Annual Changes in Global Temperature Analysis Using ARIMA Model

SYDE 631: Time Series Modelling

Yuanyuan Gan

SYDE, Master of Engineering

Abstract

Global warming is becoming a very popular topic for a long period of time. Because it is closely related to the future of human being and other creatures living on the Earth. Many researchers and institutions are focusing on this topic and they provide many available time series data and helpful academic results. This project use the data from NASA of annual changes in global temperature. With some background introduction and exploratory data analysis, the main part of this project is confirmatory data analysis based on ARIMA model. The three main steps of model construction: model identification, parameter estimation and diagnostic check are detailed presented. An ARIMA(1,1,3) model is carefully selected and fitted . After model construction, two brief applications: forecasting and simulation are used to ensure the robustness of model.

1. Introduction

1.1 Motivation and importance

Is world getting warmer? Many researchers and experts tell us the world is getting warmer, some insist human activity causes this while others not completely support this point of view. Whether this is caused by the human activity, we can feel from many aspects that the world climate is changing. One simple and direct evidence is the average globe temperature.

The Earth exists for 4.54 billion years in the Solar System which is even older than the Earth. It will rotate continually whether the global temperature increases or decreases. But for the creatures living on this planet, especially us human being, one-degree of temperature change will cause dramatic impacts. The entire living system will change significantly. In the past, a one-to two-degree drop was all it took to plunge the Earth into the Little Ice Age. A five-degree drop was enough to bury a large part of North America under a towering mass of ice 20,000 years ago.

We have seen many environmental changes in just few decades, for instance, Arctic sea ice decline, sea level rise, retreat of glaciers, extreme weather, tropical cyclones, ecosystem changes, changes in ocean properties and many others changes. One sad example is the polar bear population decline. Nowhere is the warming of our planet more apparent than in the Arctic, where the sea ice polar bears depend on is melting. This loss of multi-year sea ice is transforming the region and affecting every facet of the polar bear's life, from hunting seals to raising cubs. There are now fewer than 20,000 polar bears left on Earth. Although from the biological evolution view, with low reproductive rate, selective diet of ice-dependent seals and specialized sea-ice habitat, polar bears already live on the edge. Climate change is still pushing them over it. Global polar bear numbers are projected decline by 30% by 2050, and we may be the last generation witness this species existing on the Earth.

What should we do? As a huge challenge for all human being to face, we can almost do nothing. And this is true. Despite the politics, even the technology of modern science which many people believe in will help a little when facing the natural power. However, as an optimist, I believe if we do something now, though the global warming will not stop, it

may slow down, and global climate may be stabilized one day.

1.2 Some background

To complete this project, basic knowledge is based on the course textbook: Time Series Modelling of Water Resource and Environmental Systems, from chapter 1 to chapter 9 and chapter 22 to chapter 24. Background knowledge is from literal review of recent research papers in environmental area and webpage search engine. To better understand and apply time series modelling, I also read some blog articles from website, and these sources are listed in the reference.

The software using for this project is Jupyter 5.0.0, the programing language is Python. The main statistical time series calculation is using the package statsmodels, it provides classes and functions of estimation of many different statistical models. This project is mainly using statsmodels.tsa.arima_model.ARIMA class, which provides functions to construct model, estimate parameters, plot time series figures and etc. And other packages such as numpy, panda, scipy, matplotlib are also used.

1.3 Data sources

The data is from NASA's Goddard Institute for Space Studies (GISS), Credit: NASA/GISS, on the webpage of NASA global climate change. The data is the change in global surface temperature relative to 1951-1980 average temperatures. The latest annual average is 0.9 centigrade of 2017.

1.4 Types of models

The model using in this project is ARIMA model. ARIMA is applied in this case because the data show evident non-stationarity. No seasonal part is needed. And here one step of differencing is applied. The autoregressive part indicates that the evolving variable of interest is regressed on its own lagged, and moving average part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. Both AR part and MA part are involved here in this project of data analysis.

1.5 Overview of upcoming sections

In next section, I will give a brief introduction of the research background of this area. To analysis the data, there are two main stages, exploratory analysis and confirmatory analysis separately in section 3 and section 4. In exploratory analysis, an overall analysis without model and parameters is given, and using rolling-mean to have a general understanding of trends. In confirmatory analysis, three steps are used. And Box-Cox transformation is also introduced and applied in this part of project. Then I will give two quick examples of application of the model: forecast and simulation.

2. Background

As a very popular investigation area, many researchers have attempted to model the global temperature using variety methods. In the research paper Global Temperature Trends (Breusch, 2011), researchers are using different statistical methods to represent the trend of overall global temperature. The analysis in that paper shows that the upward movement over 130-160 years is persistent and not explained by the high correlation. The warming trend becomes steeper after the mid 1970s, but there is no significant evidence for a break in the trend in the late 1990s.

The global temperature mainly depends on how much energy the planet receives from the Sun and how much it radiates back into the space. But the most important influencing factor is the chemical composition of the atmosphere, particularly the amount of heat-trapping greenhouse gases. The primary greenhouse gases in earth's atmosphere are water vapor, carbon dioxide, methane, nitrous oxide and ozone. Carbon dioxide levels are increasing in the atmosphere. Human activities since the beginning of the Industrial Revolution have produced a 40% increase in the atmospheric concentration of carbon dioxide, from 280 ppm in 1750 to 406ppm in early 2017. Increased carbon dioxide is the primary driver of global warming. The last three decades of temperature increasing may have a causation of aerosol which effects the greenhouse gases. Some researchers find that the global temperature can be evaluated by the particular greenhouse gas emissions, such as explained in the research paper Evaluating Global Warming Potentials with historical

temperature (Tanaka, 2009).

The temperature fluctuates from region to region. Although the earth is getting warmer, it doesn't mean temperatures rose everywhere on the earth. While the global temperature is the mean of different regions, the relationship between the mean and the standard deviation is also an important research aspect. The results from the Coupled Model Intercomparison Project phase and multiple global reanalysis datasets are used to investigate the relationship between the mean and standard deviation in the surface air temperature at intra- and inter-annual timescales in the paper Surface Air Temperature Variability in Global Climate Models (Davy, 2013).

