequencies in the series, which we are more interested in. This analysis showed a significant period of around 24 years. This indicates a cyclic trend in the global temperatures, and when considering the point in time we are at now, this may indicate a potential decrease in the temperature in the next few years. As is currently controversial, the apparent decrease in the warming may be indicative of this cycle.

Differencing the series made it stationary as well as detrending with a function that has an estimated means which has been strictly increasing over the past two decades, and this is highly suggestive of a trend upwards in the temperature. However, we can not be certain of the reason for this increase. If the ARIMA(2,1,1) model is one that closely resembles that of the underlying process, it could be the temperature pattern is similar to a flavor of a random walk. This is a much lighter conclusion than the conclusion other people may want to reach about human behavior affecting the temperature. The temperature does as well seem to have slowed down in the past decade, and this could be part of some underlying periodic behavior in the temperature. These claims are consistent with those made in Benner (1999), Jones and Bradley (1992), and Vaidyanathan (2016).

References

Manley, G. (1974). Central England temperatures: monthly means 1659 to 1973. Quart. J. R. Meteorol. Soc. 100, 389-405.

Jones, P. D. & Hulme, M. (1997). The changing temperature of Central England. In Climates of the British Isles, 460 Present, Past and Future, Ed. M. Hulme and E. Barrow, London: Routledge. pp. 173-195.

Jones, P. D. & Bradly, 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. 246-268.

Benner, T. C. (1999). Central England temperatures: Long-term variability and teleconnections. Int. J. Climatol. 19, 391-403.

Harvey, D. I. & Mills, T. C. (2003). Modelling trends in central England temperatures. J. Forecasting 22, 35-47.

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. 649-65.