Statistics 199 Final Report

Edward Shiang

Contents

1 Introduction **1**

1.1 Background 1

1.2 Objectives 1

2 Results **3**

2.1 Time Series Terminology 3

2.2 Time Series Model Construction and Selection 4

2.3 Hypothesis I's Application To Current Annual GISTEMP Data 9

2.4 Bootstrap Temporal Dependence Algorithm Results Automation Script 11

3 Conclusion **12**

3.1 Reflection 12

3.2 Works Cited 13

3.3 Python Code Links 13

1 Introduction

1.1 Background

To preface this report, I would like to discuss why I chose the intersection of climate change and time series as my deep-dive topic under Mike this quarter.

If Mike asked me about potential topics I would be interested in delving into for Statistics 199 a year ago, climate change would not have even crossed my mind. However, 2019 has been different. Having jeopardized things like UCLA classes I was enrolled in and even job security within my family, I felt climate change was an appropriate subject to further investigate.

On another note, I was bummed out that I was unable to take either of the Python and time series analysis statistics electives this fall; the combination of learning time series analysis with climate data using Python through textbook readings, weekly check-ins, and Mike's research paper appeared to be the perfect opportunity to develop my statistical and programming skills while learning more about this domain.

1.2 Objectives

With heavy influence from the motivation of this project described in the previous section, there were four main objectives for the quarter:

1) Learn basic time series concepts and models with textbook readings.

2) Apply the knowledge through time series model implementation and selection in Python.

3) Explore various methodologies employed in "Debunking Climate Hiatus," a research paper Mike coauthored, with current data on the "hiatus."

4) Leave something tangible and useful for Mike and other members of the statistics community for the future.

The following results section has been respectively partitioned into work completed towards each of these goals. Beside 1), a Python notebook or script was compiled containing code written for each section respectively. Each has a GitHub link located at the last section of the paper, *Section 3.3*.

2 Results

2.1 Time Series Terminology

Note: explanations are summaries of definitions given by Robert Sumway, David Stoffer, and Michael Tsiang

Time series analysis exists as a helpful form of tackling statistical data cases which have variables of high correlation among observations due to being recorded at various points in time (thus trespassing upon the independent/identically distributed assumptions that make up a substantial amount of other statistical methods).

Time series: Group of random variables arranged by time recorded.

White noise: Group of uncorrelated random variables.

Moving average: Each value is the mean of some set which includes some past and future values. It visually "smooths" the data from a scatterplot.

Autoregression: Predicting a value of a time series using that/those preceding it.

Autocorrelation function: Measures linear predictability of a time series at a specific time using another at a different time. Intermediate series values may play indirect role in determining these values. Found by lining up necessary pairwise values and calculating correlation based on that.

Partial autocorrelation function: Measures isolated correlation between two series values x_t and x_{t+h} after removing effects from all series values in between. Found by fitting regression model on series with previous variables being the lags and returning the coefficient ϕ_{hh} of the specific lag variable.

(Weakly) Stationary Time Series: Finite variance process such that 1) the mean stays constant over time and 2) autocovariance depends only on two values s and t inputted into the function through their difference |s - t|. Basic idea is that statistical properties do not depend on time.

Lag: Time shift.

Detrending: Method for nonstationary initial series: subtract OLS predictions from actual Y observations. Good if prediction accuracy is more important than achieving stationarity.

Differencing: Method for nonstationary initial series: get each new observation value by subtracting preceding value from itself. Good if stationarity is more important than prediction accuracy.

Moving average model: Predicting a series values based on past forecast errors.

Autoregressive model: Predicting a series value based on its past values.

2.2 Time Series Model Construction and Selection

Data used for this exercise consists of NASA GISTEMP v4 annual global surface temperature anomalies data. The objective is to fit an appropriate time series model on a portion of the annual dataset (from 1998 to 2013) with monthly temperature anomaly observations.

Figure 1 shows the line plot of the annual data with a rolling mean of window = 5.