The uncertainty in global temperature changes also attracts researchers' attention. The paper Global Temperature Change and Its Uncertainty (Folland, 2001) presented the analysis of global and hemispheric surface warming trends to quantify the major sources of uncertainty. A Bayesian statistical model developed to produce probabilistic projections of regional climate change using observations and ensembles of general circulation models is applied to evaluate the probability distribution of global mean temperature change in the paper Two Approaches to Qualifying Uncertainty in Global Temperature Changes (Lopez, 2005).

From the records and evaluations, though some variations in the data obtaining, all show rapid warming in the past few decades, and the last decade was the warmest. And some predictions of future global temperature claim that the increasing will maintain until year 2050s.

Many researches about global temperature are related to other environmental aspects, for instance the greenhouse gas evaluation mentioned above, sea surface temperature, climate change and so on. But in this research project, the analysis of global temperature only focuses on the mean temperature data itself. If the temperature data can be properly modeled, many characteristics can be detected and analysist. And it can also be transferred to use in other environmental situations.

3. Exploratory data analysis

To explore the data, first to do is to look at the data itself. The data here is in the range

from year 1880 to 2017 with no missing values. The error of measurement is not clear in this situation. No known interventions have been explicitly proved influencing the data.

The simplest but very useful method of exploratory data analysis is to plot the data against time. It is shown some basic characteristics in the figure. The lowest value is -0.43 in the year 1917, and the highest value is 0.99 of 2016, range over one-degree around 1.5 degree. The mean value seems to increase over time, and the trend of data will be analysis in the later part of exploratory data analysis and also in confirmatory data analysis. It seems a quick increasing in the 1930s and around 2000, while it also could be white noise. There is no obvious variance changing by just looking at the plot. The data shows non-seasonality and no long-term cycles. No extreme value was detected. It's hard to decide whether the data should be transformed in this plot, while Box-Cox transformation is used in the confirmatory data analysis.

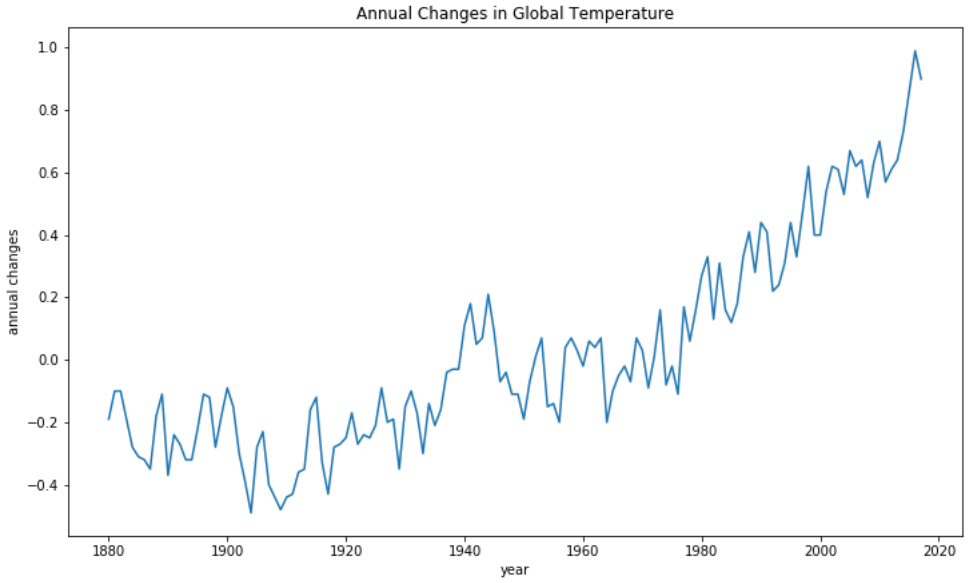

Figure 1 Annual changes in global temperature from 1880 to 2017

To explore the data trends, one way is plot moving average. Moving average also called rolling mean and running average. It is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is commonly used in time series data to smooth out short-term fluctuations and highlight longer-term trends.

And a weighted average is an average that has multiplying factors to give different weights to data at different positions in the sample window. Here using window at size 20 to calculate the moving average and weighted moving average. It is shown in figure 2, a clear overall trend of increasing. The mean value curve seems flatten during 1880 to 1900 and 1945 to 1975, a concave around 1915, and an evident linear increase after 1980.

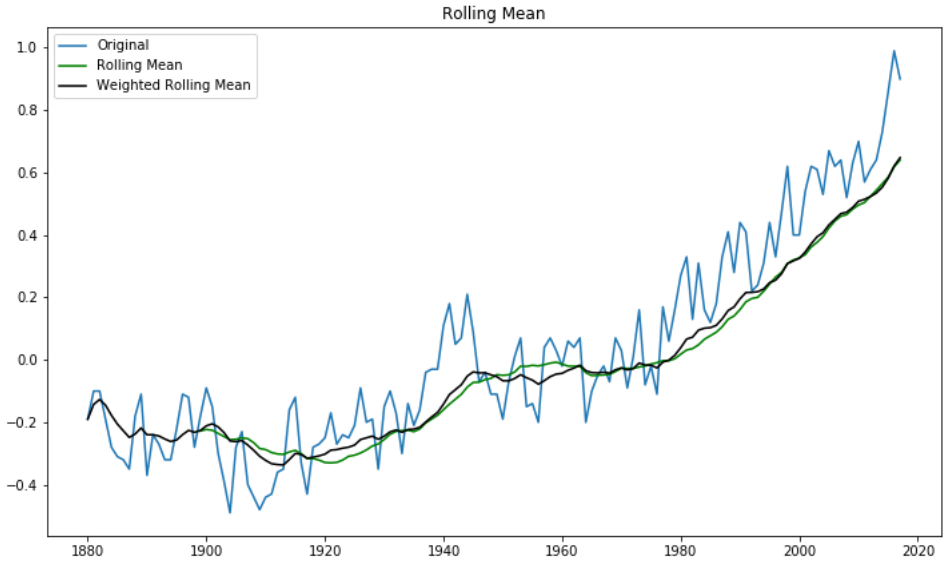

Figure 2 Trends detection of annual changes in global temperature

4. Confirmatory data analysis

After exploratory data analysis, confirmatory data analysis is to confirm statistically in a rigorous fashion the presence or absence of certain properties in the data. This part consists of model identification, Box-Cox transformation, model estimation and diagnostic check.

