

Tree Cover Variability Increases from 2005 to 2100 in Sub-Saharan Africa

Cody Carroll, Eric Kalosa-Kenyon, Amy Kim

December 11, 2017

Abstract

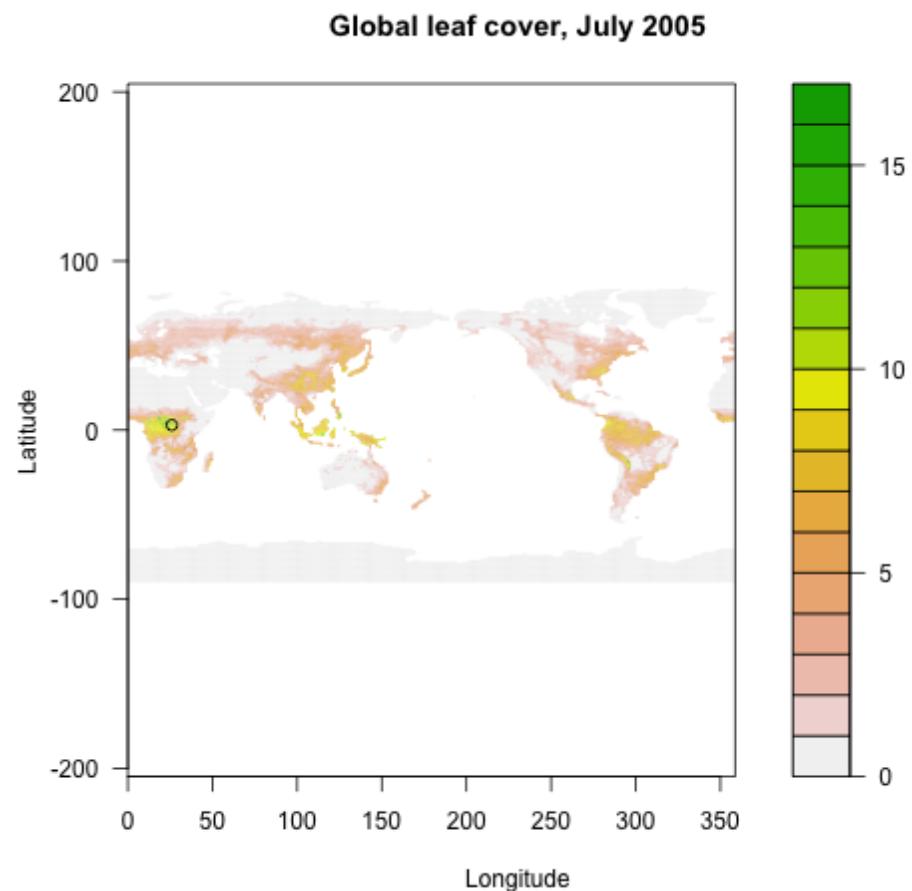
This report fulfills the requirements for the final project in Time Series Analysis (STA237A). The authors report on the behavior of simulated tree cover between the years of 2005 and 2100. The preliminary analysis using common time series methods is followed by changepoint and spatio-temporal analyses.

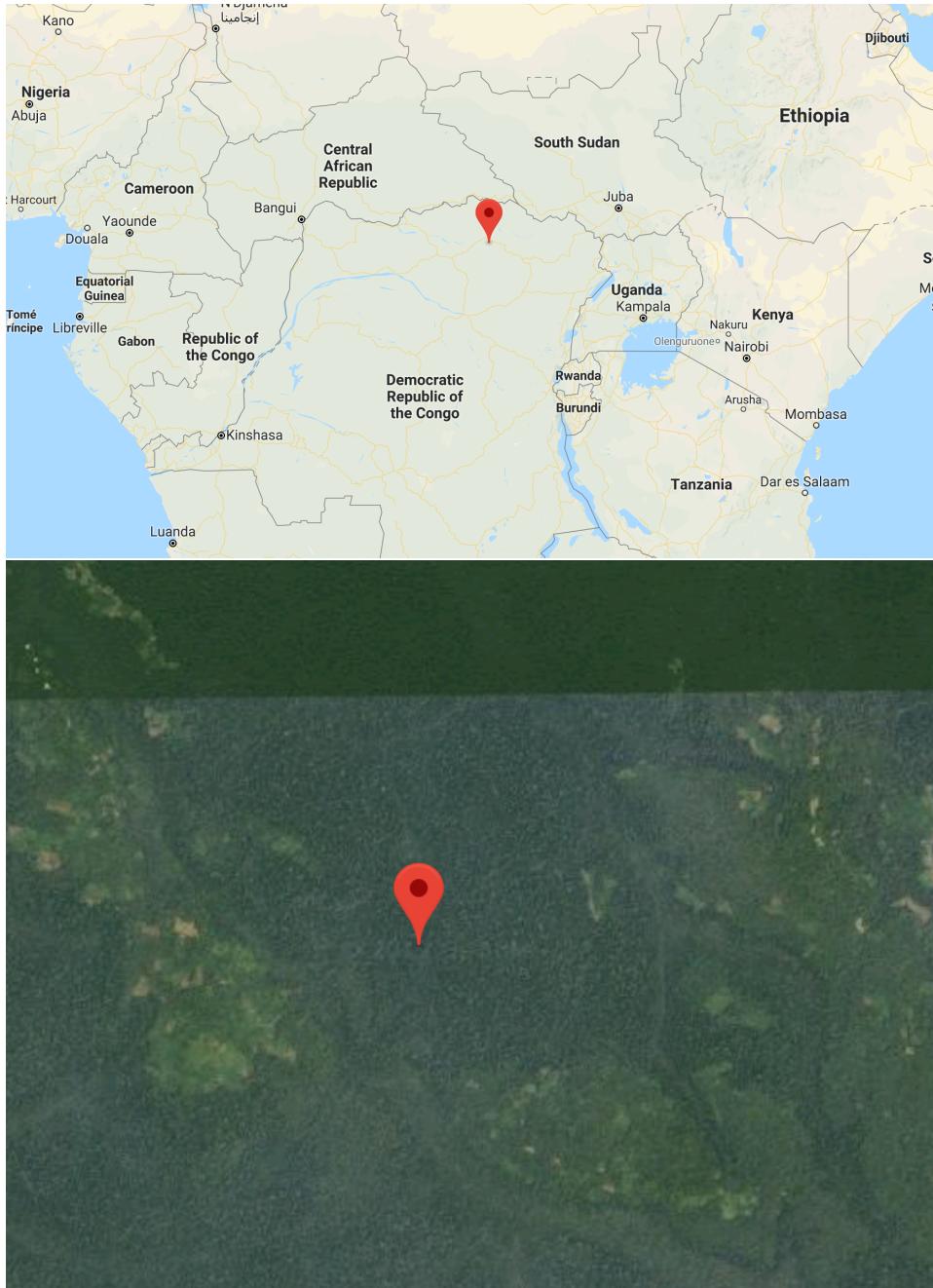
All authors contributed equally to this work.

1 Introduction

Climate change is driven by anthropogenic carbon forcing. Predicated on this inconvenient reality, the Intergovernmental Panel on Climate Change has developed representative carbon forcing pathways (RCPs) to coordinate climate research. Ensembles of computer simulations are run using pre-selected RCPs and calibrated using historical observations. These simulations are realizations of gridded meteorological PDEs in the Community Climate Model System (CCMS). The CCMS has 5 components and a coupler: atmosphere, sea, land, sea ice, and land ice. These modules each influence the global climate and each module generates output for subsequent analysis.