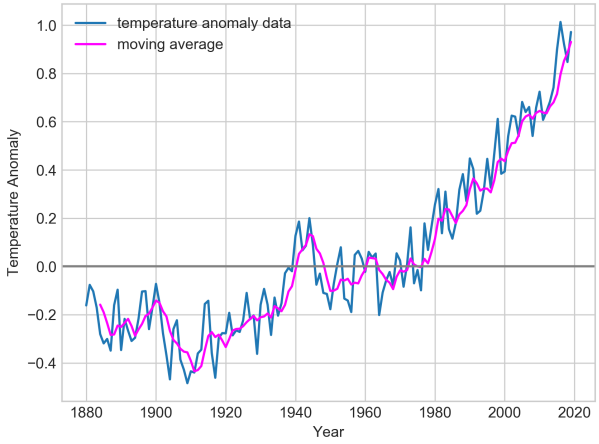

Figure 1: GISTEMP Global Temperature Anomalies Lineplot 1880-2019

It appears that differencing is needed to achieve stationarity in the series. This is confirmed with the initial ACF plot in Figure 2 (the feature of a slow, decreasing decay from left to right).

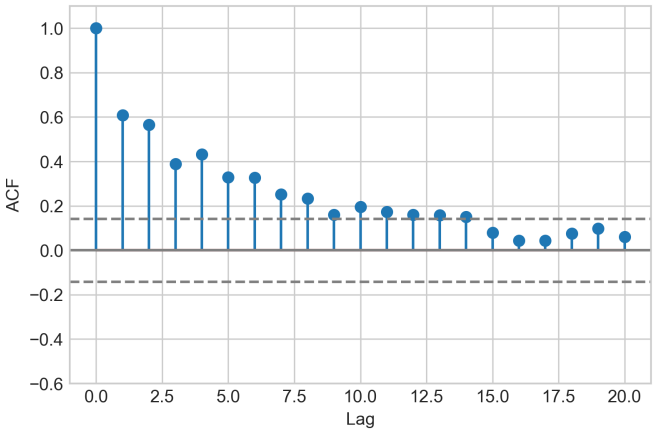

Figure 2: Initial Autocorrelation Plot

After differencing of order 1, the series seems more stationary as shown in Figure 3.

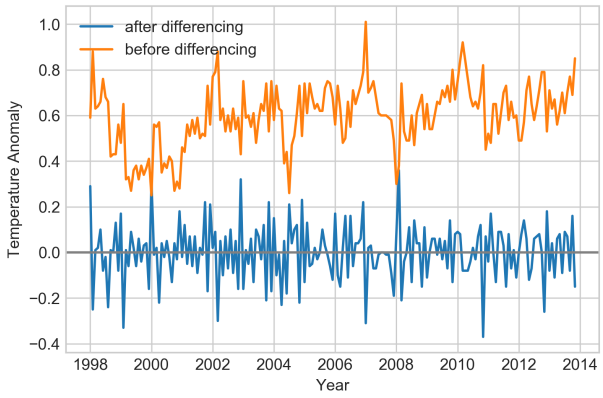

Figure 3: Pre and Post-Differencing Anomalies Line Plot 1998-2013

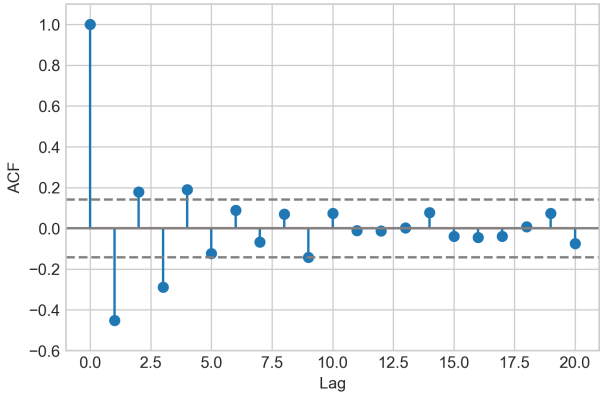

Figure 4: Autocorrelation Plot of Differenced Anomalies

Respectively examining the greatest lag where the correlation magnitude falls outside the 95% confidence levels in the ACF plot above in Figure 4 and PACF plot below in Figure 5, it is concluded that the maximum orders of p and q to consider for an autoregressive integrated moving average model are both 3.