4.1 Box-Cox transformation

To make sure the final model to have a nicely property of normality and constant variance, use Box-Cox transformation to the origin data. Base on this particular time series data, the magnitude of y is inside (-1,1), so set 1 as the constant parameter c. Then use log-likelihood function to set the parameter λ. The Box-Cox log-likelihood function is defined

here:

$$llf = (\lambda - 1)\sum_i(\log(x_i)) - N/2\log(\sum_i(y_i - \bar{y})^2/N),$$

where y is the Box-Cox transformation input x. To choose the λ with maximum log-likelihood, plot the λ from -2 to 5 while default number is 100. See the figure. The maximum result is located at λ=-0.37. Then the following model construction and calculation will use the data transformed by Box-Cox transformation with λ=-0.37 and c=1.

Note that the Box-Cox transformation will not eliminate the need for differencing, or change the property of non-stationarity. The sample ACF of original data and transformed data are showed in the next part 4.2 Model identification, and it is clear to see that the transformation will keep the property of data.

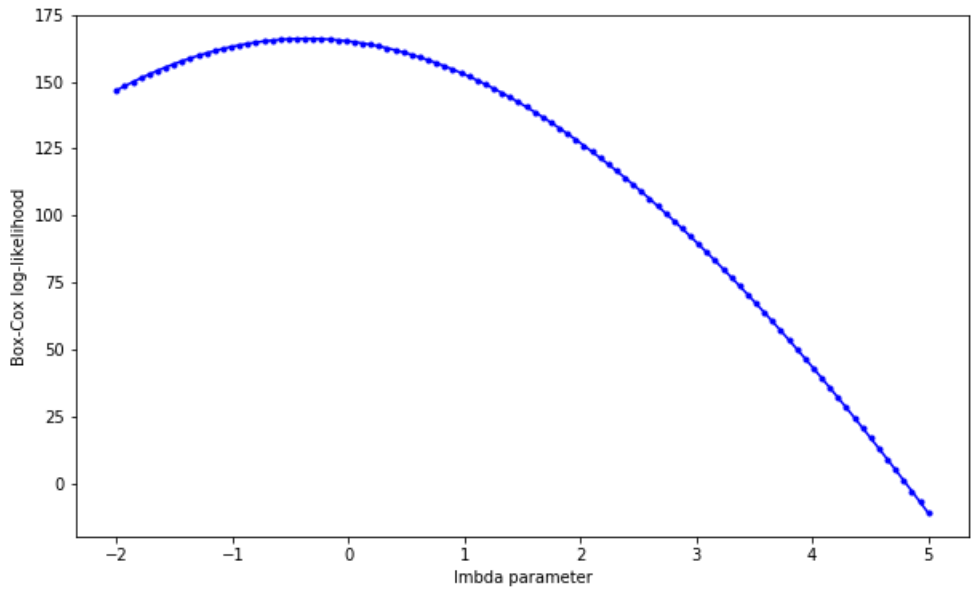

Figure 3 llf of Box-Cox transformation

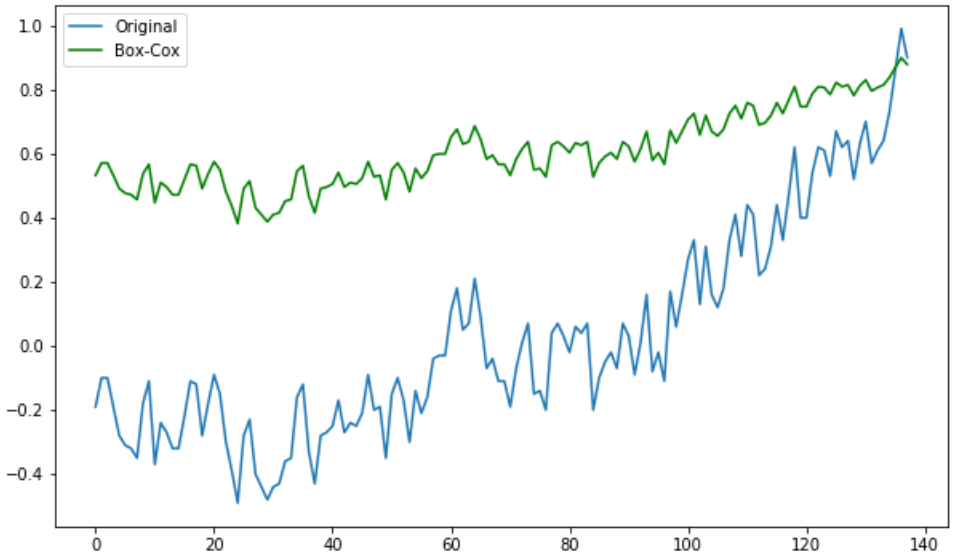

Figure 4 Data with Box-Cox transformation vs Original data

4.2 Model identification

Plot the sample ACF of original data and data after Box-Cox transformation from lag 0 to lag 40 (around 1/4 length of time series data) with 95% confidence limits. The figure shows that the sample ACF possesses large values at lower lags and slowly attenuate for increasing lag. The sample ACF does not die out quickly for larger lags, this indicate that the data should be differenced to remove homogeneous nonstationary.

After differencing once, the data is plotting in the figure. The differencing data seems has no trends and the covariance seems maintain constant over time. The figure of the sample ACF and 95% confidence limits shows the non-stationarity is removed, which means an ARIMA(0,1,0) model should be used. The large values of the first several lags in sample ACF indicate the need for MA parameters in the model. Then plot the sample PACF of data with differencing once. It is shown in the figure that some AR parameters are also needed. Both sample ACF and sample PACF show die off, indicates after differencing ARMA model with both AR part and MA part is needed here.

It is quite difficult to say how many AR or MA parameters may best represent the

model just by looking at sample ACF and sample PACF. Because no clearly truncate in either sample ACF or PACF, and the values are fluctuating at first several lags. The way to discriminate the number of parameters will show in the next section 4.3 Model estimation.