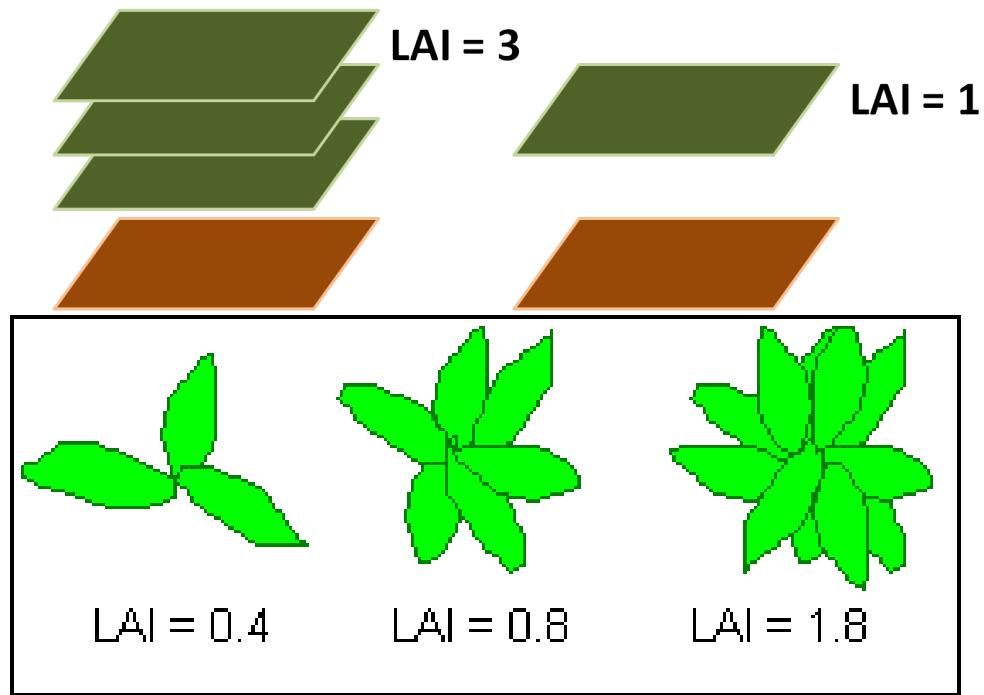
In this report, we interrogate the output of the land module both globally and over a single location in Sub-Saharan Africa. The ensemble used to generate this data follows the RCP4.5 carbon pathway. Roughly, this pathway captures the atmospheric carbon concentration induced by peak carbon emissions around 2035 with a decrease thereafter. We are interested to understand how a highly productive region's tree cover responds to this RCP. The location in question is depicted in the figures below.



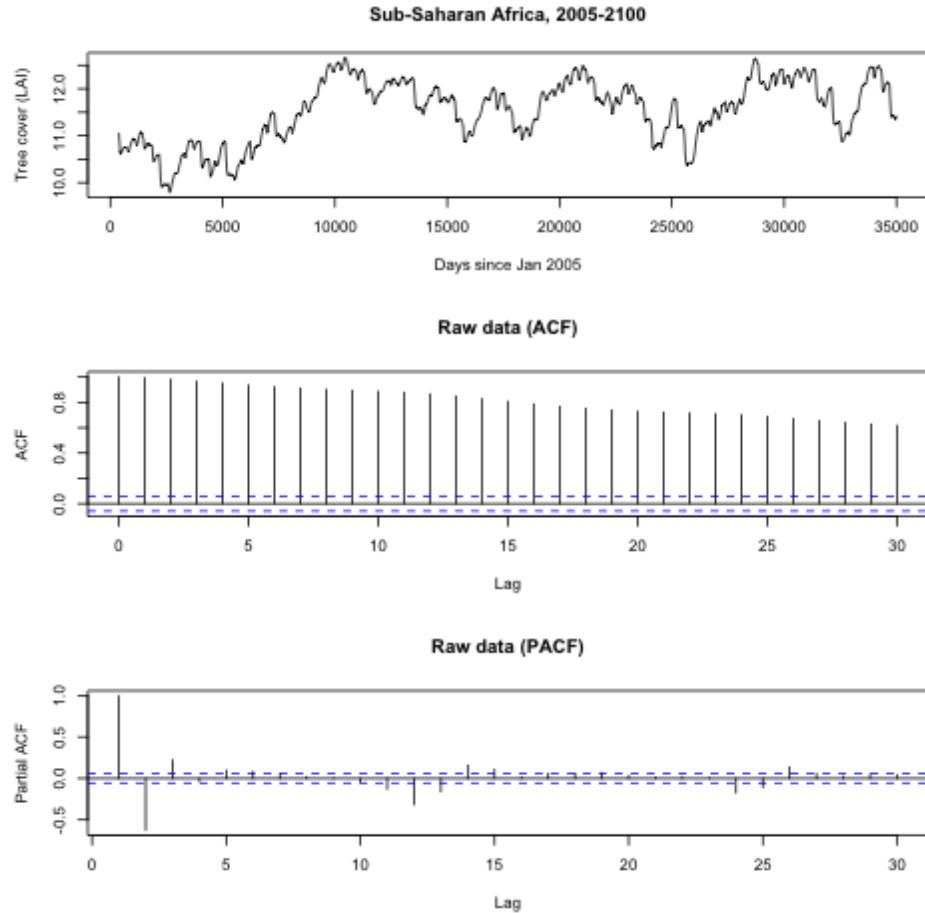


The location we selected is at 3.29 degrees North and 26.25 degrees East. It is in an equatorial/monsoon climate and is home to dense tropical rainforest.

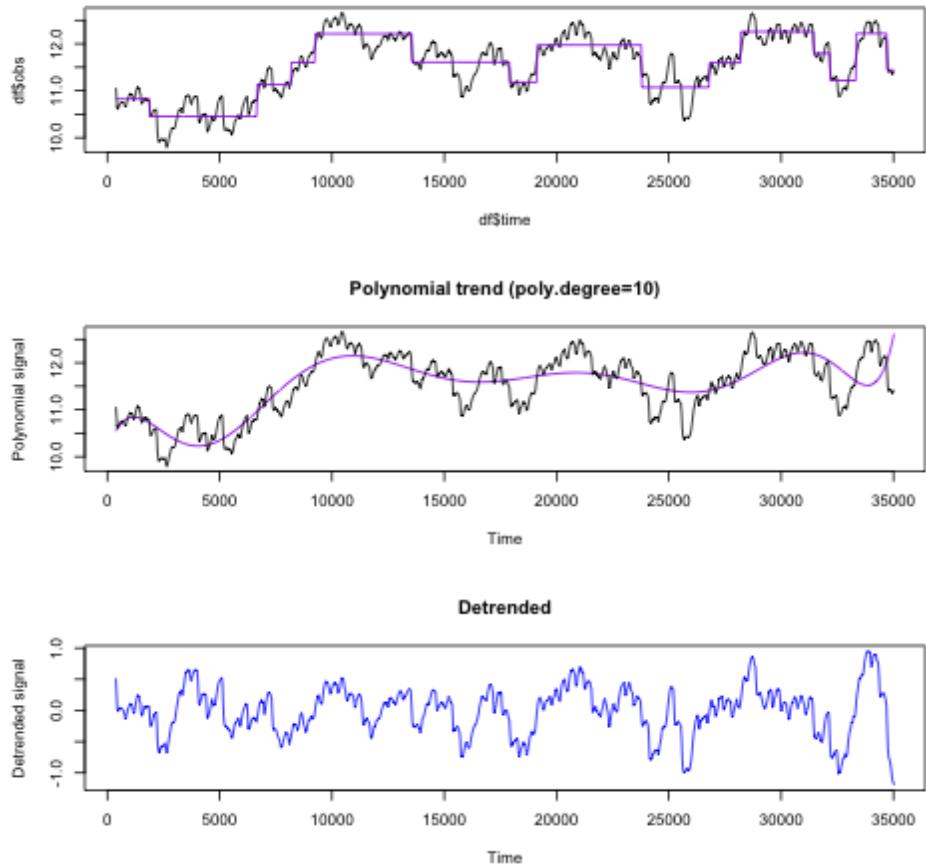
To measure projected tree cover, we study simulated leaf area indices (LAI). LAI is unitless quantity ($\text{leaf area } (m^2)/(\text{ground area } (m^2))$) first defined in 1947 by Watson in the Annals of Botany ("Comparative Physiological Studies on the Growth of Field Crops"). As a measure of tree cover, it is bounded below by 0 and above by physiological limits. The following two google-image search results are illustrative guides for the intuition behind using LAI as the metric for tree cover.