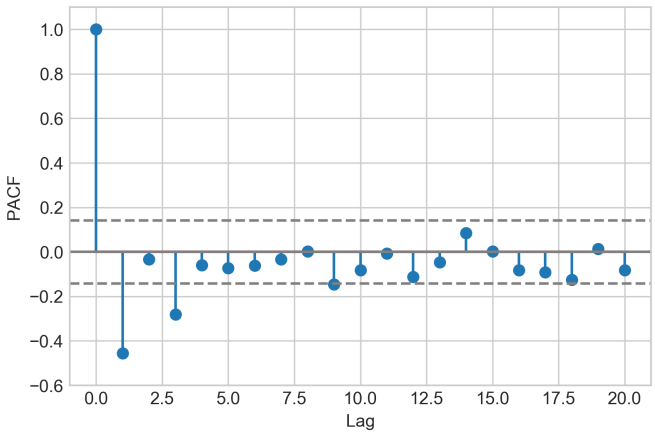

Figure 5: Partial Autocorrelation Plot of Differenced Anomalies

Using these as the upper bounds to find the set of possible ARIMA models, all models under such parameter constraints are fit (with d = 1 because of the differencing order of 1). Then, AIC, BIC, and residual standard deviation values for each are recorded. It is important to report that ARIMA(3, 1, 3) was unable to yield a successful model.

Beside that value, heatmaps of AIC, BIC, and residual standard deviations (three common metrics to compare model performance) are constructed in Figure 6.

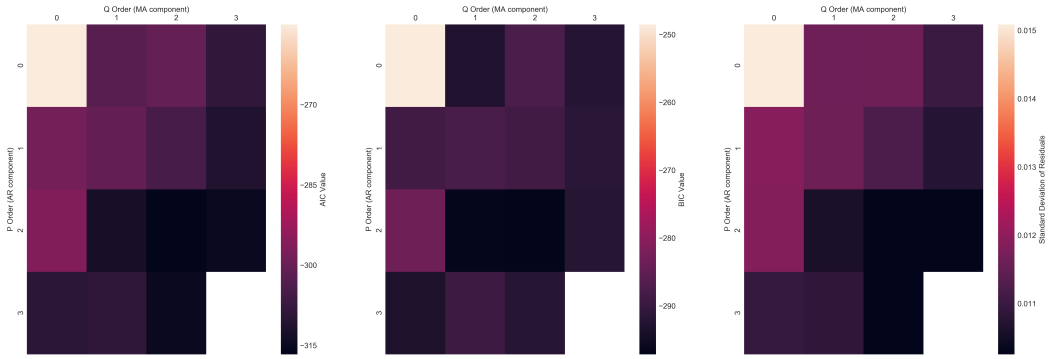

Figure 6: AIC, BIC, and Residual Standard Deviation Heatmaps for ARIMA Models of Different Parameters

Taking into consideration a cursory visual analysis of the heatmaps as well as the favorability for model simplicity, the "optimal" choice of parameters is decided to be p = 0, d = 1, and q = 1. While not holding the lowest AIC or residual standard deviation value among the candidates, its low order makes it preferable in terms of simplicity. In this case, due to the seemingly little improvement that higher order models had in the above metric performances, it is decided that simplicity takes precedence over the relative standings. The model's results table are provided in Figure 7.

ARIMA Model Results

Dep. Variable:	D.change	**No. Observations:**	191
Model:	ARIMA(0, 1, 1)	**Log Likelihood**	155.356
Method:	css-mle	**S.D. of innovations**	0.107
Date:	Mon, 16 Dec 2019	**AIC**	-304.712
Time:	02:49:13	**BIC**	-294.955
Sample:	02-28-1998	**HQIC**	-300.760
	- 12-31-2013		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0003	0.003	0.091	0.928	-0.006	0.007
ma.L1.D.change	-0.5571	0.070	-7.943	0.000	-0.695	-0.420

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	1.7949	+0.0000j	1.7949	0.0000

Figure 7: ARIMA(0, 1, 1) Model Results

Using ARIMA(0, 1, 1), a line plot containing both the GISTEMP data as well as the predictions is shown in Figure 8.