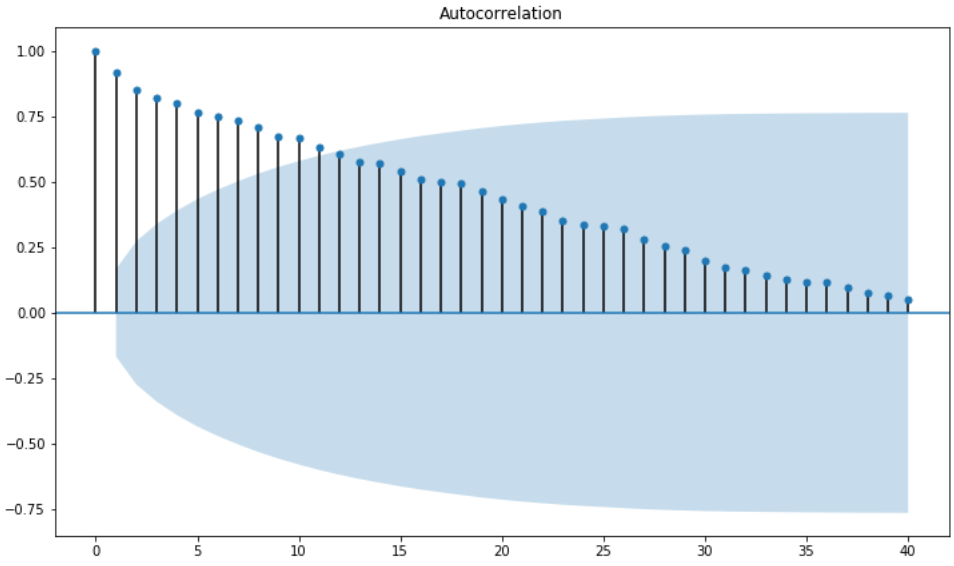

Figure 5 sample ACF of original data

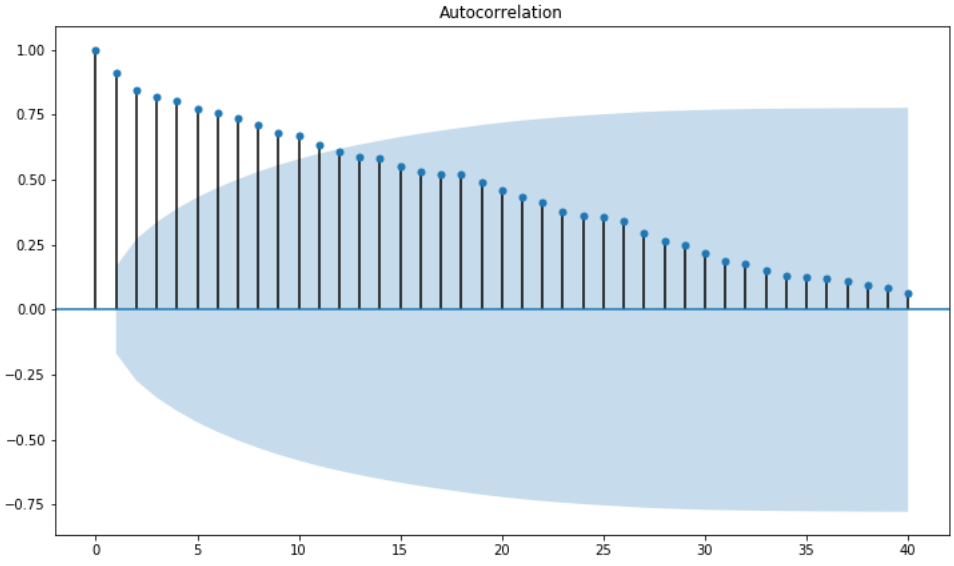

Figure 6 sample ACF of data after Box-Cox transformation

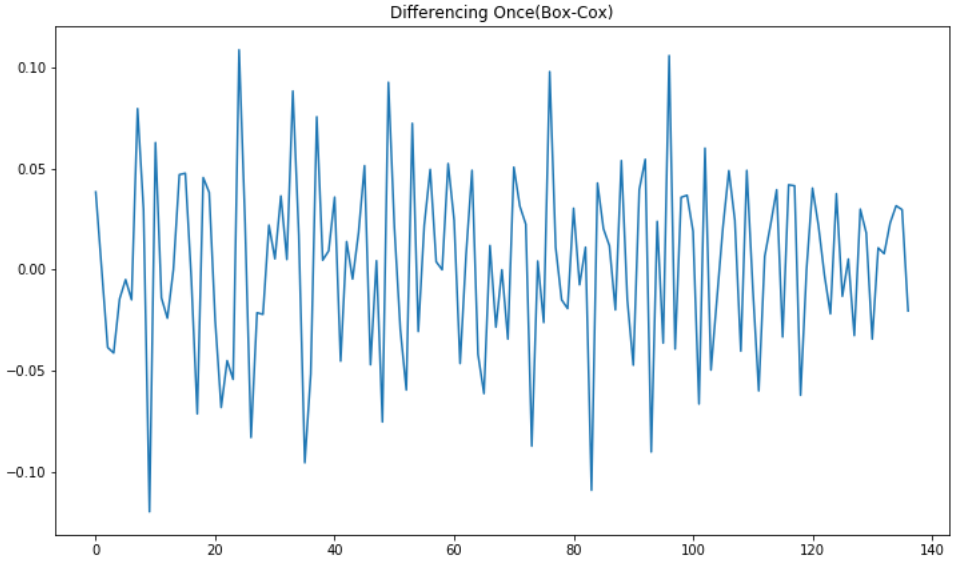

Figure 7 Data (with Box-Cox transformation) differencing once

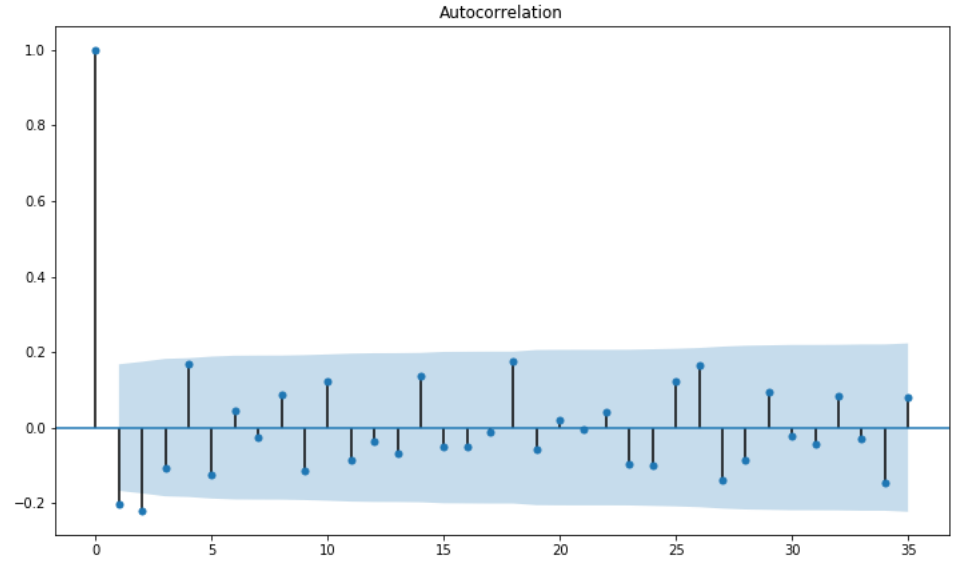

Figure 8 sample ACF of data (with Box-Cox transformation) after differencing once

