Rose-Hulman Institute of Technology
Rose-Hulman Scholar

Senior Projects - Mathematics

Mathematics

Spring 5-2019

Time Series Analysis on Satellite Observed Carbon Dioxide Data

Bochuan Lyu

Time Series Analysis on Satellite Observed Carbon Dioxide Data

A Senior Project Presented in Partial Fulfillment of
the Bachelor of Science in Mathematics

Bochuan Lyu

Abstract

Carbon dioxide is one of the most important greenhouse gas contributing to global warming [10] and the dramatic increase of carbon dioxide in recent year has been recorded. This paper mainly analyzes the carbon dioxide data from 2011 to 2017 collected by Atmospheric Infrared Sounder (AIRS) on NASA Aqua satellite. We concentrate on the area in Caribbean ocean and northeastern state of Amazonas in Brazil. The statistical models including multiple linear regression, autoregressive–moving-average models, and discrete wavelet transform are employed to study the trends and patterns in the carbon dioxide time series. This results in a partial linear model to find the time dependency, seasonal signals, and significant environmental-factor predictors.

Department of Mathematics
Under the supervision of Dr. Megan Heyman
Rose-Hulman Institute of Technology
May, 2019

Acknowledgments

I would like to acknowledge and thank my supervisor Dr. Megan Heyman for her guidance and support throughout my senior project. The datasets analyzed in this project are collected by National Aeronautics and Space Administration (NASA) and National Oceanic and Atmospheric Administration (NOAA).

Contents

List of Tables

List of Figures

1 Introduction

Global warming has drew wide attention from both academic and government institutes, such as National Oceanic and Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), and University Corporation for Atmospheric Research (UCAR). It can lead to rising sea-level [22] and more extreme weather events [20]. Carbon dioxide, CO_2, is one of the most important greenhouse gas which leads to global temperature rise [8]: a concurrent increasing atmospheric CO_2 and global temperature in the global environment [18] demonstrates a strong correspondence between them [9, 12]. A previous study demonstrated that global warming speeds up the soil organic matter to decay in which process it releases CO_2 [7] and another study predicted the serious consequences of the surging CO_2 to the ecosystem [18], which motivate us to study the relationship between the concentration of atmospheric carbon dioxide and environmental factors, like surface temperature, precipitation, and so on. We are also interested in the temporal dependency within the concentration of carbon dioxide. In this study, we will analyze the patterns and trends of the recent-year concentration of carbon dioxide with statistical methodologies. We selected a natural area without large human-effect in the northeastern state of Amazonas, Brazil, and in the Caribbean ocean to study the changes of concentration of carbon dioxide on both lands and oceans.

The global concentration of carbon dioxide data that we analyze is collected by Atmospheric Infrared Sounder (AIRS) on a NASA satellite [1]. AIRS has been launched aboard the NASA Aqua satellite since 2002 and analyzes the 3.74 μm to 15.4 μm spectral range with 2378 channels to create a fine-grained global maps of various environmental factors [1, 11]. The AIRS mid-tropospheric CO_2 Level 3 daily Gridded Retrieval Product (AIRS version 5 L3 CO_2 daily product) [1] provides 2ppm-accuracy satellite carbon dioxide concentration data in 2° latitude × 2.5° longitude grid boxes with longitudes from -180° to 177.5° and latitudes from -90° to 89.5°.

We have also obtained environmental factors like temperature, precipitation as potential predictor variables from Global Historical Climatology Network - Daily (GHCN-Daily) dataset [14, 15]. This combines daily climate observations from over 100,000 land-based stations worldwide measuring daily environmental factors such as precipitation and temperature (max, min, and average). However, the GHCN-Daily dataset only includes daily land surface observations so modeling the concentration of carbon dioxide in the Caribbean ocean area with exterior predictors would be left for future research.

2 Statistical Models

In this section, we will introduce the major statistical models: multiple linear regression, autoregressive–moving-average, and wavelets models, employed in this report.

2.1 Multiple Linear Regression

Multiple linear regression [25] is a statistical model explaining the relationship between a response variable and predictors, $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_p$ and a response variable $\boldsymbol{Y}$. p is the number of predictors, and any of the predictor variables or response variable is a vector with n observations. The multiple linear regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_p X_{p,i} + \epsilon_i, \quad \forall i \in \{1, 2, ..., n\}, \tag{1}$$

where ϵ_i is the error term. In a matrix notation, we can rewrite the multiple linear regression in Equation (1) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{y}$ is a vector with y_i in the i-th entry, $\boldsymbol{\beta}$ is a vector with β_j in the j-th entry, $\boldsymbol{\epsilon}$ is a vector with ϵ_i in the i-th entry, and $\boldsymbol{X}$ is a matrix that can be written as,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}.$$

Our model employs the method of least-squares to estimate $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

2.2 Autoregressive–moving-average models

Classical regression models usually can't explain the time dependency existing in a time series [21]. We employ autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models to describe those lagged dependencies.

An autoregressive model of order p, usually denoted as AR(p), has a form of

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t, \tag{3}$$

where X_t is a stationary time series, $\phi_1, \phi_2, ..., \phi_p$ are parameters, and ϵ_t follows a normal distribution with zero mean and constant variance σ_ϵ^2.

A moving average model of order q, usually denoted as MA(q), can be written as

$$X_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} + \epsilon_t, \tag{4}$$

where X_t is a stationary time series, $\theta_1, \theta_2, ..., \theta_p$ are parameters, and ϵ_t follows a normal distribution with zero mean and constant variance σ_ϵ^2.

A time series $\{X_t : t = 0, \pm 1, \pm 2, ...\}$ is ARMA(p, q) [21] if it is stationary and

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} + \epsilon_t, \tag{5}$$

where $\phi_1, \phi_2, ..., \phi_p$ are coefficients in AR(p) model described in Equation (3), $\theta_1, \theta_2, ..., \theta_p$ are coefficients in MA(q) model described in Equation (4).

2.3 Discrete Wavelet Transform

There are two types of wavelet transformation: continuous wavelet transform and discrete wavelet transform. However, since the time series datasets used in this project are discrete, it is more appropriate to use the discrete wavelet transform (DWT) in the statistical analysis [23]. Given a $n \times 1$ time series vector x, the coefficients of DWT: d, a $n \times 1$ vector, can be written as [17]

$$d = Wx, \tag{6}$$

where $n = 2^k$, W is an $n \times n$ orthonormal matrix. It means that $WW^T = I$, where I is a $n \times n$ identity matrix. Therefore, $W^T = W^{-1}$ and the discrete wavelet transform is inverted by

$$x = W^T d.$$

The coefficients of W depend on the wavelet function selected. The vector, d, can be written as [23]

$$d = \begin{pmatrix} c_{0,0} \\ \mathrm{d}_{0,0} \\ \mathrm{d}_{1,0} \\ \mathrm{d}_{1,1} \\ \vdots \\ \mathrm{d}_{k-1,0} \\ \vdots \\ \mathrm{d}_{k-1,2^{k-1}} \end{pmatrix},$$

where $c_{0,0}$ is used to describe the average in the time series x, $d_{0,0}$ is the coefficient of mother wavelet, and $d_{i,j}$ is the coefficient for the j-th wavelet in the i-th level. The computational effort to perform the above calculation of Wx in Equation (6) is $O(n^2)$ (quadratic computational time) but it only takes $O(n)$ (linear computational time) if pyramidal algorithm [13] is employed. This is faster than the fast Fourier transform.

There are many different wavelet functions available to perform a discrete wavelet transform, such as Haar wavelets, Meyer's wavelets, and Daubechies' wavelets [23]. The Haar wavelet is the first and simplest orthonormal wavelet basis requiring less computational power, but it is not a desirable basis for smooth functions [23]. Because of the remarkable properties: compactly supported, and orthogonal wavelet bases of $L^2(\mathbb{R})$, Daubechies' wavelets are the most popular wavelet family with various applications in signal processing [24]. Thus, all the wavelet transformations in this report are employed with the least asymmetric Daubechies' wavelets. The mother wavelet can be written as [23],

$$\left(\frac{1+e^{-i\omega}}{2}\right)^N \sum_{l=1}^{L} \left(e^{-i\omega} - z_l\right)\left(e^{-i\omega} - \bar{z}_l\right) \cdot \sum_{j=1}^{J} (e^{-i\omega} - r_j), \tag{7}$$

where $i = \sqrt{-1}$ and we have freedoms to select the values of z_l and r_j.

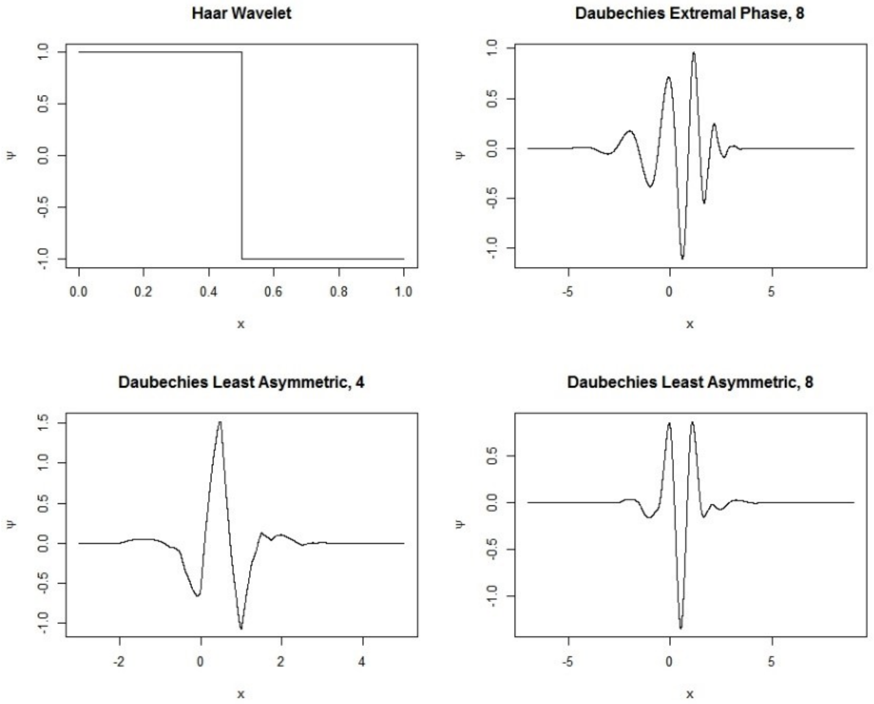


Figure 1: Different wavelet functions

3 A Bootstrapping Method for Discrete Wavelet Transform

A main challenge in applying discrete wavelet transform is performing a proper inference in statistical analysis. Which coefficients of the wavelet are statistically significant? There are several threshold technique available to determine how many levels of coefficients are significant, such as universal threshold [23]. However, a systematic method for each wavelet coefficient is still a opening question. In order to better perform a statistical inference, we have employed a technique combining both bootstrapping and threshold to do statistical inference on the coefficients. Potentially, it might also be helpful for us to select a better wavelet model during statistical analysis.