The raw data at a single location is presented in the figure below:

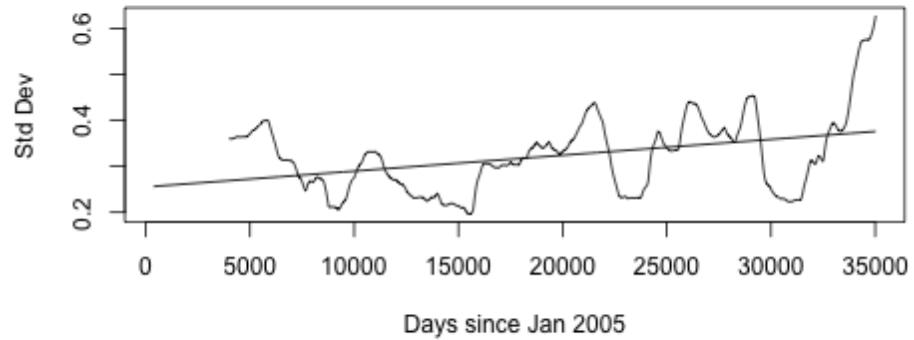


Noticing that the raw signal is highly nonstationary (the autocovariance function appears to change over time), we decided to remove long-term and seasonal trends. Two common methods for long-term detrending are polynomial fit and complexity-criterion-restricted (e.g. MIC) moving averages. Both methods are shown below. We move forward using the polynomial regression detrending method because it preserves the quasi-continuity of the data. The residual signal after detrending with the polynomial regression is shown below:

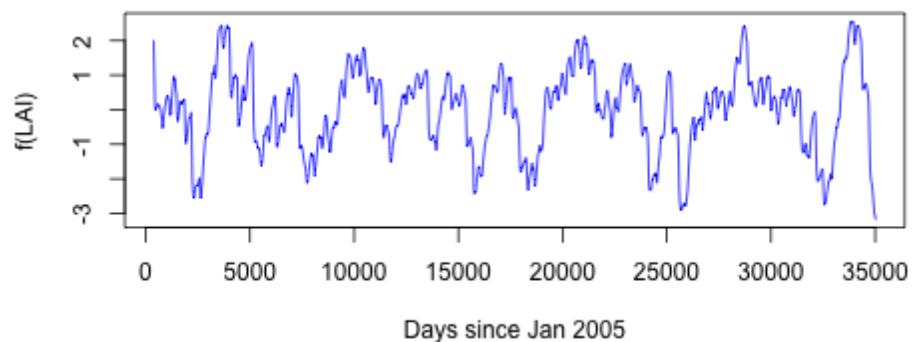


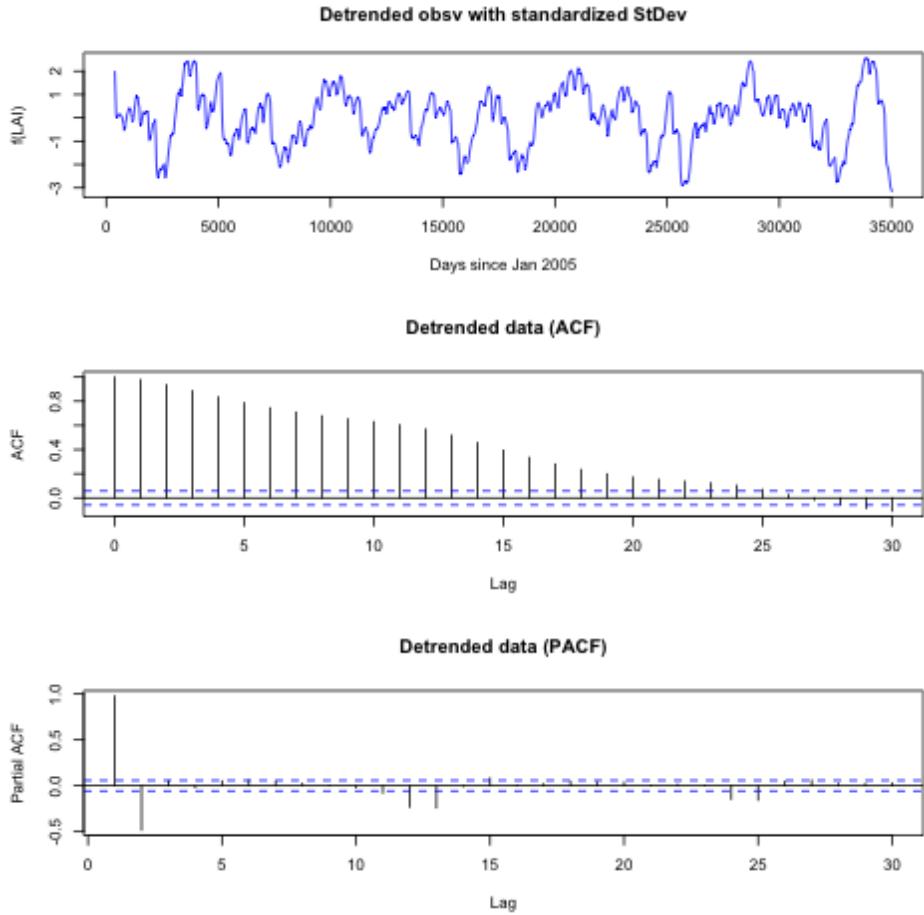
To standardize the variance of the data, we use a running standard deviation with a window-size of 10 years. We regress linearly the running standard deviation across all 95 years of signal and divide the detrended signal by the regression output. The resultant variance-standardized data at a single location is presented in the figures below:

Running standard deviation 10 year window

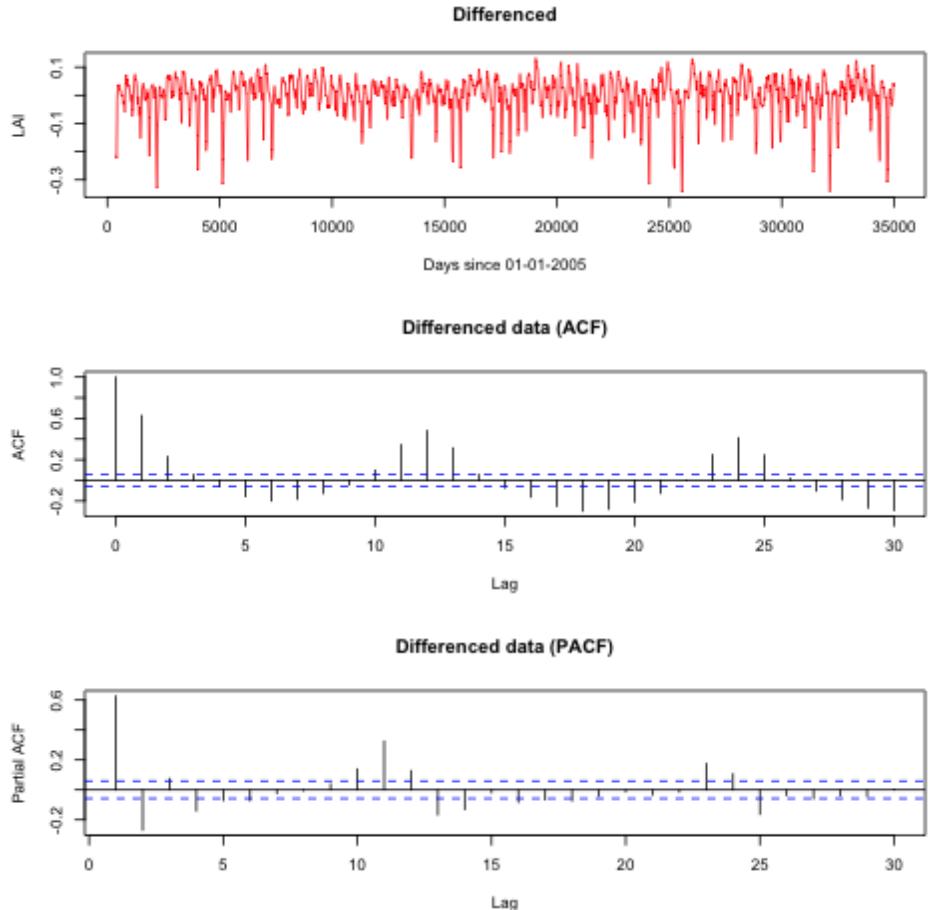


Detrended obsv with standardized StDev

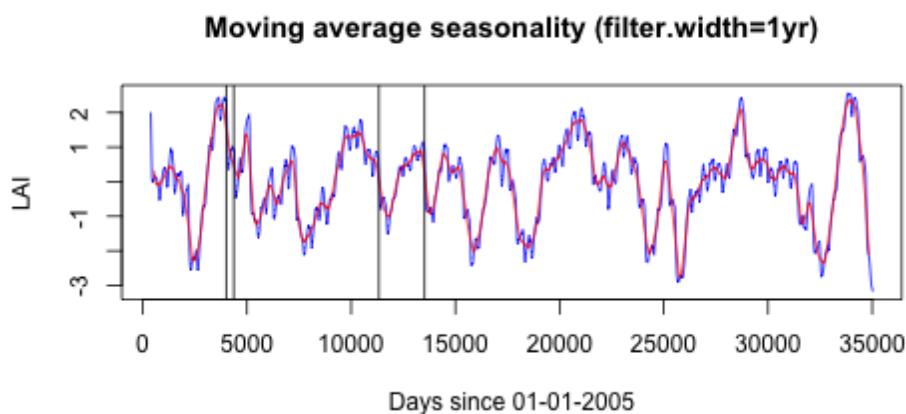
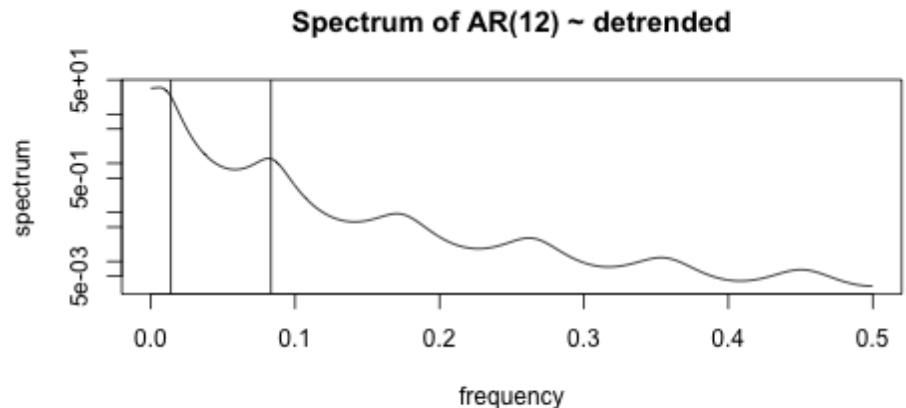




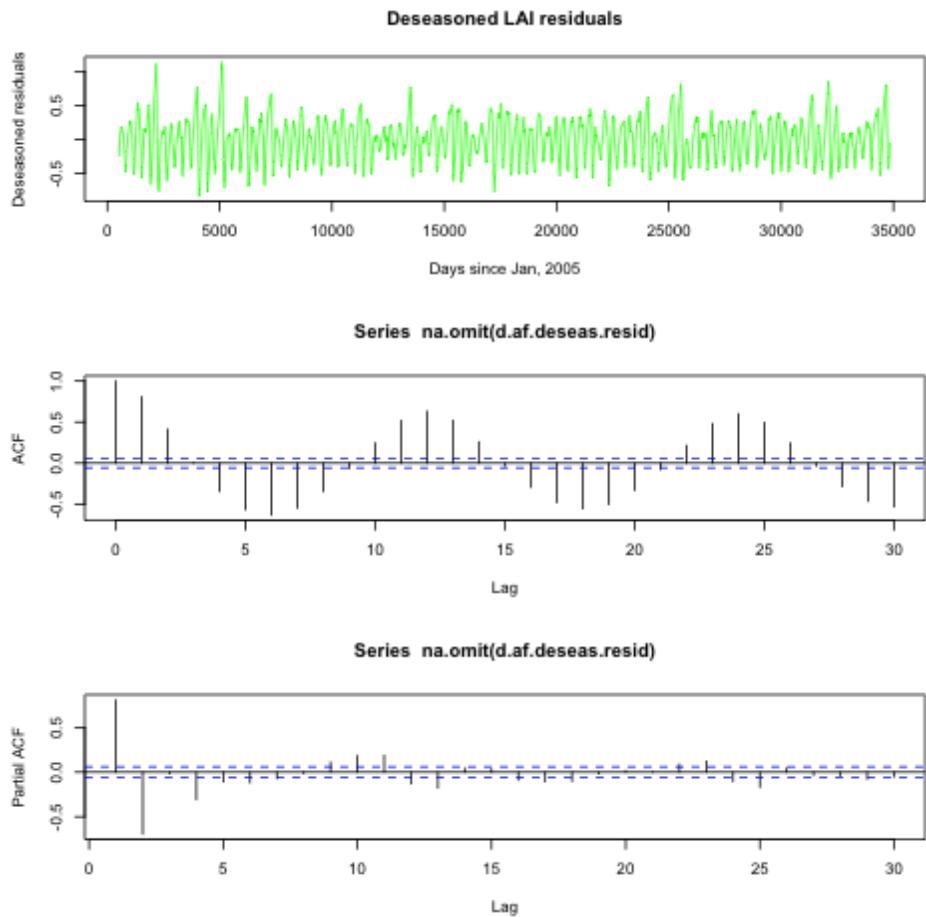
In an attempt to further stationarize the signal, we apply a first differences filter to the deseasonalized data. The differenced data at a single location is presented in the figure below:



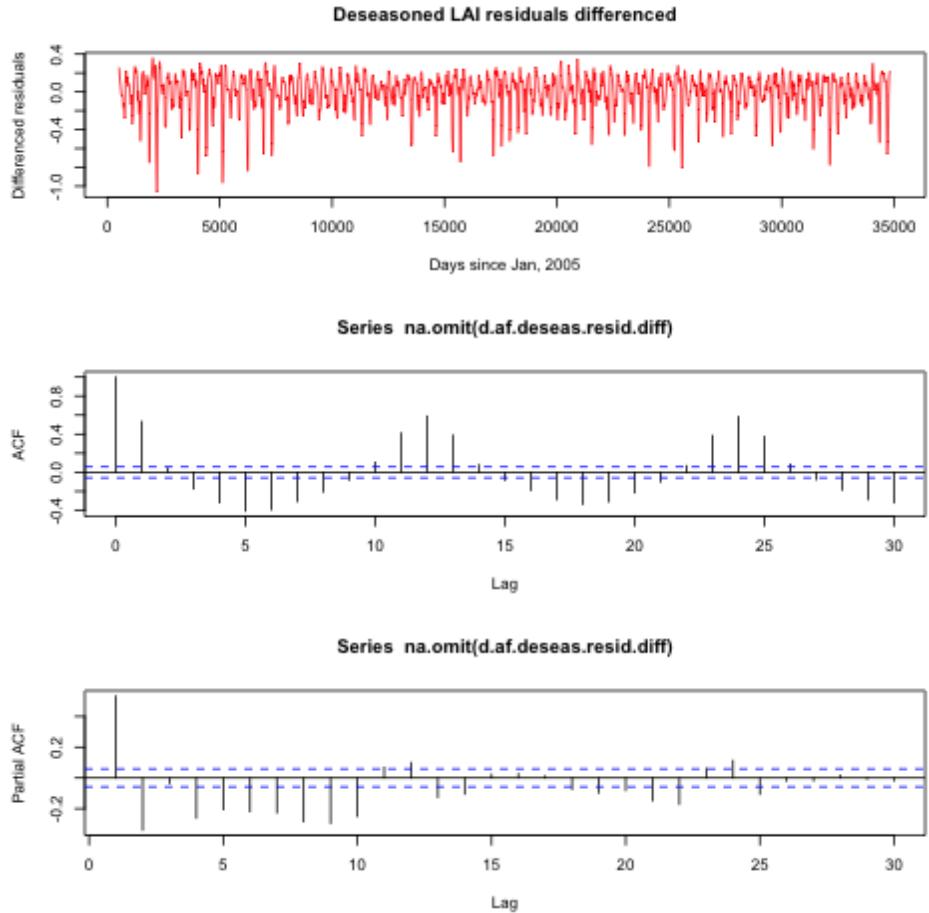
The differenced data is not stationary, so we move forward with an analysis of the deseasonalized data, as follows. The spectrum of an autoregressive process with a year's memory ($AR(12)$) is depicted in the figure below.



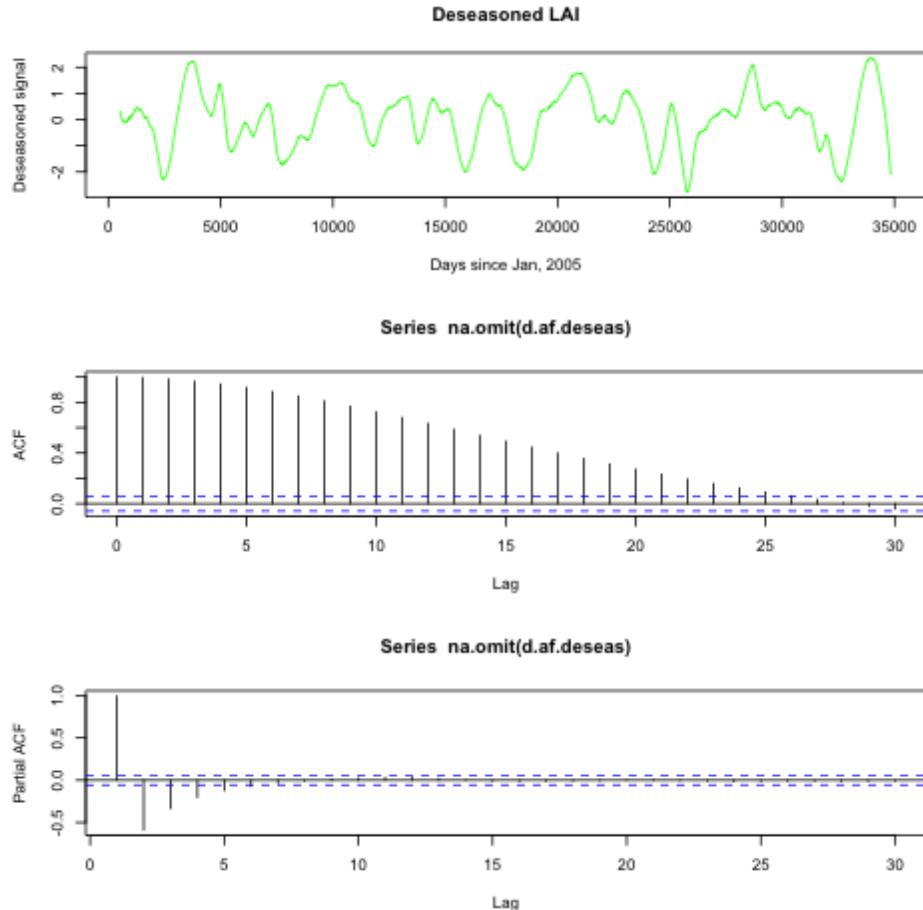
The $AR(12)$ model regressed on the deseasonalized data still shows strong signals at annual and multi-annual frequencies, indicating that there is still seasonal structure left over. To address some of this structure, we fit a running average with a window-size of one year to the data, shown above in red. The following figure depicts the residuals of the normalized Sub-Saharan LAI with long-term and annual seasonal trends removed.



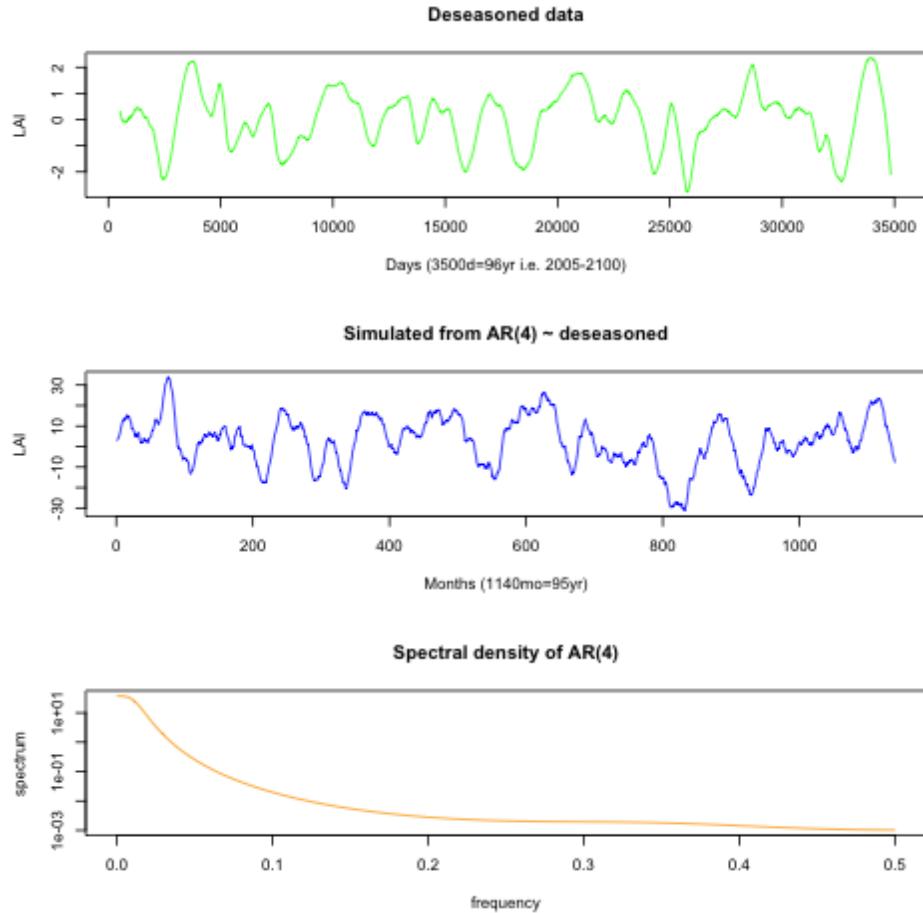
Noticing that there is still substantial structure in the deseasonalized residuals, we apply a first differences filter to the deseasonalized residuals. This differencing yields the following signal.



The residuals in this case appear to contain most of the non-stationary signal. The following figure depicts the 12-month moving average of the normalized detrended signal. There are 4 prominent legs of the PACF, so we fit an AR(4) model to the moving average.