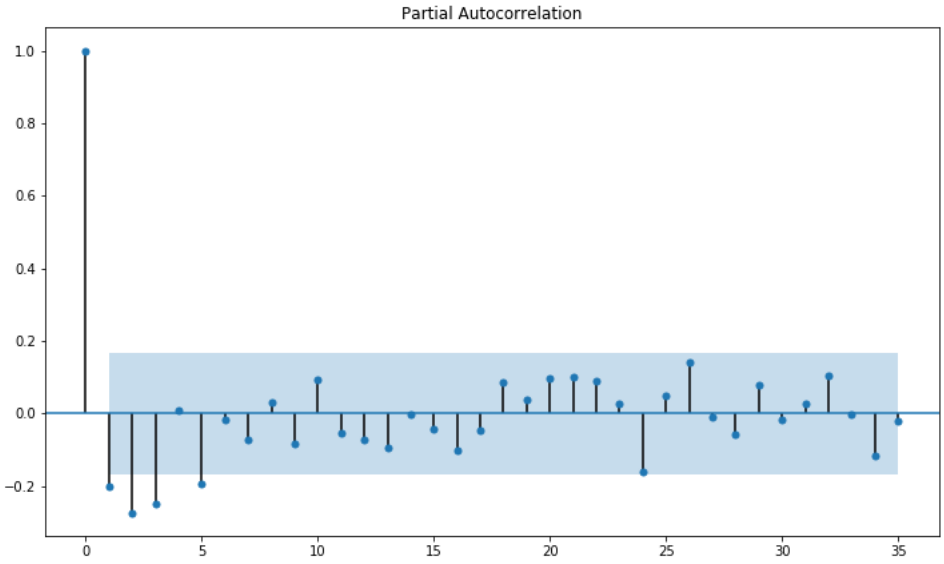

Figure 9 sample PACF of data (with Box-Cox transformation) after differencing once

4.3 Model estimation

Before fitting the model use maximum likelihood estimation, one important thing is to confirm the selected model. As I explained above, it is hard to make sure the exact number of parameters by just looking at the sample ACF and PACF, another discrimination method AIC is used here to ascertain the model. Calculating AIC of ARIMA model with different combination of AR and MA parameters. Table 1 shows the AIC of model under different transformation hyper-parameters. It is cheerful to see that for different value of transformation parameter λ, the minimum AIC is located in the ARIMA(1,1,3), and this may insist to fit the data using ARIMA(1,1,3) is a good choice.

Then using MLE to estimate model via Kalman filter. The Kalman filter is an efficient recursive filter that estimates the internal states of linear dynamic system from a series noisy measurements. It is used in a wide range of engineering and economic applications. A brief overview of estimated model is given in table 2, and the details are in Appendix.

ARIMA(p,d,q)	aic(λ=-0.37, c=1)	aic(λ=0, c=1)
(0 1 2)	-480.3559	-413.2747
(0 1 3)	-479.4559	-412.2129
(1 1 1)	-479.4480	-412.3123
(1 1 2)	-478.9584	-411.7826
(1 1 3)	-483.4551	-416.2474
(2 1 1)	-479.4111	-412.2248
(3 1 1)	-477.4262	-411.6834
(4 1 1)	-476.2852	-408.8490
(5 1 0)	-479.3795	-412.4514
(5 1 1)	-477.5612	-410.7904
(5 1 2)	-476.1324	-408.6807
(5 1 3)	-475.8898	-408.8087
(6 1 1)	-476.0991	-409.1043
(6 1 3)	-476.1873	-406.7821
(7 1 0)	-476.2023	-409.2379
(7 1 1)	-475.8020	-409.0001
(7 1 2)	-473.9002	-407.5915

Table 1 AIC values for the ARIMA model

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0025	0.001	2.132	0.035	0.000	0.005
ar.L1.D.y	-0.9275	0.067	-13.860	0.000	-1.059	-0.796
ma.L1.D.y	0.6144	0.100	6.168	0.000	0.419	0.810
ma.L2.D.y	-0.5954	0.087	-6.841	0.000	-0.766	-0.425
ma.L3.D.y	-0.3731	0.079	-4.752	0.000	-0.527	-0.219

Table 2 Estimations of the ARIMA(1,1,3) model

4.4 Diagnostic checking

4.4.1 Overfitting

Base on the Occam's Razor theory, the model parsimony is a very important property. AIC is a good way to select a suitable model. After a model being chose using the smallest AIC and then estimating parameters, one other thing must be done is to check whether the model is overfitting. If overfitting exists, either delete some parameters and re-estimate the model or choose another model.

To see whether a model is overfitting, one way is to see the estimations of the parameters and their standard errors SE's. If some parameter estimates are very close to zero and their SE's are larger than the estimates, these parameters should be considered to reduced. If the estimate is too small, which means this parameter may affect the model very slightly. The SE's present the property of whether this is a noise. In this model, parameter estimates and their SE's are listed in the table, no parameter estimates are close to zero. The absolute values of estimates are much larger than SE's, and every estimate is more than three times the value of its SE, which indicates that every parameter in this model is absolutely necessary.

	MLE's	SE's
$\phi 1$	-0.9275	0.067
$\psi 1$	0.6144	0.100
$\psi 2$	-0.5954	0.087
$\psi 3$	-0.3731	0.079

Table 3 Estimated AR and MA parameters

4.4.2 Whiteness test

Whiteness test is the most important diagnostic check. If the model doesn't satisfy whiteness test, then another model should be considered. One good way is to plot the residual autocorrelation function (RACF). And the plot of RACF of fitted model shows that there is no evident correlation after lag 0, which indices the model fitted is satisfied with whiteness test.