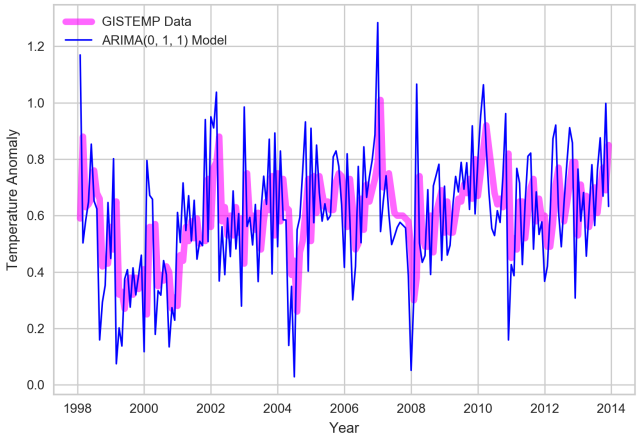

Figure 8: Line Plot of GISTEMP Data and ARIMA(0, 1, 1) Model Predictions

Now having fit the ARIMA(0, 1, 1) model on the 1998 - 2013 monthly GISTEMP temperature anomalies data, the diagnostic plots for the residuals are constructed in Figure 9. The residual plot shows no obvious pattern, suggesting that the model captured most significant properties of the anomalies data. Despite an ACF value at lag = 3 of fairly large magnitude, no significant autocorrelation amount is leftover in the residual ACF plot. For the most part, normality in the residuals can be assumed due to the overall linear nature of the QQ-Plot data points.

However, it appears that some Q-statistics are significant for the Ljung-Box-Pierce calculations; therefore, the residuals may not be entirely comparable to white noise, so the model may have a lack of fit to the anomalies data.

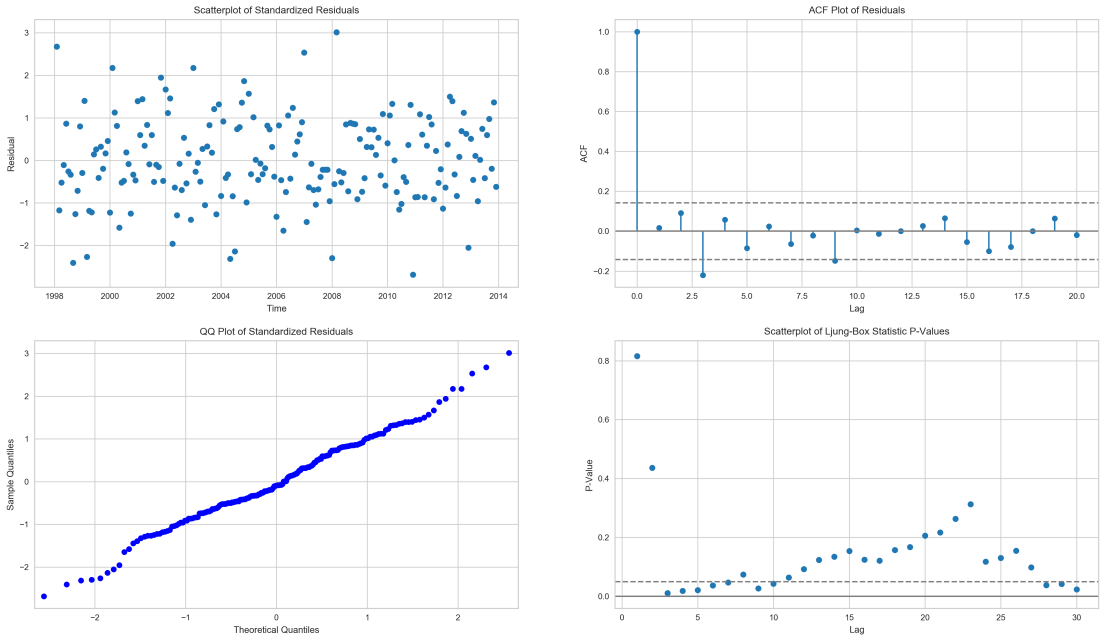

Figure 9: Diagnostic Plots for Residuals After ARIMA(0, 1, 1) Fit

2.3 Hypothesis I's Application To Current Annual GISTEMP Data

In Hypothesis I, the primary objective lies with estimating the p-value of β_1, the coefficient of time in the simple linear regression model fitted on the global temperature series during the hiatus. If the p-value turns out smaller than 0.05, there is enough evidence at a 95% level to reject the null hypothesis $H_0 : \beta_1 = 0$; the conclusion in this case would be that assuming that the temperature series follows a linear model, there is sufficient evidence suggesting that there existed a linear trend during the hiatus time period. Otherwise, H_0 would fail to be rejected: the appropriate conclusion here would be that the evidence of this linear trend during the hiatus is not strong enough.

Various plots are reconstructed using the current annual GISTEMP temperature series data. Figure 10 and 11 are the ACF and PACF plots of the residuals for a simple linear model fitted during the 1950 - 2013 series data.