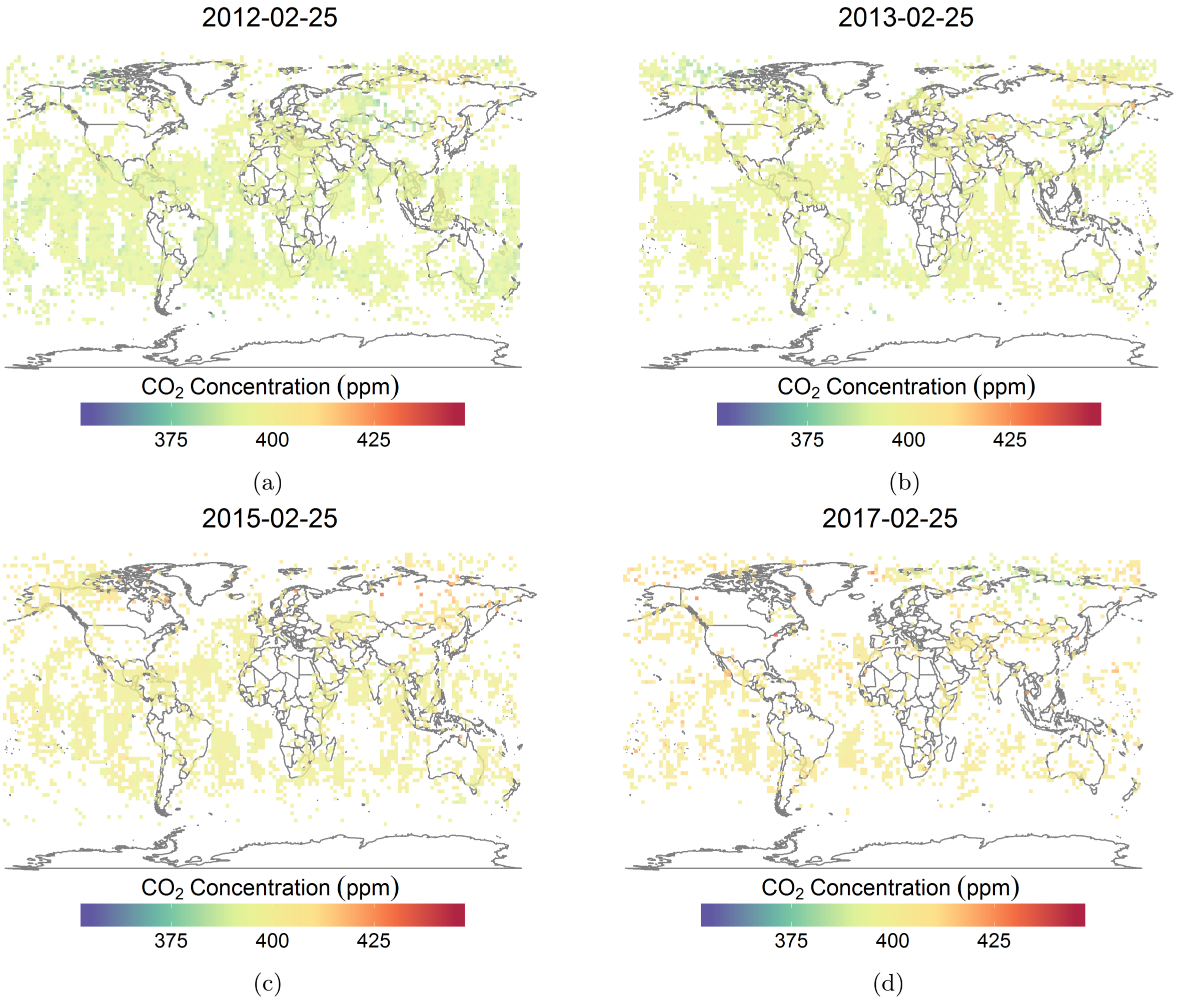
The procedure of our method is described as follows. First, we need to select a wavelet model with certain levels. Then, use this wavelet model to extract the signal from the time series data so as to obtain the residuals. If it is reasonable for us to assume the independence, mean of 0, and constant variance of those residuals, we could use residual bootstrapping to obtain bootstrapped time series data and corresponding coefficients of wavelets. We could use those coefficients to construct confidence intervals or obtain p-values.

3.1 Simulation

We have also construct a simulation study to examine the performance the method in Section 3. In the simulation, we have used first three levels of the coefficients of "DaubLeAsymm" wavelet with 10 filters to generate the truth time-series data and then added some noise from different distributions with different signal-to-noise ratio to generate simulated data. We have studied the effects of signal-to-noise ratio, noise distributions, and the selection of wavelets family and the number filters on the performance of the discrete wavelet transform. In all of the simulations, true coefficients are captured by the bootstrapped distribution constructed by our new method. The results are attached in the appendix B.

4 Data Cleaning

The daily carbon dioxide data from AIRS version 5 L3 CO_2 daily product is stored in a large matrix, where each row represents different longitudes from -180° to 177.5° and each column represents different latitudes from -90° to 89.5° from Jan. 1, 2010 to Feb. 28, 2017. Measured from Google Maps [4], the latitude of Caribbean Sea is approximately from 12 to 17, and the longitude is from about -70 to -80. The latitude of the green area of interest around the northeastern state of Amazonas in Brazil is approximately from -10 to -2, and the longitude is from about -53 to -62. We take the average of daily carbon dioxide concentration over the grids in those areas to obtain the daily CO_2 concentration time series in the northeastern state of Amazonas, Brazil, and the Caribbean shown in Figure 3 (b) and (c).



(a)

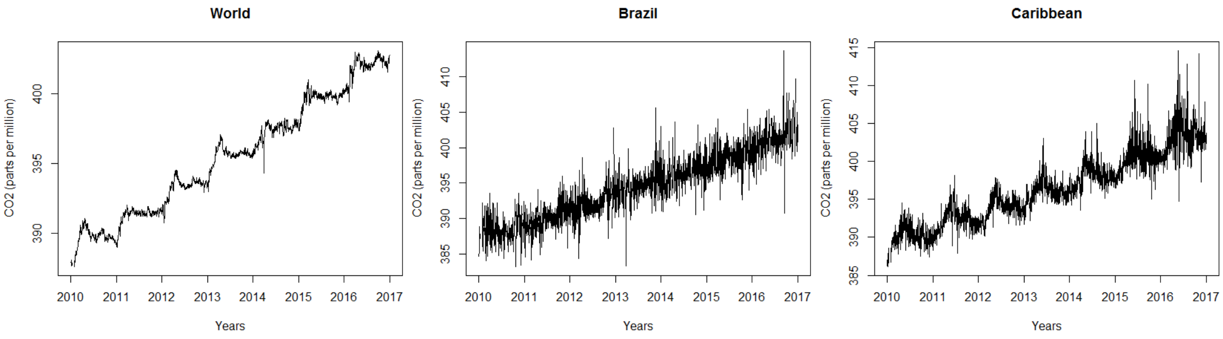
(b)

(c)

(d)

Figure 2: The visualization of CO_2 concentration (parts per millions) from 2012 to 2017. There aren't any CO_2 concentration data in the grids without any colors.

The carbon dioxide concentration appears to increase globally from 2011 to 2017 as shown in Figure 2. We can better visualize the overall increasing trends in Figure 3 (a). Our data matches the previous study [18] and we have also noticed the clear seasonal patterns in the global average CO_2 concentration and Caribbean average CO_2 concentration. It is difficult to accurately measure the CO_2 concentration from the satellite at every single location on Earth every day so NASA only keeps the stable and trustworthy data in AIRS L3 CO_2 daily product [1]. There are some missing values in the CO_2 concentration, which we can also see in Figure 2. Furthermore, we can see in Figure 2 that there tends to be more missing CO_2 concentration data over land than oceans and more missing data over time.



(a) (b) (c)

Figure 3: The changes of average carbon dioxide concentration in the World, Brazil, and Caribbean over time.

In the Caribbean CO_2 concentration data, there are 34 missing values out of 2048, and there are 172 missing values out of 2048 in the Brazil CO_2 concentration data. The actual missing dates are provided in Table 1 and 2 in the Appendix. Bennett maintained that statistical analysis is likely to be biased when more than 10% of data are missing [2]. Both of the missing values take up less than 10 % in total data. Therefore, it is still reasonable to employ statistical models to analyze the current dataset to produce reliable results.

Since the ARMA models and discrete wavelet transformation, which are employed in the Section 5, require no missing values in the time series data, we need to figure out a way to impute those missing values [19]. In order to impute the missing values, we need to check whether the data is missing at random. There are 7 missing CO_2 values everywhere globally. Thus, I checked the online archive of AIRS L3 CO_2 daily product and realized that there are no data files on those dates. There are no obvious document recording the reasons for those missing data files. There could be several potential reasons, such as the maintenance of the satellite. Furthermore, the rest of the values seem not to be randomly missing as well, because there are some missing data points in rows both in the northeastern state of Amazonas, Brazil, and the Caribbean. Although it isn't fully reasonable to believe the missing values are random, we need to make that assumption to impute the missing values in order to do further study and leave the effects of this assumption for future work.

There are several common imputation methods: missing value imputation by last observation carried forward, weighted moving average, mean value, random samples, interpolation, or Kalman smoothing and state space models [16]. The previous study [3] has done simulations to test the performance of those imputation methods and the results show that structural models using Kalman smoothing and linear interpolation handle missing data in univariate time series with the best performance. According to the results [3], the imputation method using Kalman Smoothing on structural time series models has the best performance at the missing rate of 0.1. Hence, we employ Kalman Smoothing on structural time series models to impute the missing data. After we impute the data, we attain the data with a length of 2616. As we mentioned in Section 2.3, the discrete wavelet transform requires the length of time series to be a power of 2. Therefore, we have kept the last 2048 elements in the CO_2 time series, which is from July 23, 2011, to Feb. 28, 2017.

Additionally, we also obtain daily data of minimum, average, and maximum temperature, and precipitation from GHCN-Daily dataset. Since the data is collected by station-based measurements, it only contains land area data. Furthermore, there are too many missing values (over 20%) of minimum, maximum temperature, or precipitation data in the northeastern state of Amazonas, Brazil. Thus, we could use those predictor variables in our models. However, the average temperature could be used if 56 missing values are imputed properly. When we investigate more on those missing values, most of them are missing because of no data are collected during those days. But, there are no evidences that there are some extreme weather happening in those days. It is reasonable for us to assume that those values can be imputed or predicted by the values collected in a similar time. Similarly as we do for CO_2 data, we use Kalman Smoothing on structural time series models to impute the missing data.

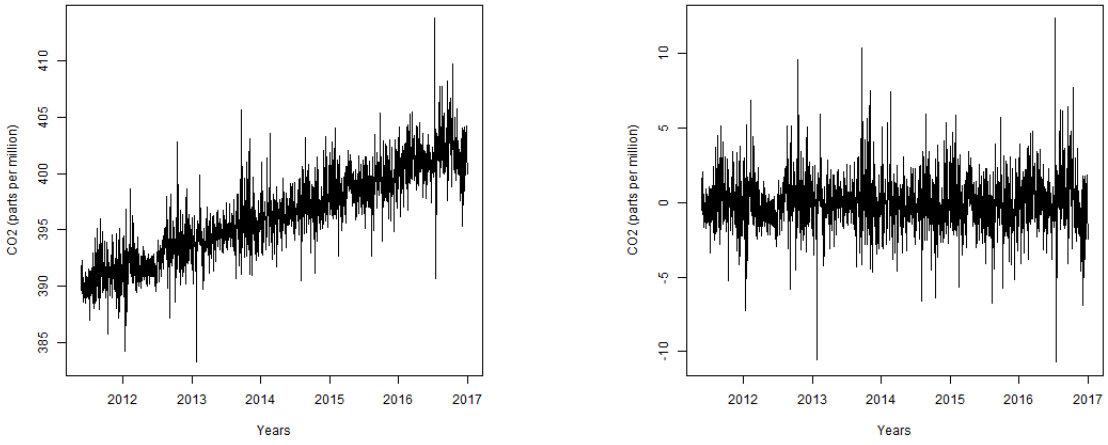
5 Data Analysis

In this section, we will analyze the carbon dioxide time series data in the northeastern state of Amazonas, Brazil, and the Caribbean area mentioned in Section 4 with statistical models.