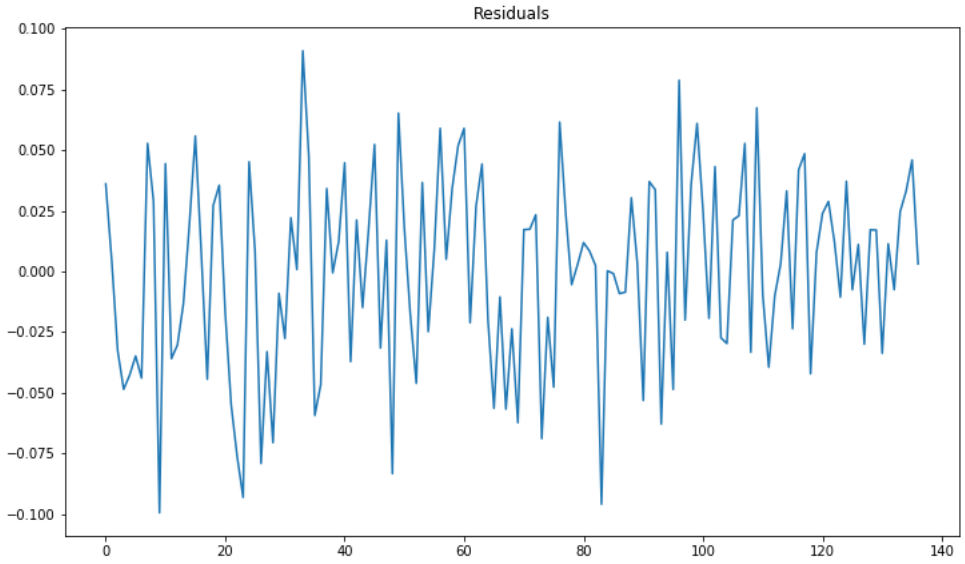

Figure 10 Residuals of estimated ARIMA(1,1,3) model

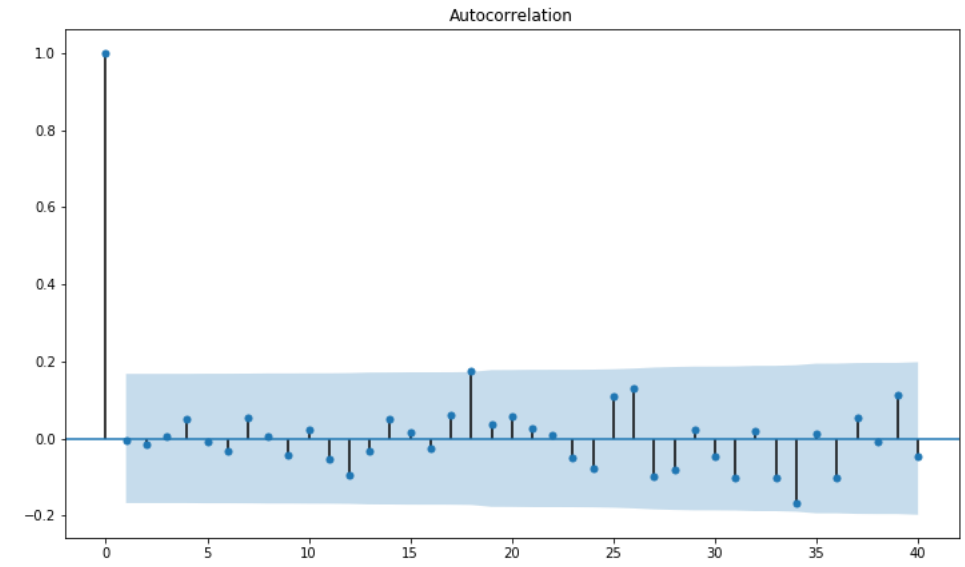

Figure 11 RACF of fitted model

To ensure the test more convincing, another method is also used to test whiteness. Durbin-Watson test is a test used to detected autocorrelation. The null hypothesis of the test is that there no series correlation. The Durbin-Watson test statistics is defined as:

$$\sum_{t=2}^{T}((e_t - e_{t-1})^2) / \sum_{t=1}^{T} e_t^2$$

The test statistic is approximately equal to 2*(1-r) where r is the sample autocorrelation of the residuals. Thus, for r == 0, indicating no serial correlation, the test statistic equals 2. This statistic will always be between 0 and 4. The closer to 0 the statistic, the more evidence for positive serial correlation. The closer to 4, the more evidence for negative serial correlation. And the output of fitted model is 2.0028, is very close to 2, indicates no correlation of residuals.

4.4.3 Normality test

Here using Q-Q plot to test the residuals' normality. Q-Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. In using a normal probability plot, the quantiles one uses are the rankits, the quantile of the expected value of the order statistic of a standard normal distribution. To see the Q-Q plot of residuals, almost all the scatters lay on the diagonal and others are very close to the diagonal line. Most are grouped in the middle of the plot. This will verify that the residuals are almost normal distribution.

After several diagnose tests, the model shows good performance. No change of model or transformation of data is needed. This model fits this particular time series data pretty well.