As described above, the normalized, deseasonalized data has an AR(4) structure suggested by its PACF. We fit an AR(4) and simulated some data from the process, depicted in the following figure.



The coefficients for this AR(4) model fit to the deseasonalized signal are $\Psi_{1:4} = 1.316, -0.073, -0.054$, and -0.203 . The peak frequency on the lefthandside of the spectral density plot suggests that the only prominent periodic signal left over is the long term trend.

2 Changepoint Analysis

Instead of dealing with the non-stationarity of our series by differencing and using the ARIMA framework, another approach is to notice that there are specific points in time where the time series changes behavior in one fell swoop. An example of this is around the year 2026, where the series jumps up to a mean level of around 11.75 and stays there. Figure 1 illustrates this "jump" with a red dashed line.

Sub-Saharan Africa LAI, Changepoint in Mean at 2026

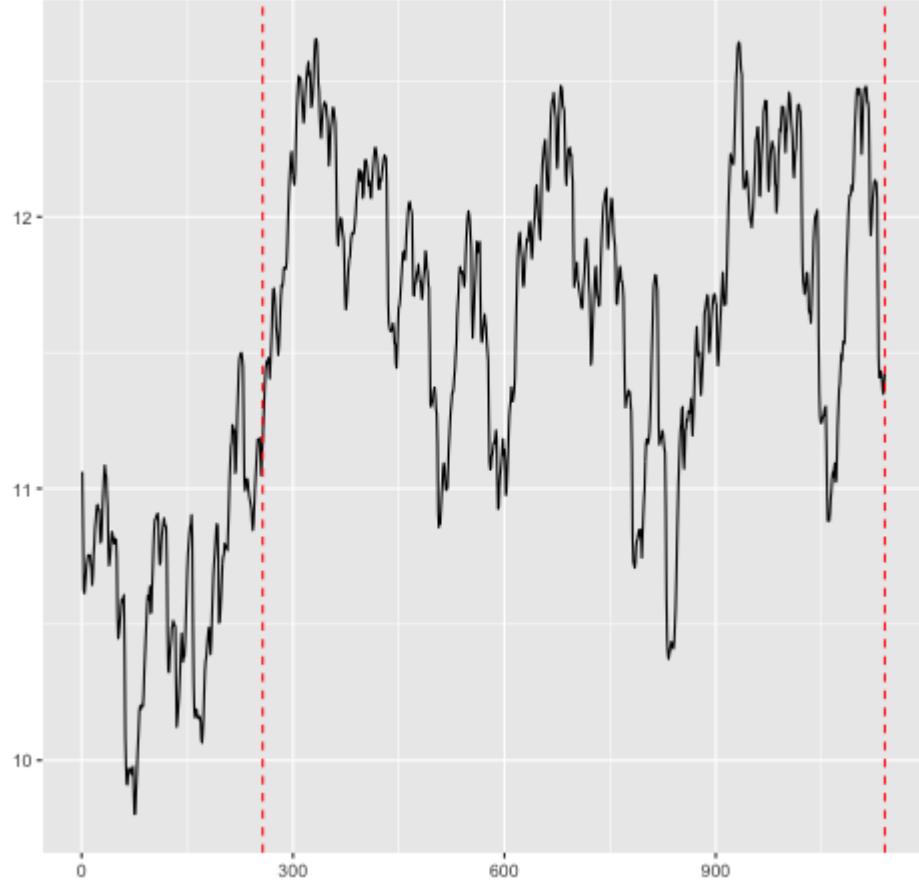


Figure 1: LAI TS

This point in time is called a changepoint. We give a short introduction on changepoint detection methods.

2.1. Changepoint Detection

Conceptually, for data Z_1, \dots, Z_n , a changepoint τ is a point in time, such that Z_1, \dots, Z_τ differ from $Z_{\tau+1}, \dots, Z_n$. For the sake of illustration, we take a simple example. Assume the parametric form

$$Z_t | \theta_t \sim N(\theta_t, 1). \quad \text{where } \theta_t = 1 \text{ if } t \leq \tau \text{ and } \theta_t = 0 \text{ if } t > \tau$$

An example of such a process is depicted in Figure 2.

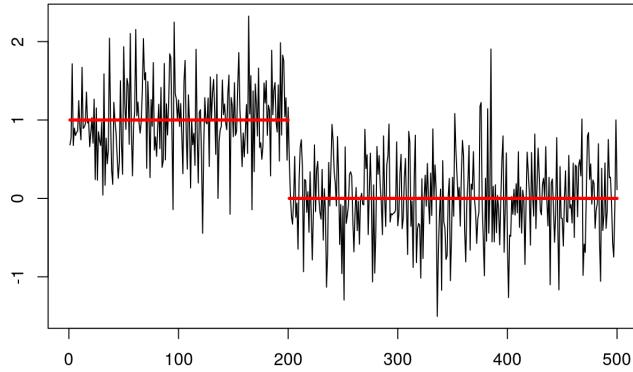


Figure 2: Example of a Changepoint

It's intuitively clear from the picture that the change occurs at $\tau = 200$, but how do we detect where a changepoint occurs, mathematically? One approach is to use a Likelihood Ratio Test. We state the hypotheses as:

$$H_0 : \theta_t = \theta \quad \forall t \quad \text{vs.} \quad H_1 : \theta_t = \theta_1, \quad t \leq \tau; \quad \theta_t = \theta_2, \quad t > \tau$$

To construct the likelihood ratio statistic, we need to maximize the likelihood under the null and alternative hypotheses. We let $p(\cdot)$ be the probability density function associated with the distribution of the data and $\hat{\theta}$ is the maximum likelihood estimate of the parameters. Under H_0 , the maximum log-likelihood is $\log p(y_{1:n}|\hat{\theta})$. Under H_1 with a changepoint τ , the maximum log-likelihood is $ML(\tau) = \log p(y_{1:\tau}|\hat{\theta}_1) + \log p(y_{\tau+1:n}|\hat{\theta}_2)$. Therefore the likelihood ratio statistic is:

$$\Lambda(\tau) = \frac{\max_{\tau} ML(\tau)}{\log p(y_{1:n}|\hat{\theta})}$$

A more commonly used version of this statistic is the log-LR statistic:

$$2 \log \Lambda(\tau) = 2 \left[\max_{\tau} ML(\tau) - \log p(y_{1:n}|\hat{\theta}) \right]$$

We compare this statistic with a cut-off value, λ . If $LR > \lambda$, then we reject H_0 and estimate the changepoint as

$$\hat{\tau} = \arg \max_{\tau} LR(\tau)$$

There are many different ways to select λ . Selecting an optimal value of λ remains an open research topic in changepoint analysis.

2.2. Changepoint Analysis on LAI Time Series

We use the package 'changepoint' to investigate if our LAI time series has a changepoint, using a specific λ value that takes into consideration the inherent correlation between observations of a time series.