5.1 Brazil

In Figure 4 (a), we can still see a positive linear trend so we can't fit the discrete wavelet transform to this time series. Hence, we fit a linear regression model with a predictor variable, time, to remove the linear trend,

$$CO_{2,\text{Brazil}} = \beta_0 + \beta_1 \cdot \text{Time} + \epsilon. \tag{8}$$



(a)

(b)

Figure 4: The carbon dioxide time series data in the northeastern state of Amazonas, Brazil after imputation (Figure (a)) and the residuals after fitting a linear regression in Equation (8) (Figure (b)).

The time series after removing the linear trend is shown in Figure 4 (b). However, without any increasing or decreasing overall trends, there might be some moderate seasonal patterns. Therefore, we employ a discrete wavelet transform to detect and extract the seasonal trends. As demonstrated in Figure 5, there are the signals found by DWT by levels from 0 to n of coefficients, $\forall n \in \{0, 1, ..., 11\}$, since $2048 = 2^{11}$. In the first 5 plots of Figure 5, the fitted curves are too smooth so that the discrete wavelet transform doesn't extract enough information of the seasonal patterns. Figure 5 is made by functions in WiSEBoot package [5]. In the last 4 plots of Figure 5, the fitted curves might overfit and the discrete wavelet transforms extract not only the signals but also the noise in the time series. $J_0 + 1 = 5$ and $J_0 + 1 = 6$ might be the best two options. $J_0 + 1 = 5$ captures the signals with a minimum unit of 64 days and $J_0 + 1 = 6$ captures the signals with that of 32 days, which is approximately one month. Therefore, I use the fitted curved by keeping the levels from 0 to 6 of the coefficients in DWT.

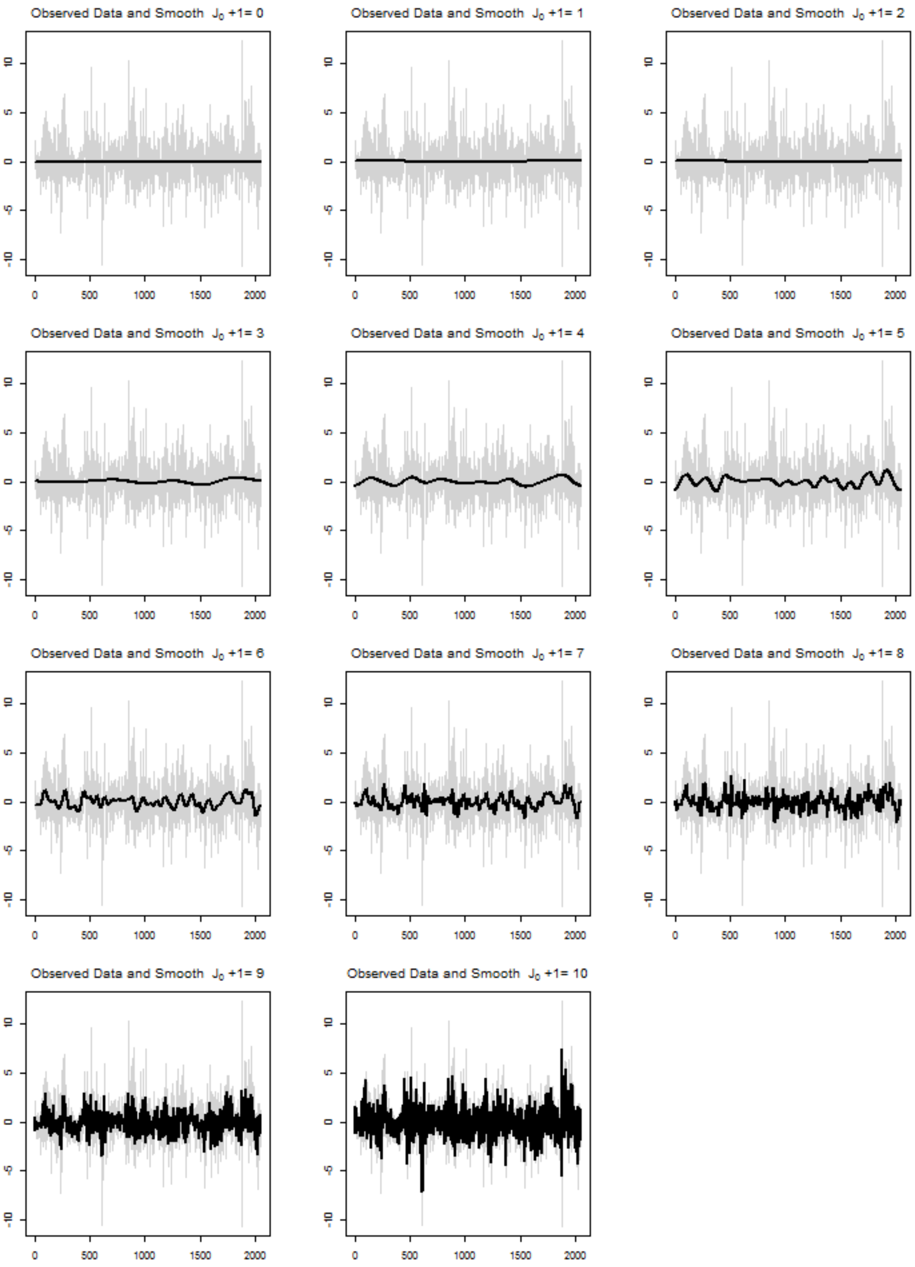


Figure 5: The signals in carbon dioxide time series data of Brazil found via discrete wavelet transform by keeping levels from 0 to n of coefficients, $\forall n \in \{0, 1, ..., 11\}$ [5].

After the detrending of discrete wavelet transforms, we can see the time series in Figure 11 has no clear overall increasing or decreasing trends, and there are only very small and ignorable seasonal patterns. Then, we check the assumptions for statistical inference as shown in Figure 15 in Section A: Appendix and notice that the errors are not independent. We might want to fit an ARMA model to explain the time dependency within the time series. Additionally, we make ACF and PACF plots to determine the parameters of the orders of AR and MA in ARMA model.

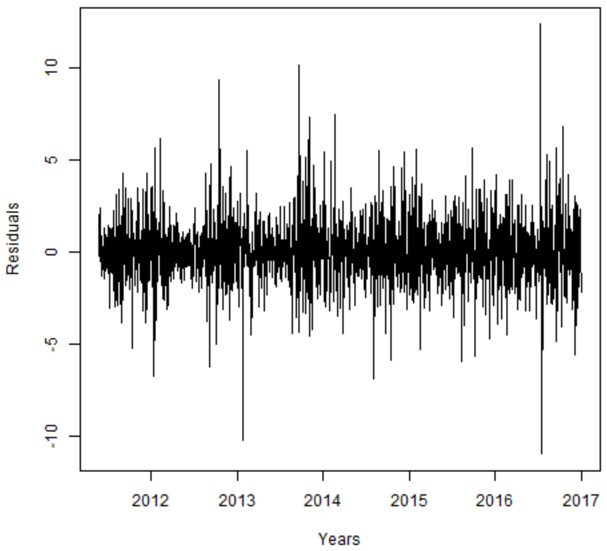


Figure 6: The carbon dioxide time series data in the northeastern state of Amazonas, Brazil after imputation, linear detrending, and wavelet detrending.

There are no significant lag-values in ACF plot of Figure 7 but some lag-values might be significant in PACF plot of Figure 7. Thus, we might expect an ARMA model with more than one autoregressive terms and moving average terms. For now, we still don't have any predictor variables processed and cleaned, which have the high priority to do in the next term. We can only fit an ARMA model with time and the coefficients of a discrete wavelet transform without predictor variables of environmental factors, like precipitation, and temperature.

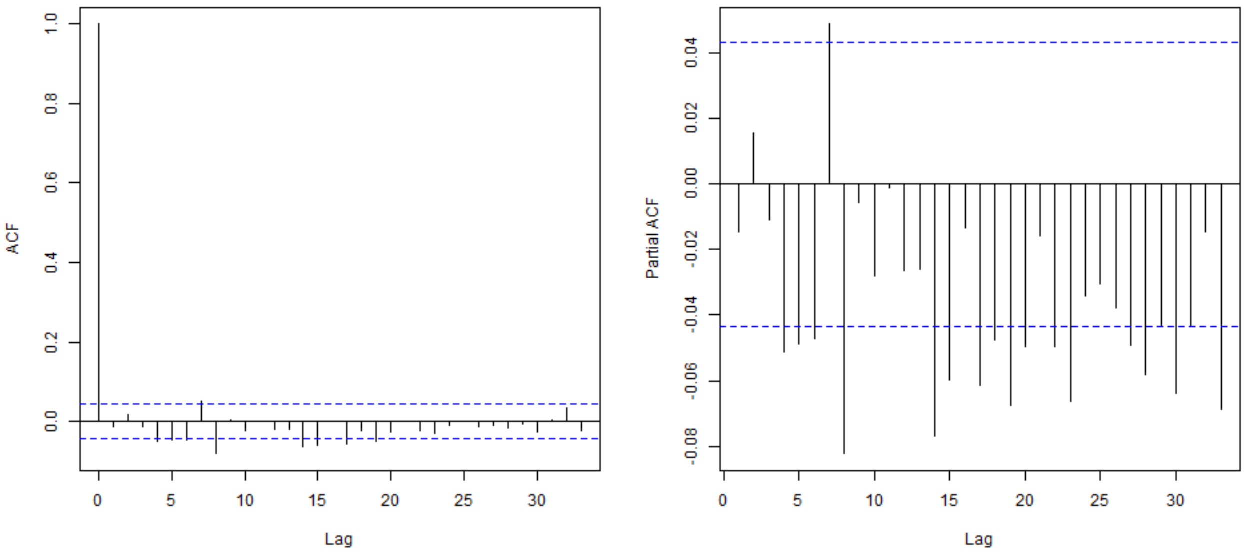


Figure 7: The autocorrelation function (left) and partial autocorrelation function (right) of carbon dioxide time series data in the northeastern state of Amazonas, Brazil.

The final statistical model for Brazil is

$$\begin{aligned} X_t =& \beta_0 + \beta_1 \cdot T_{1,t} + \beta_2 \cdot T_{2,t} + (W^T d_B)_t + \phi_1 X'_{t-1} \\ &+ \phi_2 X'_{t-2} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t, \end{aligned} \tag{9}$$

where T_1 represents the date, T_2 represents the temperature, W is a matrix using the least asymmetric Daubechies' wavelets, d_B are the truncated wavelet coefficients for Brazil by keeping the levels from 0 to 6 of the coefficients in DWT, $X'_t = X_t - (\beta_0 + \beta_1 \cdot T_{1,t} + \beta_2 \cdot T_{2,t} + (W^T d_B)_t)$, and ϵ are error terms.