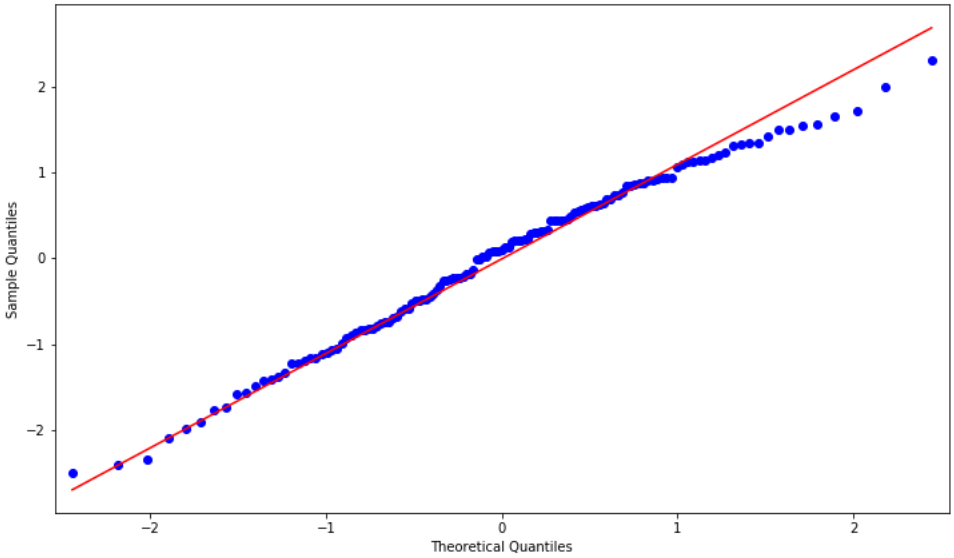

Figure 12 Q-Q plot of residuals

5. Applications and insights

5.1 Forecast

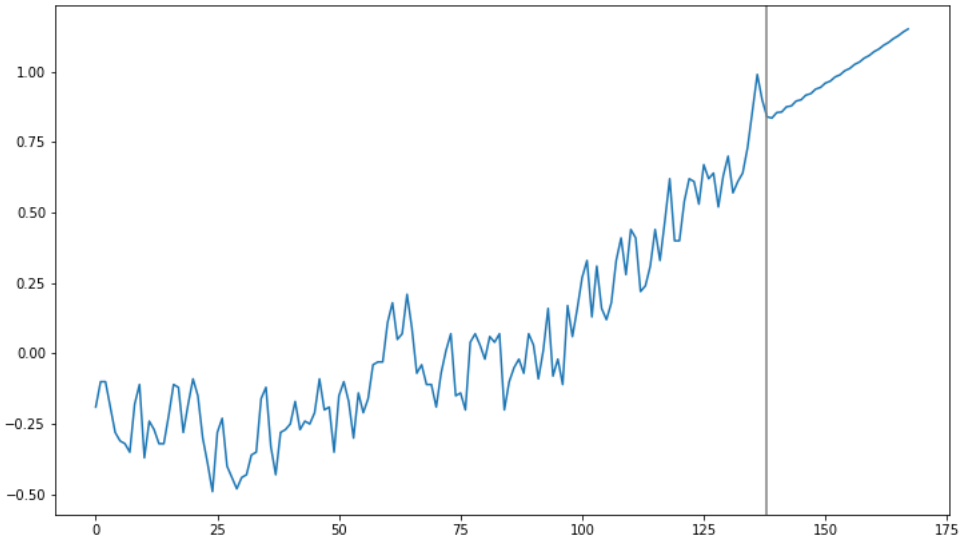

Figure 13 Forecast for 30 years

Using one step ahead forecast to forecast next 30 years' data. The forecast data preserve the trend of original data. Clear away the interference of white noise, the forecasting values show a small wave of zig-zag shape while increasing. The slop of forecasting value seems a little smaller than last part of original data. While forecasting, the standard errors of forecasts are also calculated, and attached in the Appendix.

5.2 Simulation

Using simulation sequences can help to ensure the robustness of fitted model. And simulated value will also preserve the properties of time series data. The random generator using to simulate here is built-in method of randn, it will generate an array of random floats samples from a univariate "normal" (Gaussian) distribution of zero mean and unit variance.

Using the built-in function and fitted values of ARIMA(1,1,3) model parameters, one example of simulated values is plotting here. And also reconstructed the data with the inverse of differencing. The reconstructed data of simulation seems to have the same pattern of original time series data, while the magnitude is quite larger.

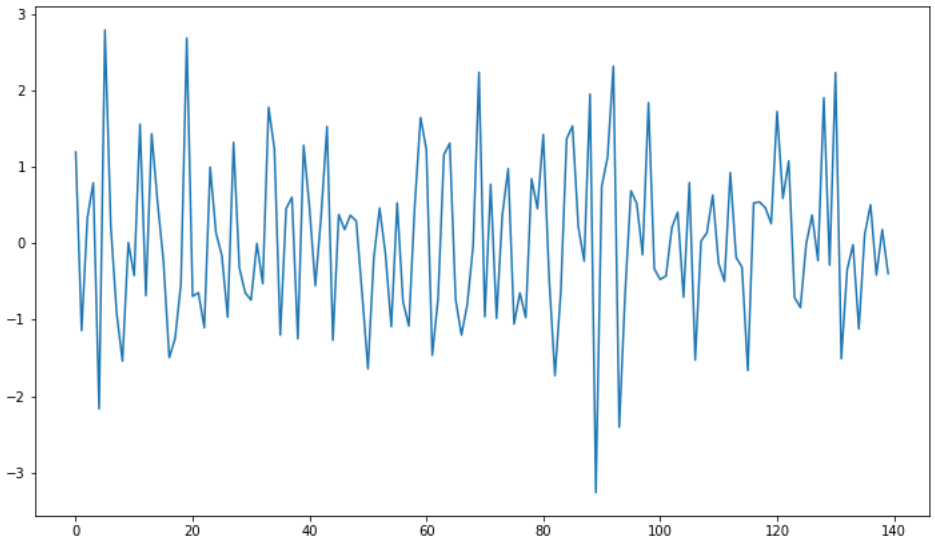

Figure 14 Simulation of ARMA(1,3) model with parameters fitted to globe temperature

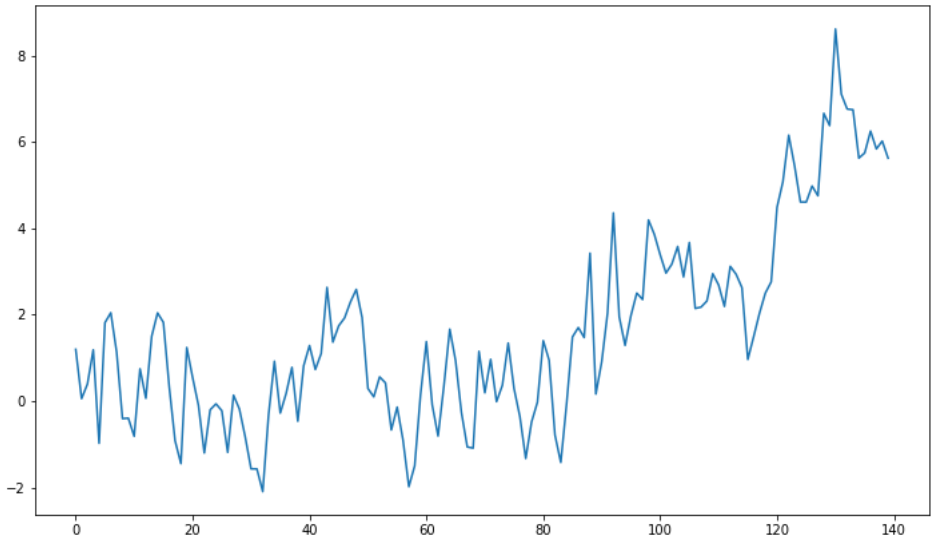

Figure 15 Data constructed with simulated values

6. Conlusion

This project is mainly focus on the model construction. After carefully model selection, estimation and diagnostic check, an ARIMA(1,1,3) is fitted well to annual changes of global temperature. The coefficient d in ARIMA(p,d,q) is equal to one, indices the non-stationarity in the data. The number of parameters is not large. This model is simple but preserve the main properties of data.