We find an estimated changepoint of $\hat{\tau} = 237$ months after Jan. 2005, i.e. 2026 is when LAI experiences a changepoint in mean. The test rejected the null hypothesis that there is no changepoint with a p-value < 0.01 . Looking at the series after 2026 though, we question there may be more time points. So we segment the series and iterate the changepoint estimation. Since we are doing several tests now, we need to adjust our α level. I rerun the original test and also test on the subsegment after the changepoint at $\alpha^* = 0.05/2$, so that tests have an overall level of $\alpha = 0.05$. The first test finds the same changepoint is still significant. However, the second test finds no significant changepoint (Figure 3) in the remaining data:

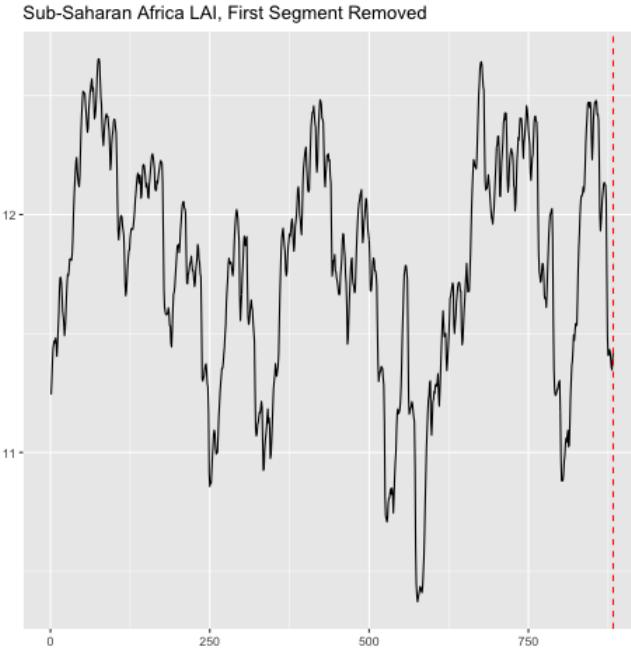


Figure 3: LAI Time Series After 2026

We conclude that the only significant changepoint in leaf area index in sub-Saharan Africa occurs in 2026.

This data is simulated under the setting that carbon emissions peak in the year 2040. Observing a changepoint in 2026 illustrates the ramifications of carbon emissions at the current rate, before they even reach their expected maximum level. We stay cautious with our conclusions and rely on the climatologists and domain experts to spell out the actual science, but the results from our analysis are highly suggestive of real effects of increasing carbon emissions on the global environment.

3 Principal Component Analysis

We have done the time series analysis for the one location, and we can expand our analysis into global scale. This becomes the analysis of space-time with a large dataset (12818 locations and 1140 time points). In order to extract the underlying trends, we can consider Principal Component Analysis for examining both the spatial and temporal variation here.

Principal Component Analysis

$$\mathbf{Z}(s, t) = \mathbf{U}\Lambda\mathbf{V}^T \quad (1)$$

where \mathbf{U} is a $T \times k$ orthogonal matrix with columns \mathbf{u}_j , \mathbf{V} is a $S \times k$ orthogonal matrix with columns \mathbf{v}_j and Λ is a $k \times k$ diagonal matrix with diagonal entries λ_j . This \mathbf{Z} :

$$\mathbf{Z}(s, t) = \begin{pmatrix} z(s_1, t_1) & z(s_2, t_1) & z(s_3, t_1) & \dots & z(s_{12818}, t_1) \\ z(s_1, t_2) & z(s_2, t_2) & z(s_3, t_2) & \dots & z(s_{12818}, t_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z(s_1, t_{1140}) & z(s_2, t_{1140}) & z(s_3, t_{1140}) & \dots & z(s_{12818}, t_{1140}) \end{pmatrix} \quad (2)$$

We decide to report PCA on a detrend dataset ¹ since its first few Principal Components(PCs) have more cumulative explained variances. We mainly examine the first Principal Component(PC) since it explains 43% of variances and others do less than 10% (Table 2) ².

3.1 Spatial Pattern

Spatial pattern explains how strong the PCs depend on some locations, and it is represented by the loadings of each principal components. The mean

¹We used cubic splines for each time series to extract variations

²419 PCs achieve to explain 90% of variations.

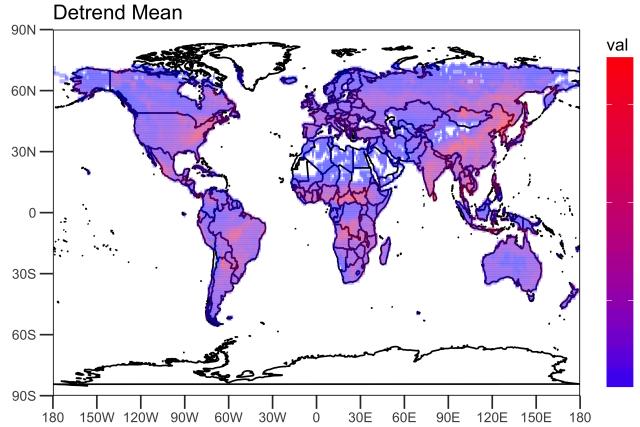


Figure 4: Overall Spatial Mean

spatial structure, Figure 4, indicates locations known forest area higher mean values(red) and desert areas have lower values (blue). We can interpret PC1 implies main variance across the all locations and over the 95 years in Figure 5. It shows forest areas have negative effects (blue) and infertile lands have positive effects (red) which contracts to mean structure.

3.2 Temporal Pattern

Temporal pattern explains the dominant temporal variation of time series in the all locations, and it is represented by principal components (PCs, a number of time series) of PCA. We can confirm the detrend first three PC are stationary in Figure 6, and the first PC has widest range of oscillation(black) and ranges of oscillation get smaller (Blue is the second PC, and red is the third.)

We have found the PCs have seasonality through ACF in Figure 10. PC1 and PC2 have annual seasonality, PC3 and PC4 have semi-annual, and PC5 and PC6 have quarterly seasonality, which get supported by periodograms, Figure 13. It has positive and negative sides, which makes sense because the dataset has covered both north and south hemispheres. Additionally, PC1 and PC2 seems to have similar structures behind as well as PC3 and PC4 and PC5 and PC6 via the Cross-Covariance Functions in Figure 12. Since those have the obvious seasonality, we conduct spectral analysis on PC 1.

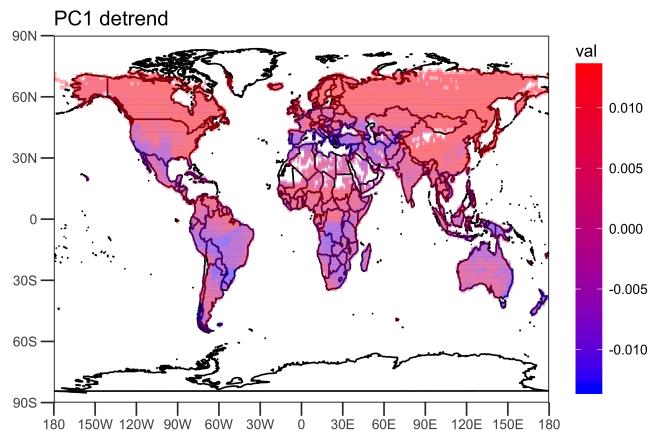


Figure 5: Spatial Pattern of PC 1

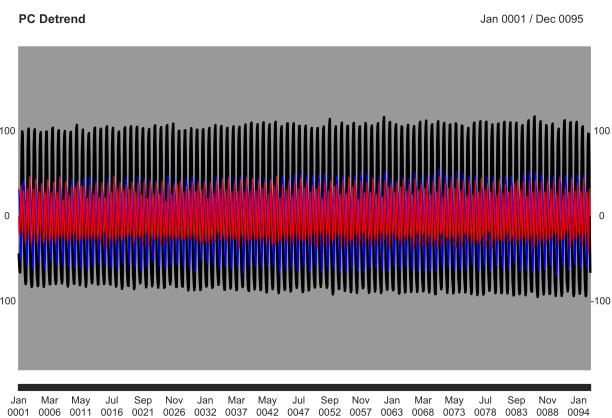


Figure 6: PC temporal patterns of Detrend Data

Spectral Analysis We model the first principal component (PC1) which has annual seasonality as:

$$X_t = A \cos(2\pi \frac{1}{12}t) + B \sin(2\pi \frac{1}{12}t) \quad (3)$$

$$= R \sin(2\pi \frac{1}{12}t + \varphi) \quad (4)$$

$$\gamma(h) = \sigma^2 \cos(2\pi \frac{1}{12}h) \quad (5)$$

where $R^2 = A^2 + B^2$, $\varphi = \arctan(\frac{A}{B})$.