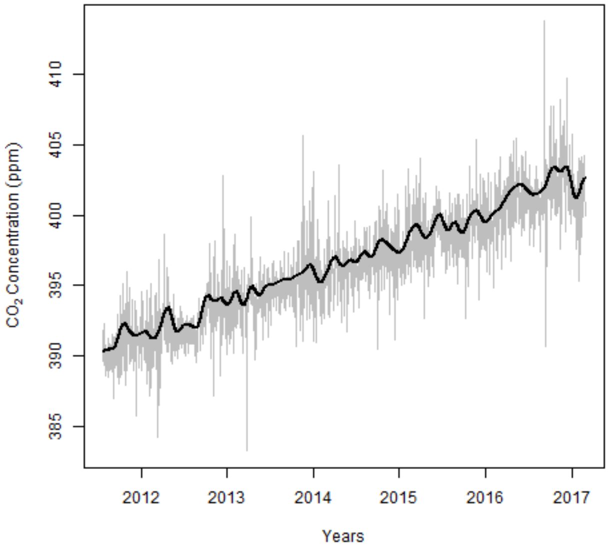
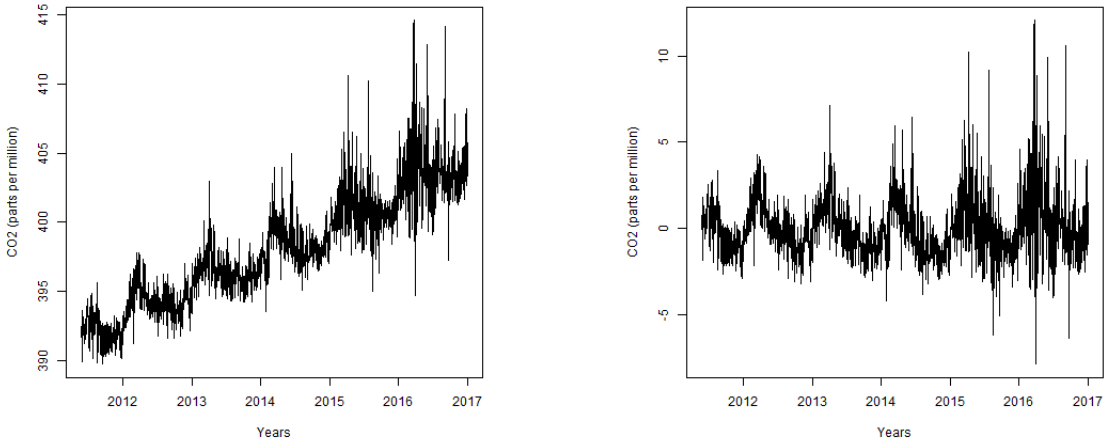


Figure 8: The signal captured by final model in Equation (9)

5.2 Caribbean

In Figure 9 (a), we can still see an overall positive linear trend so we can't fit the discrete wavelet transform to this time series. Hence, similarly as the Equation (8), we fit a linear regression model with a predictor variable, time, to remove the linear trend,

$$CO_{2,\text{Caribbean}} = \beta_0 + \beta_1 \cdot \text{Time} + \epsilon. \tag{10}$$



(a)

(b)

Figure 9: The carbon dioxide time series data in the Caribbean after imputation (Figure (a)) and the residuals after fitting a linear regression in Equation (10) (Figure (b)).

The time series after removing the linear trend is shown in Figure 9 (b). However, since linear regression model in Equation (10) can not explain the seasonal patterns within the time series, there are still a clear oscillating seasonal trend as shown in Figure 9 (b). Therefore, we can similarly employ a discrete wavelet transform to detect and extract the seasonal trends. As demonstrated in Figure 10, there are the signals found by DWT by levels from 0 to n of coefficients, $\forall n \in \{0, 1, ..., 11\}$, since $2048 = 2^{11}$. Figure 10 is made by functions in WiSEBoot package [5]. Similarly as Section 5.1, in the first 6 plots of Figure 10, the fitted curves don't contain enough information of the seasonal patterns and in the last 4 plots, there might be too much noise rather than signals. Therefore, I use the fitted curved by keeping the levels from 0 to 6 of the coefficients in DWT, which has minimum unit of 32 days (approximately one month).

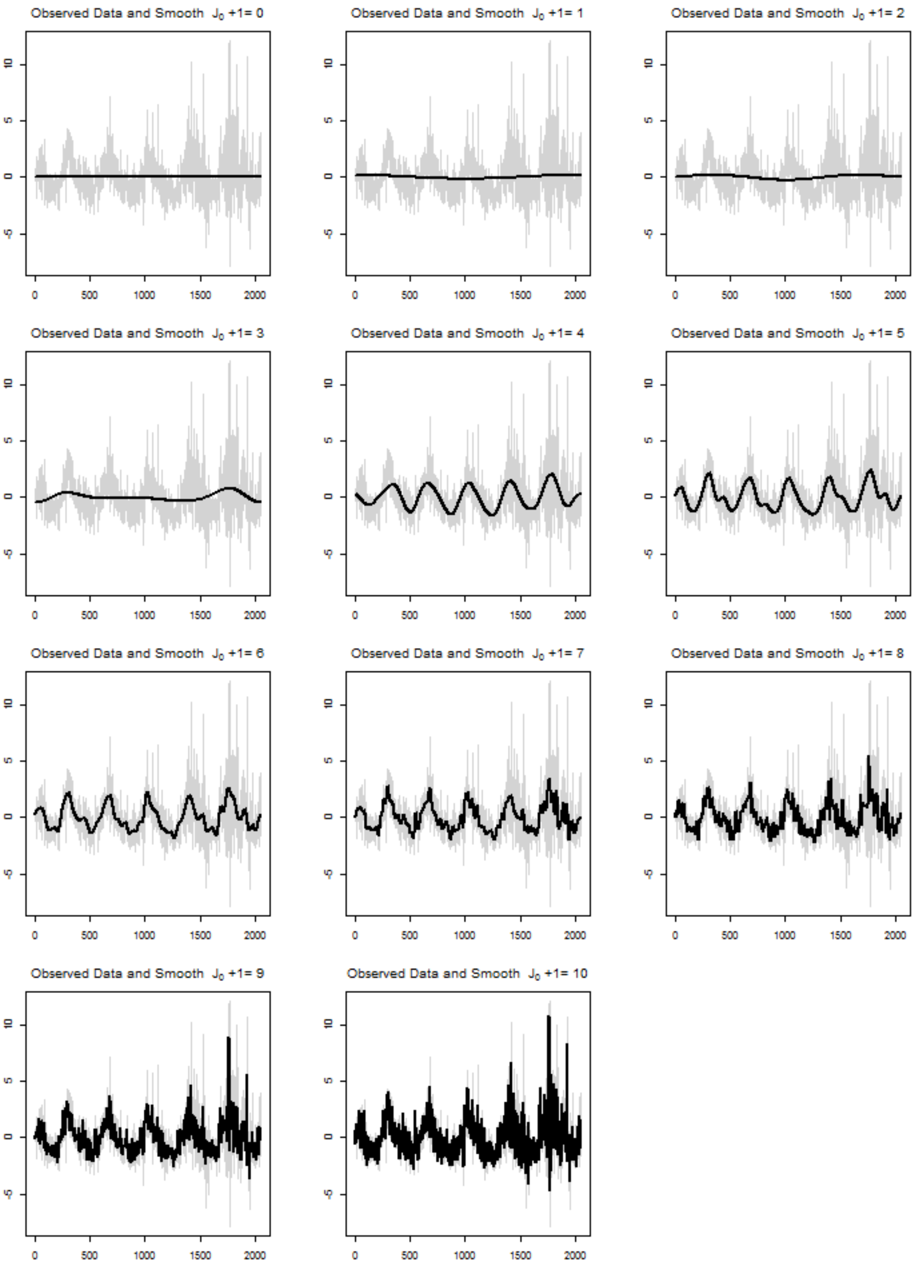


Figure 10: The signals in carbon dioxide time series data of Caribbean found via discrete wavelet transform by keeping levels from 0 to n of coefficients, $\forall n \in \{0, 1, ..., 11\}$ [5].

After the detrending of discrete wavelet transforms, we can see the time series in Figure 11 has no clear overall increasing or decreasing trends, and there are still some clear seasonal trends (a fan pattern). Then, we check the assumptions for statistical inference as shown in Figure 14 in Section A: Appendix, and we can notice that the errors are not independent. Although there are still some seasonal patterns, we might still want to fit an ARMA model to explain the time dependency within the time series. Furthermore, we make ACF and PACF plots to determine the parameters of the orders of AR and MA in ARMA model.

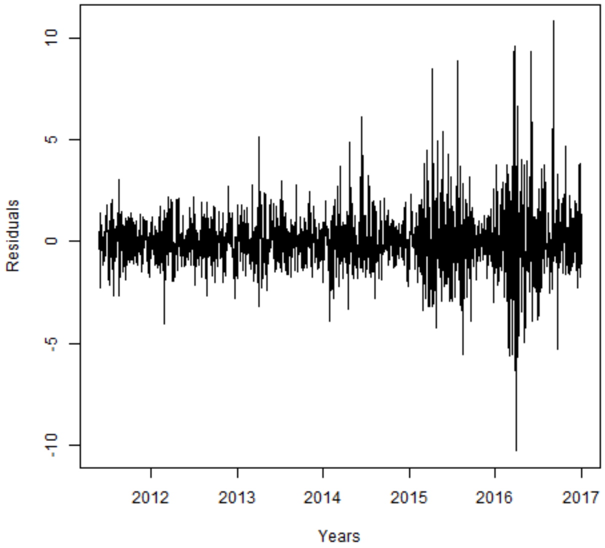


Figure 11: The carbon dioxide time series data in the Caribbean after imputation, linear detrending, and wavelet detrending.

There might be only one significant lag-value (lag-1) in ACF plot of Figure 12 and one lag-value (lag-1) might be significant in PACF plot of Figure 12. Thus, we might expect an ARMA model with one autoregressive term and moving-average term. Since GHCN-Daily dataset only contains land-based stations measurement, we need to investigate more about whether this dataset is appropriate to provide predictor variables in the Caribbean Ocean, and we might want to seek another dataset. For now, we can only fit an ARMA model with time and the coefficients of a discrete wavelet transform without predictor variables of environmental factors, like precipitation, and temperature.

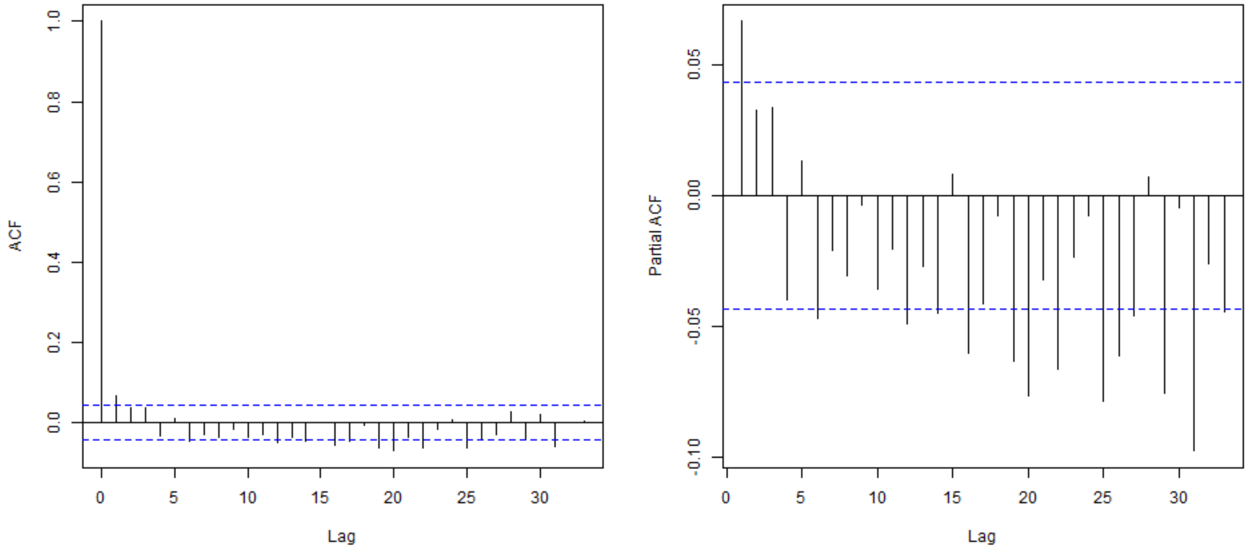


Figure 12: The autocorrelation function (left) and partial autocorrelation function (right) of carbon dioxide time series data in the Caribbean.