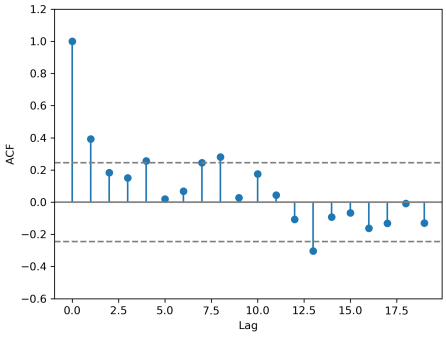

Figure 10: Residuals ACF Plot 1950 - 2013

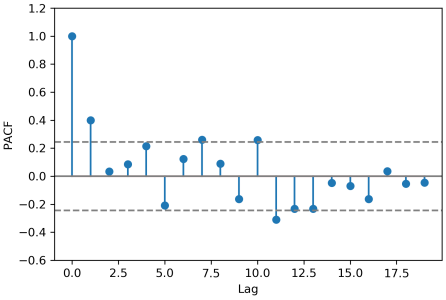

Figure 11: Residuals PACF Plot 1950 - 2013

It is discernable that for this time window, there exists significant autocorrelation and partial autocorrelation among the temperature series values at lag 1. Therefore, it is reasonable to doubt the success of simple linear regression in capturing the temperature anomalies by itself, as there seemingly exists substantial "temporal dependence."

For **Method IA**, a simple linear regression model is constructed for the temperature series during the hiatus time window. The fit on the data is shown in Figure 12.

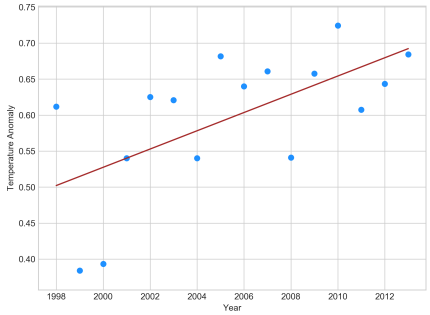

Figure 12: Annual Hiatus Temperature Series Scatterplot With Least Squares Regression Fit Line

The p-value in this case for β_1 is 0.01. This indicates that even with 99% level of significance, there exists enough evidence to believe that the climate hiatus did not occur.

It was decided that **Method IB** would be omitted from this section after examining the lack of significant partial autocorrelation of lag = 1 displayed by the residuals PACF plot below.

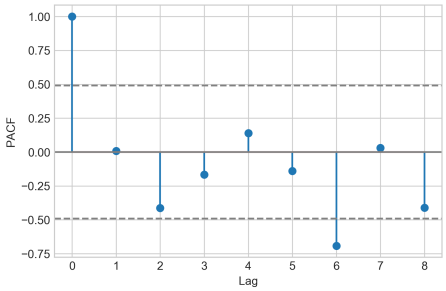

Figure 13: Residuals PACF Plot 1998 - 2013

Next, **Method IC** was carried out via an automation script explained in the next section. The results table is shown in Figure 14.

	Block Size	Standard Error	t14	N(0,1)	Bootstrap
0	1	0.0035702959627325896	0.004333602068418808	0.0006798180484748207	0
1	2	0.003543615985826081	0.007658357092669609	0.0018625812732413365	0.004
2	3	0.0035926424827007683	0.00331614571517882	0.0004121244512405484 4	0.001
3	4	0.0037308368660786337	0.0009529336484089448	3.1108239473942725e-05	0
4	5	0.003647278643288166	0.0017748350622325726	0.0001191283983060900 7	0
5	6	0.0036868941764604804	0.0010243947276532012	3.657825452807091e-05	0
6		0.004508635835238663	0.009958699034198217	0.0028925496588326784	

Figure 14: Results Table

Even when accounting for temporal dependence, the small p-values regardless of bootstrap block size indicate the strength in evidence against the hiatus claims.

However, these results are not guaranteed upon conducting these procedures on other data sources (and perhaps with observations of a different time frequency level). After fitting a linear regression model on the HADCRUT annual temperature series data within the hiatus time window, the p-value for β_1 is 0.19. This means the opposite conclusion (failing to reject the claim of $\beta_1 = 0$) in Method IA when compared to that reached from the GISTEMP data. More research and analysis needs to be conducted for further insight on how to reconcile these results into an overall conclusion.