While there are also some limitations in this project. The works after model construction needs further analysis. The model constructed based on ARIMA model while is there any other models can better represent the properties is not sure. And this data analysis very relies on the calculation inside the software and program package, so it may lead different results when using different program language and packages. Despite these, this project provides a detailed example of ARIMA model usage and global temperature analysis. And the ARIMA(1,1,3) fits the data well.

Appendix

1. Details of fitted ARIMA(1,1,3) model

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0025	0.001	2.132	0.035	0.000	0.005
ar.L1.D.y	-0.9275	0.067	-13.860	0.000	-1.059	-0.796
ma.L1.D.y	0.6144	0.100	6.168	0.000	0.419	0.810
ma.L2.D.y	-0.5954	0.087	-6.841	0.000	-0.766	-0.425
ma.L3.D.y	-0.3731	0.079	-4.752	0.000	-0.527	-0.219

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.0781	+0.0000j	1.0781	0.5000
MA.1	1.2903	+0.0000j	1.2903	0.0000
MA.2	-1.3706	+0.0000j	1.3706	0.5000
MA.3	-1.5156	+0.0000j	1.5156	0.5000

2. Standard error of forecasting values

```
[0.03956266 0.0479968  0.05031855 0.05162476 0.05371809 0.05499441
 0.05690645 0.05815497 0.05991991 0.06114197 0.06278542 0.06398196
 0.06552355 0.06669538 0.06815048 0.06929829 0.07067914 0.07180361
 0.07312008 0.07422189 0.07548204 0.07656188 0.07777236 0.07883092
 0.07999729 0.08103527 0.08216219 0.08318029 0.0842717  0.08527063]
```

3. Original Data

year	data	year	data	year	data
1880	-0.19	1930	-0.15	1980	0.27
1881	-0.1	1931	-0.1	1981	0.33
1882	-0.1	1932	-0.17	1982	0.13
1883	-0.19	1933	-0.3	1983	0.31
1884	-0.28	1934	-0.14	1984	0.16
1885	-0.31	1935	-0.21	1985	0.12
1886	-0.32	1936	-0.16	1986	0.18
1887	-0.35	1937	-0.04	1987	0.33
1888	-0.18	1938	-0.03	1988	0.41
1889	-0.11	1939	-0.03	1989	0.28
1890	-0.37	1940	0.11	1990	0.44
1891	-0.24	1941	0.18	1991	0.41
1892	-0.27	1942	0.05	1992	0.22
1893	-0.32	1943	0.07	1993	0.24
1894	-0.32	1944	0.21	1994	0.31
1895	-0.22	1945	0.09	1995	0.44
1896	-0.11	1946	-0.07	1996	0.33
1897	-0.12	1947	-0.04	1997	0.47
1898	-0.28	1948	-0.11	1998	0.62
1899	-0.18	1949	-0.11	1999	0.4
1900	-0.09	1950	-0.19	2000	0.4
1901	-0.15	1951	-0.07	2001	0.54
1902	-0.3	1952	0.01	2002	0.62
1903	-0.39	1953	0.07	2003	0.61
1904	-0.49	1954	-0.15	2004	0.53
1905	-0.28	1955	-0.14	2005	0.67
1906	-0.23	1956	-0.2	2006	0.62
1907	-0.4	1957	0.04	2007	0.64

1908	-0.44	1958	0.07	2008	0.52
1909	-0.48	1959	0.03	2009	0.63
1910	-0.44	1960	-0.02	2010	0.7
1911	-0.43	1961	0.06	2011	0.57
1912	-0.36	1962	0.04	2012	0.61
1913	-0.35	1963	0.07	2013	0.64
1914	-0.16	1964	-0.2	2014	0.73
1915	-0.12	1965	-0.1	2015	0.86
1916	-0.33	1966	-0.05	2016	0.99
1917	-0.43	1967	-0.02	2017	0.9
1918	-0.28	1968	-0.07		
1919	-0.27	1969	0.07		
1920	-0.25	1970	0.03		
1921	-0.17	1971	-0.09		
1922	-0.27	1972	0.01		
1923	-0.24	1973	0.16		
1924	-0.25	1974	-0.08		
1925	-0.21	1975	-0.02		
1926	-0.09	1976	-0.11		
1927	-0.2	1977	0.17		
1928	-0.19	1978	0.06		
1929	-0.35	1979	0.16		

Reference

1. K.W. Hipel and A.I. Mcleod. (1994). *Time Series Modelling of Water Resources and Environmental Systems.*

2. Trevor Breusch and Farshid Vahid. (2011). *Global Temperature Trends*.

3. Chris K. Folland, Daniel Erik Everhart and P.D. Jones. (2001). *Global Temperature change and Its Uncertainties since 1861.*

4. Richard Davy and Igor Esau. (2013). *Surface Air Temperature Variability in Global Climate Models*.

5. Ana Lopez, Claudia Tebaldi and Mark New. (2006). *Two Approaches to Quantifying Uncertainty in Global Temperature Changes.*

6. Katsumasa Tanaka, Dmitry Rokityanskiy, Richard S. J. Tol. (2009). *Evaluating Global Warming Potentials with historical temperature.*

7. James Hansen, Makiko Sato, Reto Ruedy, Ken Lo, David W. Lea, and Martin Medina-Elizade. (2006). *Global Temperature Change.*

8. Marianna G. Shepherd, Wayne F. J. Evans, G. Hernandez, Dirk Offermann and Hisao Takahashi. (2004). *Global Variability of Mesospheric Temperature: Mean Temperature Field.*

9. Search engine. (2018). https://climate.nasa.gov/vital-signs/global-temperature.

10. Search engine. (2018). https://earthobservatory.nasa.gov/WorldOfChange/DecadalTemp.

11. Search engine. (2018). https://www.cnblogs.com/foley/p/5582358.html.

12. Search engine. (2018). https://blog.csdn.net/u010414589/article/details/49622625.