The final statistical model for Caribbean is

$$X_t = \beta_0 + \beta_1 \cdot T_{1,t} + (W^T d_C)_t + \phi_1 X'_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \tag{11}$$

where T_1 represents the date, W is a matrix using the least asymmetric Daubechies' wavelets, d_C are the truncated wavelet coefficients for Caribbean by keeping the levels from 0 to 6 of the coefficients in DWT, $X'_t = X_t - (\beta_0 + \beta_1 \cdot T_{1,t} + (W^T d_C)_t)$, and ϵ are error terms.

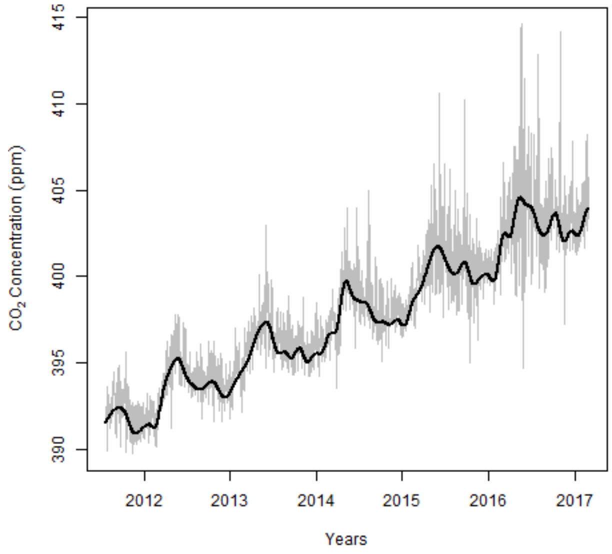


Figure 13: The signal captured by final model in Equation (11)

6 Conclusion

First, the carbon dioxide concentration in the northeastern state of Amazonas, Brazil, Caribbean, and the global is increasing dramatically from 2010 to 2017. We also confirm the known result of a linear increasing trend of CO_2 concentration over time. Second, the discrete wavelet transform detects monthly signals of carbon dioxide concentration in the chosen areas of both Brazil and the Caribbean. We also find that there isn't any significant correlation between temperature and carbon dioxide concentration.

7 Future Work

In the future, we can figure out a more systematic method to do statistical inference for the final models in Equations (9) and (11). Besides statistical inference and explanations on the coefficients, prediction of carbon dioxide concentration in the future can be a potential research direction. Furthermore, there is only one predictor variables available for Brazil and no predictors available for Caribbean up to now. We should look for additional datasets so as to find out better predictors for CO_2.

References

[1] AIRS Science Team/Joao Texeira (2009), AIRS/Aqua L3 Daily CO2 in the free troposphere (AIRS-only) 2.5 degrees x 2 degrees V005, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: October 31, 2018, 10.5067/Aqua/AIRS/DATA335

[2] Bennett, D. A. (2001). How can I deal with missing data in my study?. *Australian and New Zealand journal of public health*, 25(5), 464-469.

[3] Elissavet, R. (2017), Missing Data in Time Series and Imputation Methods.

[4] Google Maps. (n.d.). Retrieved November 1, 2018, from https://www.google.com/maps

[5] Heyman, M. and Chatterjee, S. "WiSEBoot: Wild Scale-Enhanced Bootstrap." R package version 1.4.0, 2016.

[6] Hyndman, R. J., Khandakar, Y. (2007). Automatic time series for forecasting: *the forecast package for R* (No. 6/07). Monash University, Department of Econometrics and Business Statistics.

[7] Jenkinson, D. S., Adams, D. E., Wild, A. (1991). Model estimates of CO2 emissions from soil in response to global warming. Nature, 351(6324), 304.

[8] Joos, F., Plattner, G. K., Stocker, T. F., Marchal, O., Schmittner, A. (1999). Global warming and marine carbon cycle feedbacks on future atmospheric CO2. *Science, 284*(5413), 464-467.

[9] Jouzel, J., Masson-Delmotte, V., Cattani, O., Dreyfus, G., Falourd, S., Hoffmann, G., ... Fischer, H. (2007). Orbital and millennial Antarctic climate variability over the past 800,000 years. *science, 317*(5839), 793-796.

[10] Krupa, S. V., Kickert, R. N. (1989). The greenhouse effect: impacts of ultraviolet-B (UV-B) radiation, carbon dioxide (CO2), and ozone (O3) on vegetation. *Environmental Pollution, 61*(4), 263-393.

[11] Laboratory, N. J. (2017, May 04). What's in the Air: NASA's Atmospheric Infrared Sounder. Retrieved from https://www.youtube.com/watch?v=SWCs3BBzPoU

[12] Lüthi, D., M. Le Floch, B. Bereiter, T. Blunier, J.-M. Barnola, U. Siegenthaler, D. Raynaud, J. Jouzel, H. Fischer, K. Kawamura, and T.F. Stocker. 2008. High-resolution carbon dioxide concentration record 650,000-800,000 years before present. *Nature*, Vol. 453, pp. 379-382, 15 May 2008.

[13] Mallat, S. (1999). *A wavelet tour of signal processing.* Elsevier.

[14] Menne, Matthew J., Imke Durre, Bryant Korzeniewski, Shelley McNeal, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S. Vose, Byron E.Gleason, and Tamara G. Houston (2012): Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. Version 3.24. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ [November 7, 2018].

[15] Menne, M.J., I. Durre, R.S. Vose, B.E. Gleason, and T.G. Houston, 2012: An overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29, 897-910, doi:10.1175/JTECH-D-11-00103.1.

[16] Moritz, S., Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. *The R Journal, 9*(1), 207-218.

[17] Nason, G. P. (2008). *Wavelet methods in statistics with R.* New York: Springer.

[18] Norby, R. J., & Luo, Y. (2004). Evaluating ecosystem responses to rising atmospheric CO2 and global warming in a multi-factor world. *New Phytologist, 162*(2), 281-293.

[19] Percival, D. B., Walden, A. T. (2008). *Wavelet methods for time series analysis.* Cambridge: Cambridge University Press.

[20] Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, P. R., Chivian, E. (2001). Climate change and extreme weather events; implications for food production, plant diseases, and pests. *Global change & human health, 2*(2), 90-104.

[21] Shumway, R. H., Stoffer, D. S. (2011). Time series regression and exploratory data analysis. In *Time series analysis and its applications* (pp. 47-82). Springer New York.

[22] Vermeer, M., Rahmstorf, S. (2009). Global sea level linked to global temperature. *Proceedings of the national academy of sciences, 106*(51), 21527-21532.

[23] Vidakovic, B. (2013). *Statistical modeling by wavelets.* Oxford: Wiley-Blackwell.

[24] Vonesch, C., Blu, T., Unser, M. (2007). Generalized Daubechies wavelet families. *IEEE Transactions on Signal Processing, 55*(9), 4415-4429.

[25] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley Sons.

A Appendix

The Date of Missing Values						
9/27/2011	11/7/2011	1/4/2012	2/15/2012	10/4/2012	10/31/2012	11/26/2012
12/14/2012	1/13/2013	1/28/2013	2/9/2013	3/6/2013	3/16/2013	3/23/2013
4/21/2013	11/1/2013	11/27/2013	1/20/2014	2/14/2014	3/5/2014	3/22/2014
3/26/2014	4/10/2014	11/22/2014	2/1/2015	2/20/2015	3/17/2015	3/24/2015
10/29/2015	12/25/2015	2/25/2016	3/6/2016	3/25/2016	4/11/2016	5/31/2016
9/25/2016	10/21/2016	11/14/2016	11/25/2016	12/4/2016	12/16/2016	12/25/2016
1/26/2017	10/11/2011	11/10/2011	1/20/2012	2/16/2012	10/9/2012	11/12/2012
11/28/2012	12/15/2012	1/15/2013	1/29/2013	2/11/2013	3/7/2013	3/17/2013
4/3/2013	5/7/2013	11/4/2013	12/1/2013	1/25/2014	2/15/2014	3/7/2014
3/23/2014	3/27/2014	4/11/2014	11/25/2014	2/4/2015	2/22/2015	3/18/2015
4/2/2015	11/6/2015	1/13/2016	2/26/2016	3/14/2016	3/26/2016	4/12/2016
8/21/2016	9/26/2016	10/23/2016	11/16/2016	11/27/2016	12/12/2016	12/20/2016
1/3/2017	2/9/2017	10/18/2011	12/16/2011	1/26/2012	4/11/2012	10/14/2012
11/21/2012	12/2/2012	12/23/2012	1/20/2013	2/7/2013	2/14/2013	3/14/2013
3/18/2013	4/4/2013	10/9/2013	11/5/2013	12/24/2013	2/3/2014	2/21/2014
3/8/2014	3/24/2014	3/28/2014	6/19/2014	11/27/2014	2/6/2015	2/27/2015
3/19/2015	4/27/2015	11/27/2015	1/31/2016	3/3/2016	3/21/2016	4/1/2016
4/23/2016	8/28/2016	9/27/2016	11/3/2016	11/23/2016	11/30/2016	12/13/2016
12/21/2016	1/13/2017	2/15/2017	10/30/2011	12/19/2011	1/30/2012	4/12/2012
10/27/2012	11/23/2012	12/11/2012	1/12/2013	1/24/2013	2/8/2013	3/4/2013
3/15/2013	3/22/2013	4/6/2013	10/31/2013	11/17/2013	1/11/2014	2/12/2014
3/2/2014	3/11/2014	3/25/2014	4/8/2014	11/11/2014	1/19/2015	2/17/2015
3/5/2015	3/21/2015	10/28/2015	11/28/2015	2/14/2016	3/5/2016	3/22/2016
4/10/2016	5/17/2016	9/2/2016	10/20/2016	11/12/2016	11/24/2016	12/2/2016
12/14/2016	12/22/2016	1/17/2017	2/20/2017			

Table 1: The missing Dates in the northeastern state of Amazonas, Brazil CO_2 Time Series Data

The Date of Missing Values						
10/24/2012	4/6/2013	3/23/2014	3/25/2014	3/27/2014	6/19/2014	3/30/2015
7/8/2015	10/1/2015	4/23/2016	5/27/2016	6/6/2016	6/10/2016	9/25/2016
9/27/2016	10/10/2016	11/21/2016	2/9/2013	3/22/2014	3/24/2014	3/26/2014
3/28/2014	6/22/2014	5/16/2015	9/6/2015	11/17/2015	5/17/2016	6/1/2016
6/8/2016	8/27/2016	9/26/2016	10/3/2016	11/12/2016	11/24/2016	

Table 2: The missing Dates in the Caribbean CO_2 Time Series Data

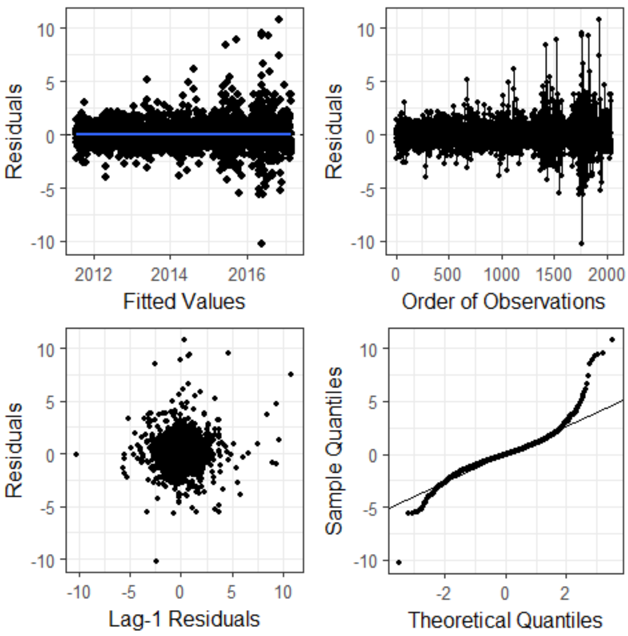


Figure 14: Checking the assumptions of Caribbean CO_2 residuals for statistical inference.