2.4 Bootstrap Temporal Dependence Algorithm Results Automation Script

Method IC (3.1.3) on page 22 of the research paper utilizes circular block bootstrap, a technique of generating a series preserving stationarity. By executing this method to create "bootstrap" temperature series, the sampling distribution of $\hat{\beta}_1$ can be estimated (then standard error and p-value calculations). A step-by-step walkthrough of the algorithm as well as the standard error and p-value calculation descriptions can be found on page 21 of the research paper.

A Click command line interface script in Python was built automating the bootstrap process as well as the standard error and p-value calculations, outputting the final results in a .csv or .xlsx file. A brief set of instructions on sample usage has been provided below.

1) Download script from Github link provided in Section 3.3 and move it into an appropriate directory.

2) Open Terminal, and set current directory to be that containing the script.

3) Type and enter a specific command into the console of the following format:

`python3 climate_boot.py --time_freq` *FREQ_TYPE* `--start` *START_TIME* `--end` *END_TIME* `--source` *DATA_SOURCE* `--block` *BLOCK_SIZE* `--file` *FILE_TYPE*

FREQ_TYPE : time frequency level of interest (either `year` or `month`)
START_TIME : start time of interest (`month_number/year` if `month`, `year` otherwise)
END_TIME : end time of interest (`month_number/year` if `month`, `year` otherwise)
DATA_SOURCE : climate data source of interest (either `gistemp` or `hadcrut`)
BLOCK_SIZE : maximum block size to conduct circular bootstrap on
FILE_TYPE : results table file type (either `csv` or `xlsx`)

For example, let's say the goal is to generate an excel file containing the standard error and p-value results table (from block size = 1 to 6) for monthly GISTEMP temperature series data during the hiatus window.

The appropriate command then would be `python3 climate_boot.py --time_freq month --start 1/1998 --end 12/2013 --source gistemp --block 6 --file xlsx`

4) A csv or excel file will be generated in the same directory of the script containing the results pertaining to the parameters specified above.

For a reminder of the specific `--tags`, execute steps 1) and 2), then type in `python3 climate_boot.py --help`. Specific information about the script's implementation has been described in code documentation of the script itself.

3 Conclusion

3.1 Reflection

I definitely have several areas for self-improvement looking back. The first would be asking and finishing the course paperwork before the quarter starts; if I did that, I would have had several more weeks to work on the paper. Beyond that, I would have liked to have documented and organized my code on a routinely basis. Coding in Python was not easy for me; in addition to having to manually code up things from the textbook like ACF and PACF plots that were one-line commands in R, I sat through various debugging sessions that eventually turned out to be because of simple errors like a index mismatch or wrong data type. When typing up the final report, I spent a lot of time putting chunks of many Python notebooks together and figuring out what I did: a fair amount of code ended up unused in the final report. Moreover, I wish I made more consistent progress per week. I feel I had weeks that were catalysts, and there were others that seemed a bit stagnant on my end.

In retrospect, I feel I am taking away quite a bit from my Statistics 199 experience. I was always curious about time series analysis: after textbook readings and Python notebooks, I am pleased to say I have gained a rough understanding for time series models and concepts. Beyond statistics, I have sharpened my Python skills by implementing steps in Mike's research paper, having produced a command line interface script that can execute a specific algorithm outlined in the research paper using user-inputted parameters of time duration, frequency, maximum block size, results file type, and dataset. Moreover, I was satisfied with summarizing this entire journey into one research paper using snippets of Latex, Python, and R.

A special thank you to Mike for being so thoughtful, accommodating, and knowledgeable during this process. I am really grateful for this experience fall quarter, and I look forward to further developing and expanding my skillset in Python, experience with time series analysis, and knowledge on climate change.

3.2 Works Cited

Rajaratnam, Bala, Joseph Romano, and Michael Tsiang. "Debunking the Climate Hiatus." Climatic Change, September 2015. https://link.springer.com/article/10.1007/s10584-015-1495-y.

Shumway, Robert H., and David S. Stoffer. Time Series Analysis and Its Applications: with R Examples. Cham, Switzerland: Springer, 2017.

3.3 Python Code Links

2.2 `Time Series Model Construction and Selection`

https://github.com/eshiang21/time_series_climate_data/blob/master/arima_hiatus.ipynb

2.3 `Hypothesis I's Application To Current Annual GISTEMP Data` https://github.com/eshiang21/time_series_climate_data/blob/master/hyp_1.ipynb

2.4 `Bootstrap Temporal Dependence Algorithm Results Automation Script`

https://github.com/eshiang21/time_series_climate_data/blob/master/climate_boot.py