By looking at the Residuals vs. Order of observations plot, we can see a increasing variance over time. Thus, we can't reasonably assume that the error terms are independent. Since the assumption that errors are independent isn't reasonably met, we don't have any methods for statistical inference.

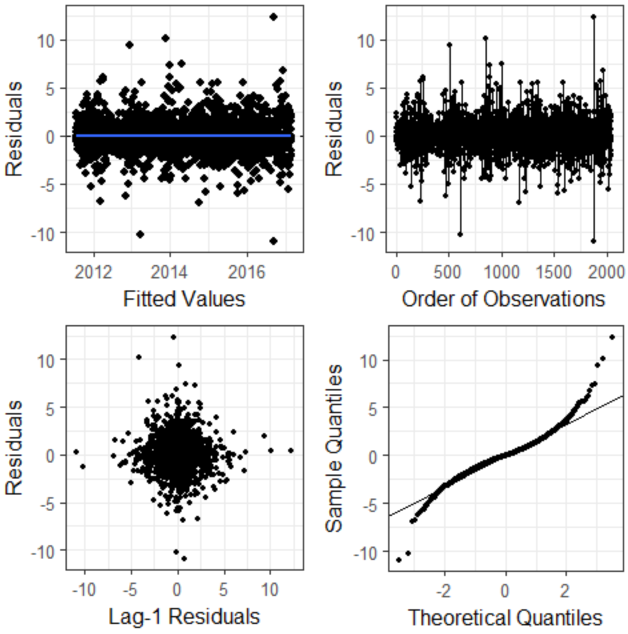


Figure 15: Checking the assumptions of Brazil CO_2 residuals for statistical inference.

By looking at the Residuals vs. Order of observations plot, we can see a increasing variance over time. Thus, we can't reasonably assume that the error terms are independent. Since the assumption that errors are independent isn't reasonably met, we don't have any methods for statistical inference.

B Appendix

B.1 Tuning the Signal-to-noise (Error) Ratio

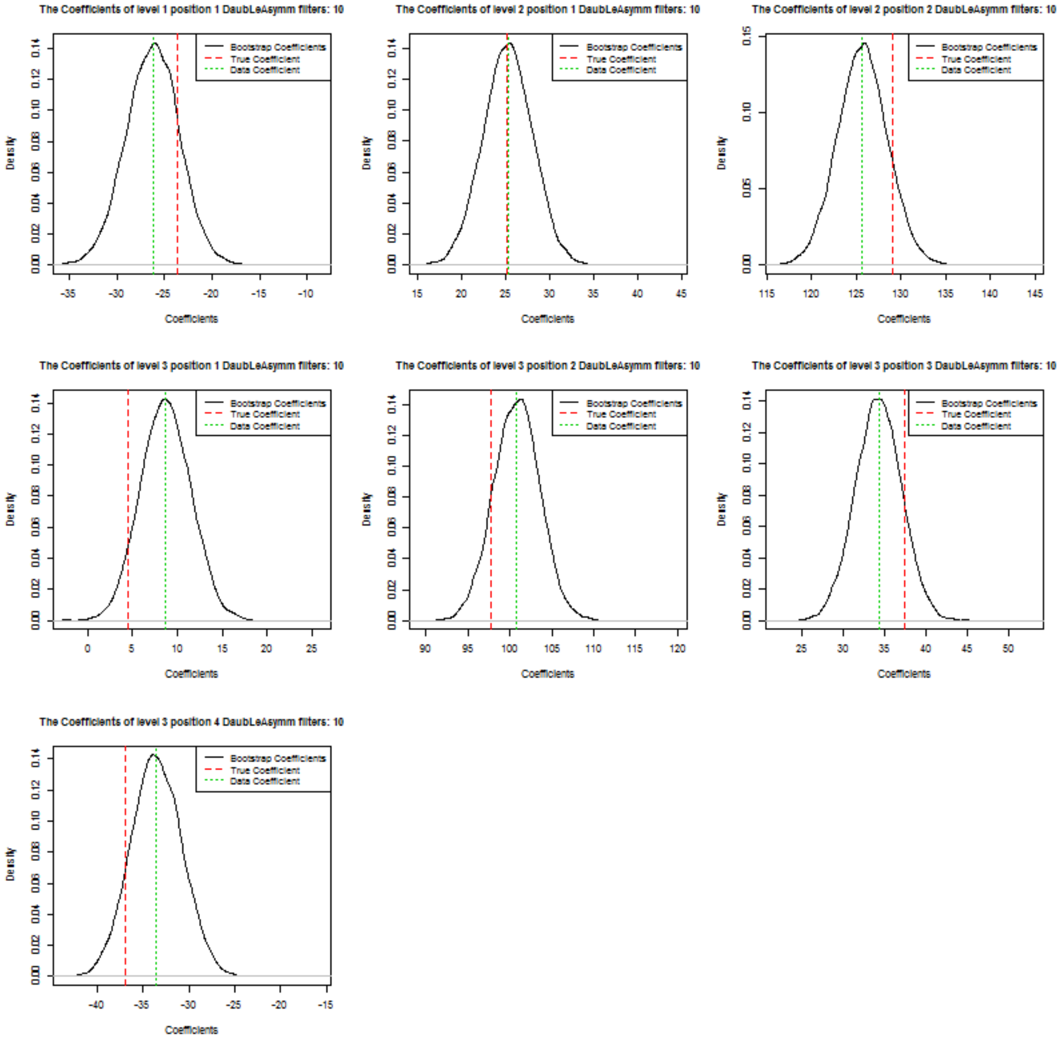


Figure 16: Signal-to-noise Ratio is 2.

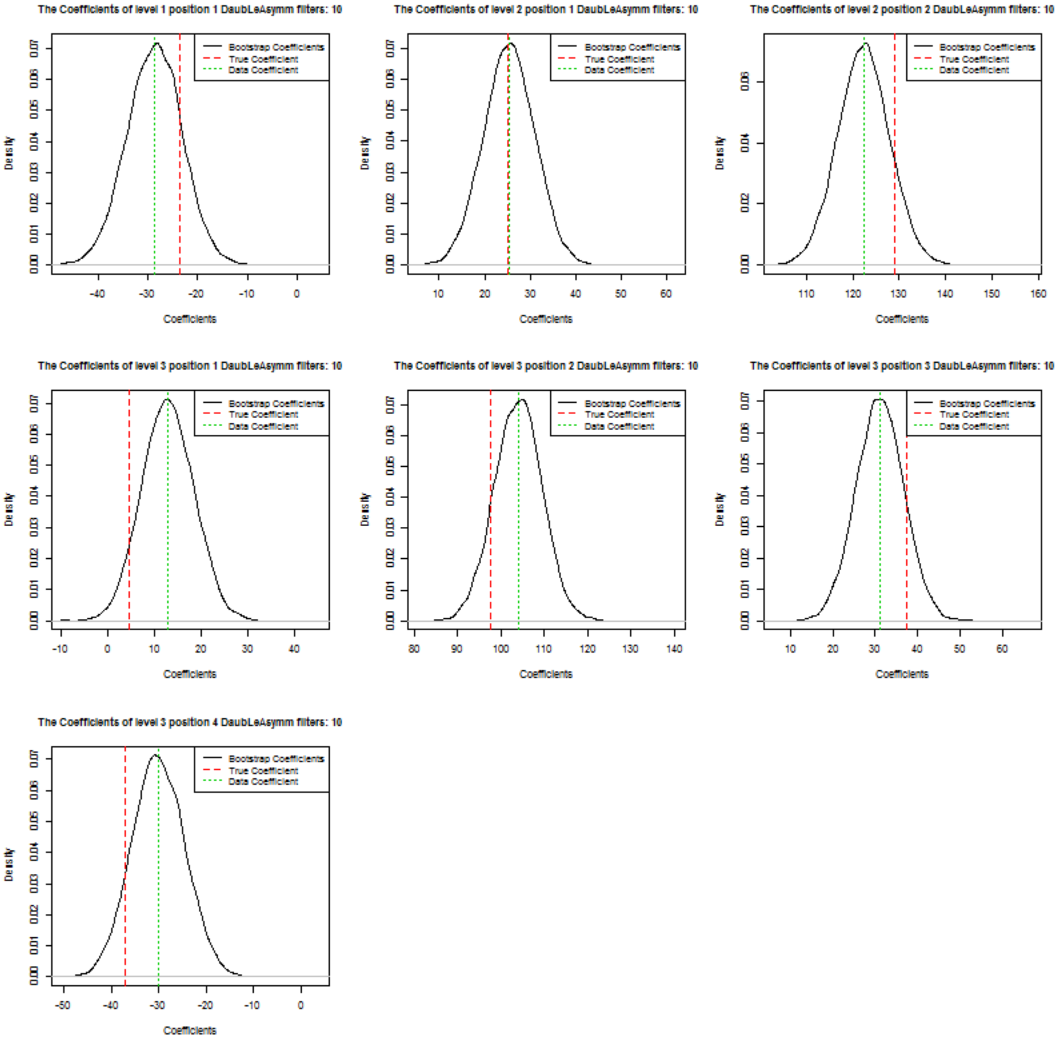


Figure 17: Signal-to-noise Ratio is 1.

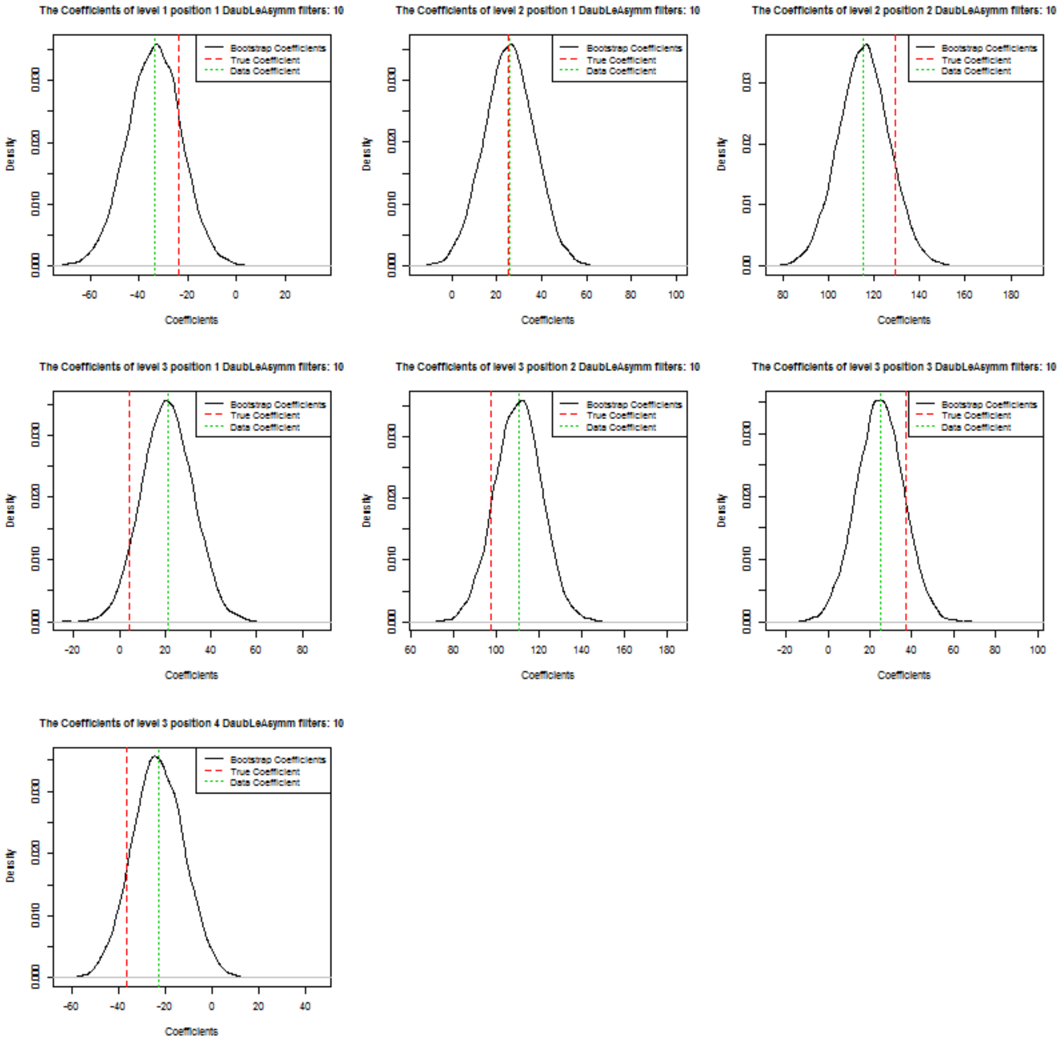


Figure 18: Signal-to-noise Ratio is 0.5.

In the Figures 16, 17, and 18, we can see all distributions are bell-shaped, symmetric, and centered at the wavelet coefficients of the simulated data. All the bootstrap coefficients capture the truth wavelet coefficients. As the signal-to-noise ratio becomes smaller and smaller, the standard deviation of the bootstrap coefficients becomes larger and larger, but the changes in the standard deviation of the bootstrap coefficients are not very dramatically. Furthermore, the differences between the truth coefficient and simulated data coefficient trend to become larger as the signal-to-noise ratio becomes smaller. Based on the results of the bootstrapping, there might not be sufficient evidence, by constructing a confidence interval, to support that the coefficients in the first, third and fourth positions of level 3 are statistically significant if the signal-to-noise ratio is very small. However, there is sufficient evidence to support that all the other coefficients are statistically significant with different signal-to-noise ratio in this simulation. Thus, based on the simulation results, residual bootstrap works very well on the wavelets if the noise comes from normal distribution and we selected the correct wavelet family and filter.

B.2 Tuning the Error (Noise) Distributions

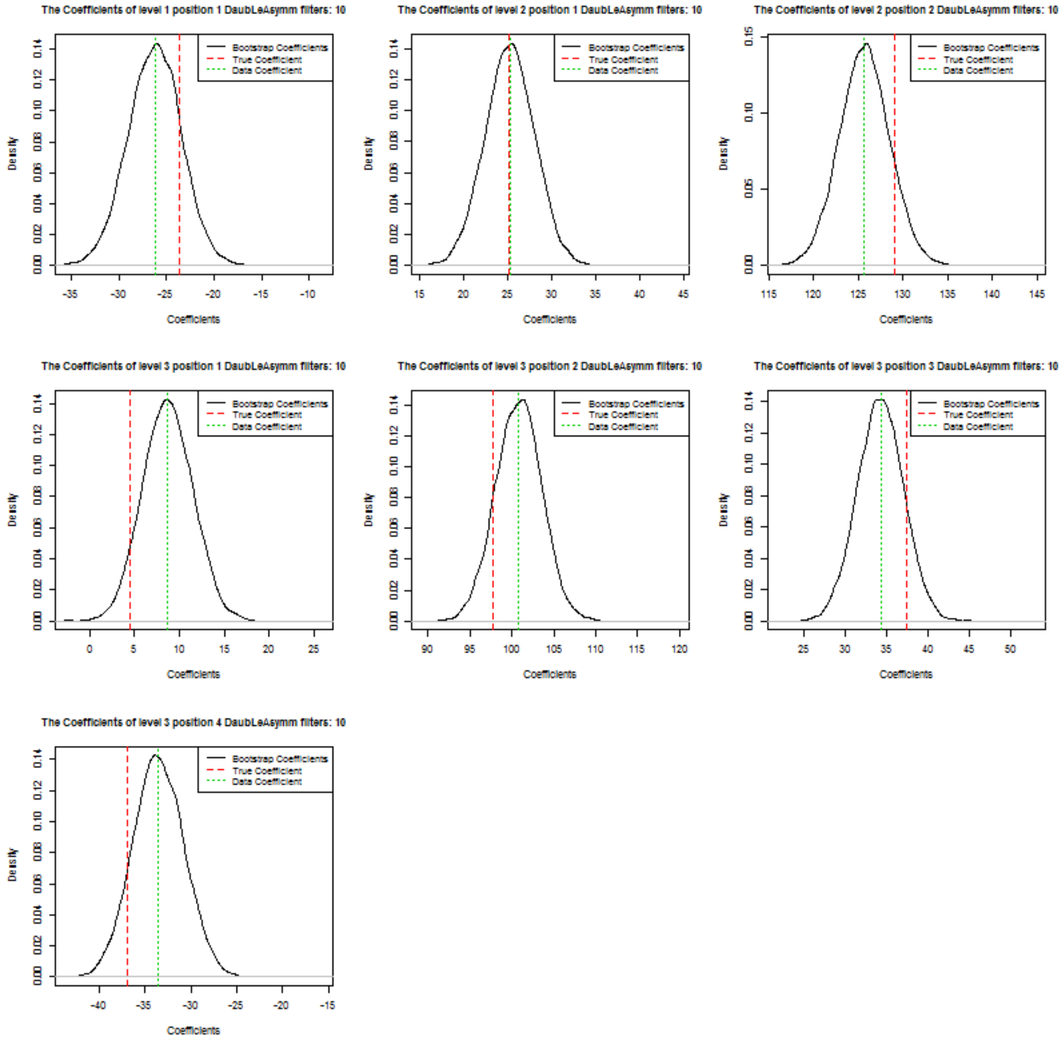
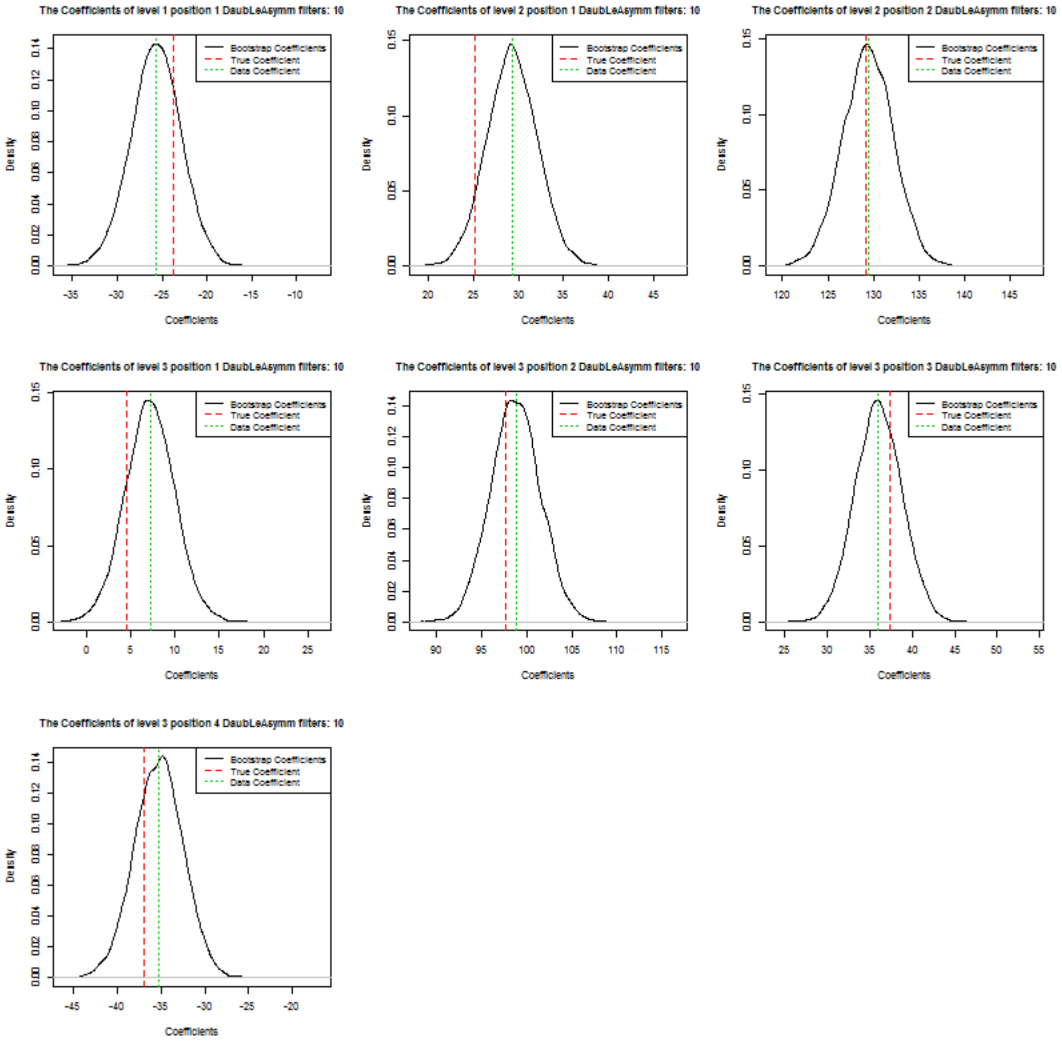


Figure 19: The errors come from normal distribution.



The Coefficients of level 1 position 1 DaubLeAsymm filters: 10

The Coefficients of level 2 position 1 DaubLeAsymm filters: 10

The Coefficients of level 2 position 2 DaubLeAsymm filters: 10

The Coefficients of level 3 position 1 DaubLeAsymm filters: 10

The Coefficients of level 3 position 2 DaubLeAsymm filters: 10

The Coefficients of level 3 position 3 DaubLeAsymm filters: 10

The Coefficients of level 3 position 4 DaubLeAsymm filters: 10

Figure 20: The errors come from uniform distribution.

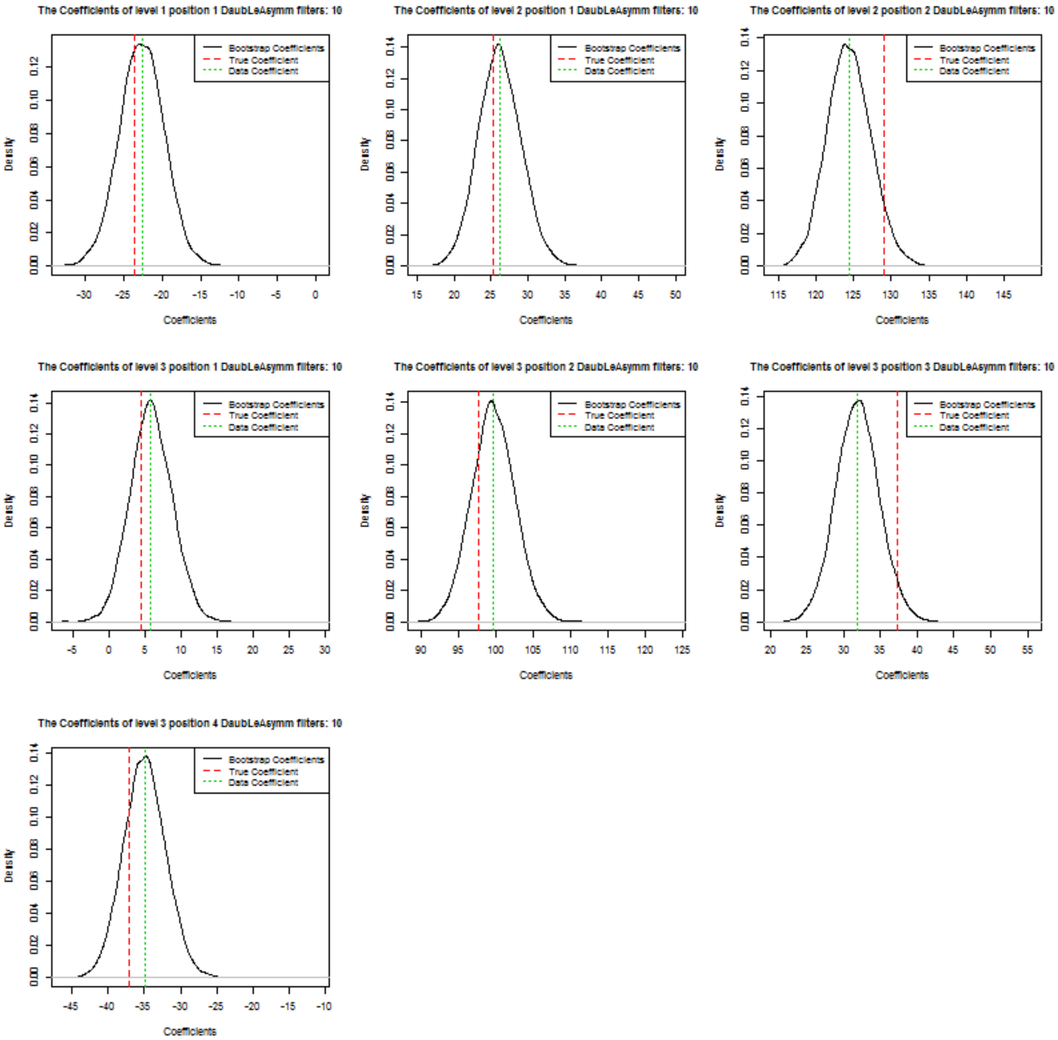


Figure 21: The errors come from lognormal distribution.

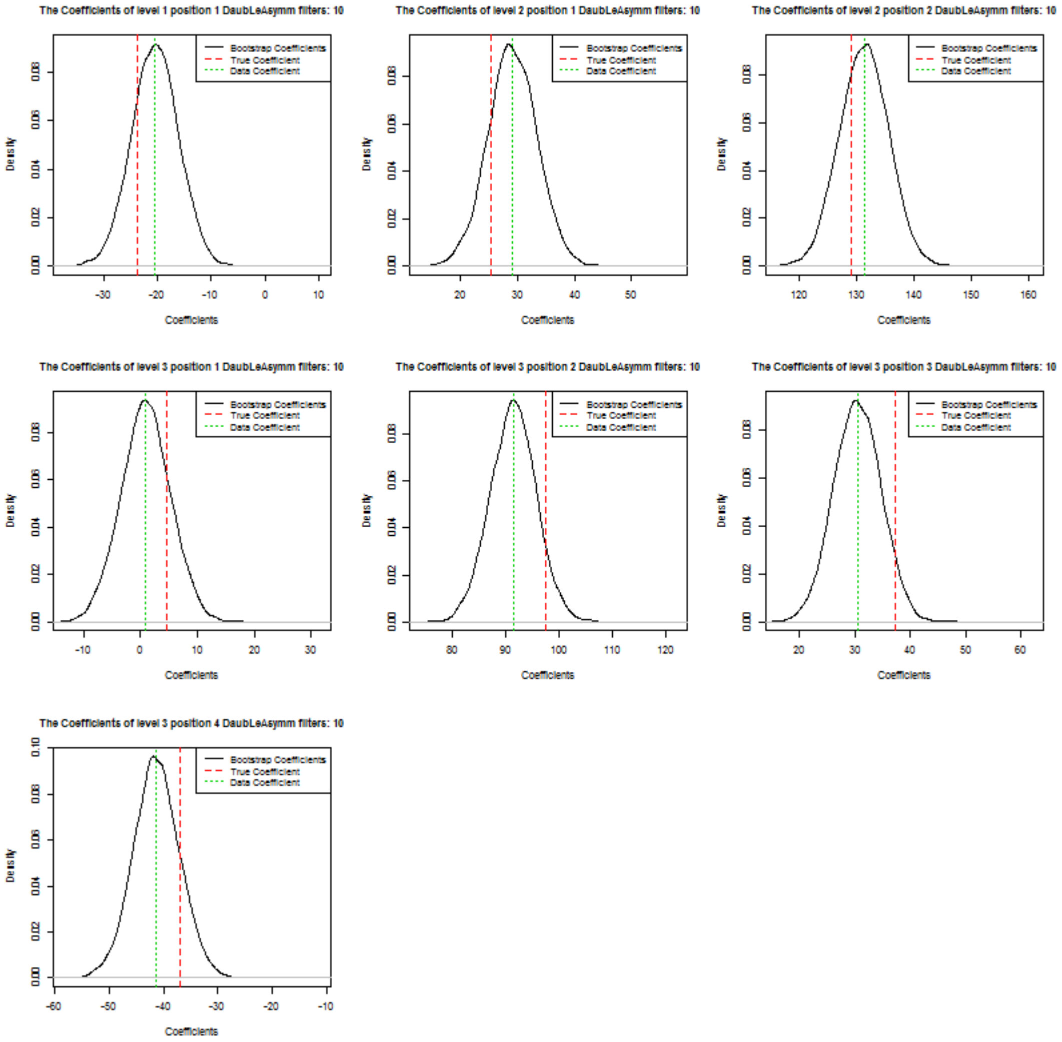


Figure 22: The errors come from T Distribution with Degree of Freedom: 2.

In the Figures 19, 20, 21, and 22, we can see all distributions are bell-shaped, symmetric, and centered at the wavelet coefficients of the simulated data. All the bootstrap coefficients capture the truth wavelet coefficients. With each noise distribution, the differences between the truth coefficients and the simulated data coefficients trends to get larger in the higher levels. In this simulation, the differences between the truth coefficients and the simulated data coefficients seem to be the smallest with uniform noise distribution among four distributions. The signal-to-noise ratios are set to 2 for all four distributions. With normal or uniform noise distributions, there is sufficient evidence, by constructing a confidence interval, to support that all the coefficients are statistically significant. However, there isn't sufficient evidence to support that the coefficients in the position 1 of level 3 are statistically significant if the noise is lognormal distribution or t-distribution with 3 degrees of freedom. Thus, based on the simulation results, residual bootstrap works very well on the wavelets if the noise comes from normal or uniform distributions, the signal-to-noise ratio is larger than 1, and we selected the correct family of wavelet and the number of filters. However, if the noise comes from t-distribution with 3 degrees of freedom or lognormal distribution, the ranges of bootstrap coefficients are larger than the normal or uniform distributions. It might fail to recognize the coefficients with small values.

B.3 Employing the Wrong Family of Wavelets/Filter Numbers

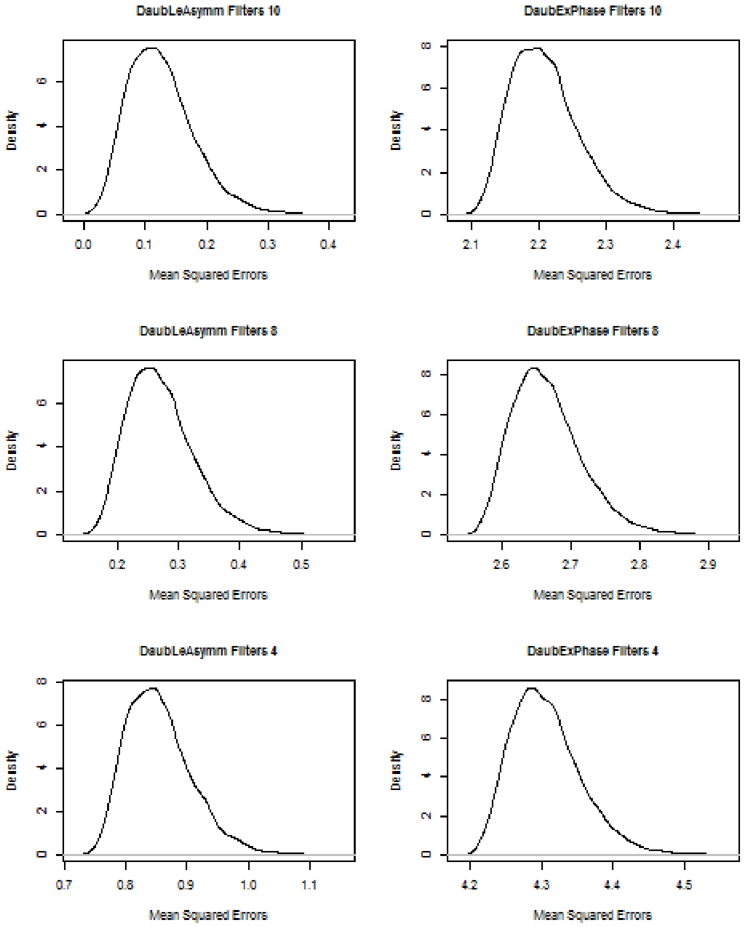


Figure 23: The bootstrap distribution of mean squared errors between true signal and signal captured by DWT

In the Figure 23, we can see all distributions of mean squared errors between true signal and wavelet signal with different wavelet family and the number of filters are bell-shaped and right-skewed. The spread and the shape of all the distributions above are almost the same. Wrong wavelet family or the number of filters only shift the curves to the right. The centers of the mean squared errors get larger as the difference between the true filter number and the utilized filter number gets larger or wrong wavelet family is employed. The wrong filter numbers only slightly increase the mean squared errors, but the wrong wavelet family will boost the mean squared errors dramatically.