Here is our fitted model:

$$\hat{X}_t = -32.652 \cos(2\pi \frac{1}{12}t) - 97.1938 \sin(2\pi \frac{1}{12}t) \quad (6)$$

$$= 102.5319 \sin(2\pi \frac{1}{12}t + \frac{\pi}{10}) \quad (7)$$

$$\hat{\gamma}(h) = 13.93^2 \cos(2\pi \frac{1}{12}h) \quad (8)$$

This model can explains 96% variances (Adjusted R^2 is 0.9644). The PC 2 can be fitted by

$$\hat{X}_t = 48.96085 \sin(2\pi \frac{1}{12}t - \frac{2\pi}{5}), \hat{\sigma}^2 = 8.17 \quad (9)$$

which is shifted and smaller oscillations with respect to PC1 model.

Interestingly, we find another cycles from the residuals of the PC1 model. If we allow to add more cyclic variables, we could end up:

$$\begin{aligned} X_t &= c_1 \cos(2\pi \frac{1}{12}t) + c_2 \sin(2\pi \frac{1}{12}t) + c_3 \cos(2\pi \frac{1}{6}t) + c_4 \sin(2\pi \frac{1}{6}t) \\ &\quad + c_5 \cos(2\pi \frac{1}{4}t) + c_6 \cos(2\pi \frac{1}{3}t) + c_7 \sin(2\pi \frac{1}{3}t) + c_8 \sin(2\pi t) \end{aligned} \quad (10)$$

this model can explain 99.62% variances.

A Simulation

All analysis in this report was performed on a single ensemble submission to the NCAR CMIP database. The ensemble's Project ID in the CMIP5 database is cmip5.output1.NCAR.CCSM4.rcp45.mon.land.Lmon.r2i1p1.v20140403—esgf-data.ucar.edu. Its version number is 20140403. To download the bash script

that will download the data for you, follow this link: https://esgf-node.llnl.gov/esg-search/wget/?distrib=false&dataset_id=cmip5.output1.NCAR.CCSM4.rcp45.mon.land.Lmon.r2i1p1.v20140403|esgf-data.ucar.edu

B PCA

B.1 Tables

	PC1	PC2	PC3	PC4	PC5	PC6
StDev	66.650	28.003	24.356	20.348	17.832	12.301
Prop. of Var.	0.347	0.061	0.046	0.032	0.025	0.012
Cum. Prop.	0.347	0.408	0.454	0.486	0.511	0.523

Table 1: Explained Variations of PC: Raw Dataset

	PC1	PC2	PC3	PC4	PC5	PC6
StDev	73.857	35.586	23.852	14.581	11.859	10.669
Prop. of Var.	0.426	0.099	0.044	0.017	0.011	0.009
Cum. Prop.	0.426	0.524	0.569	0.585	0.596	0.605

Table 2: Explained Variations of PC: Detrendset

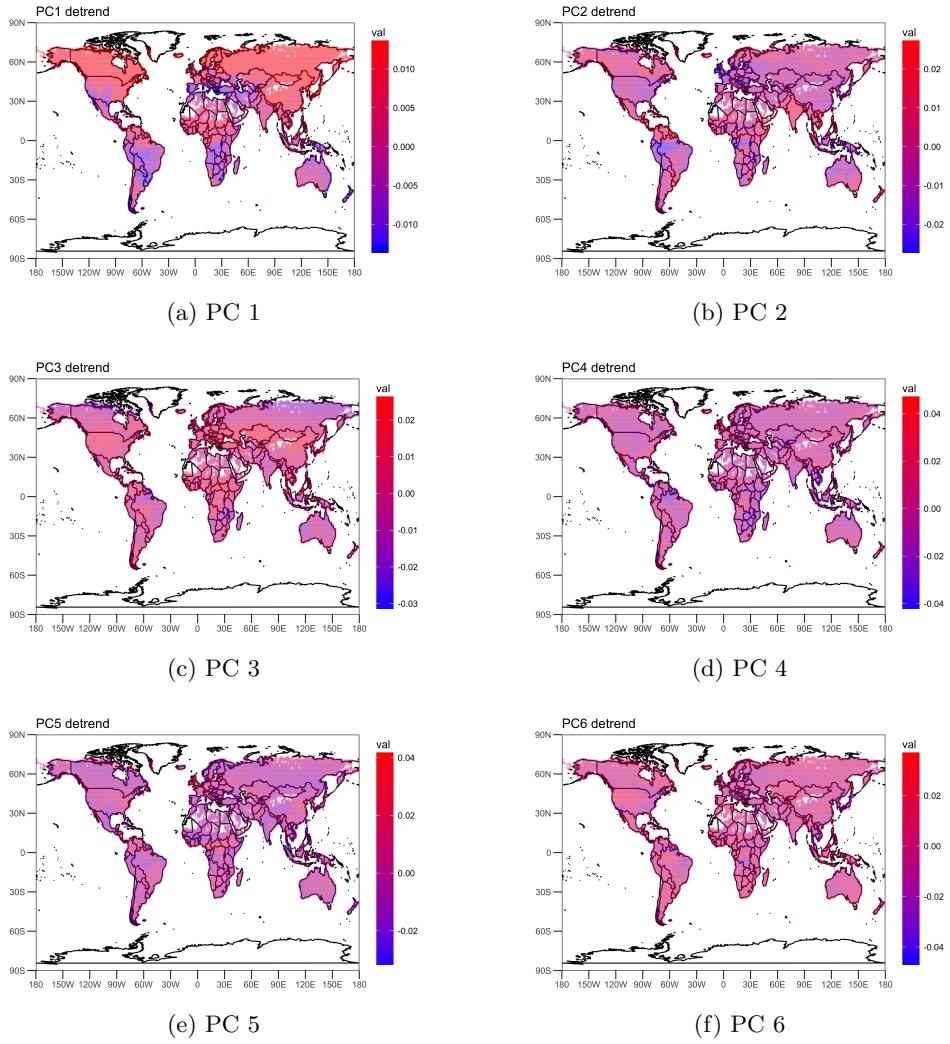


Figure 7: Spatial Patterns

B.2 Plots

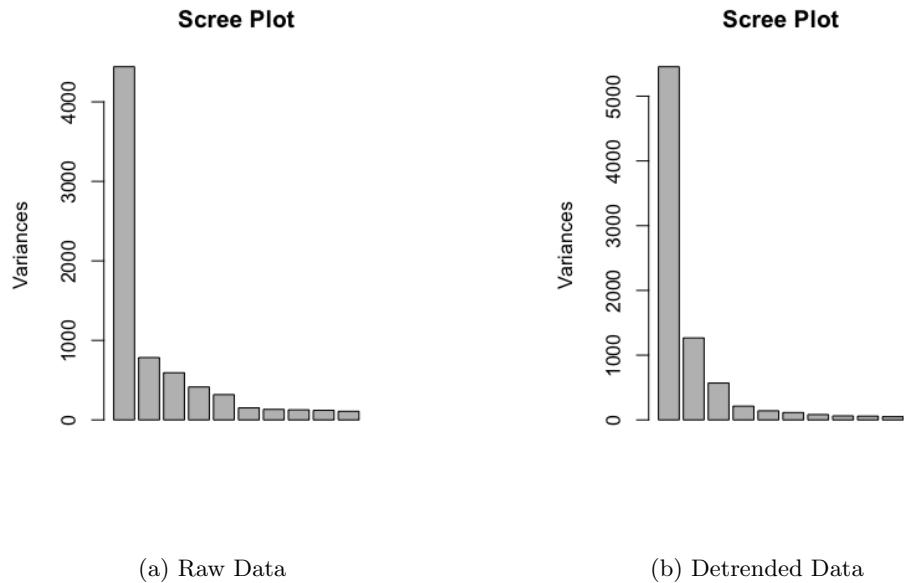


Figure 8: Scree Plots

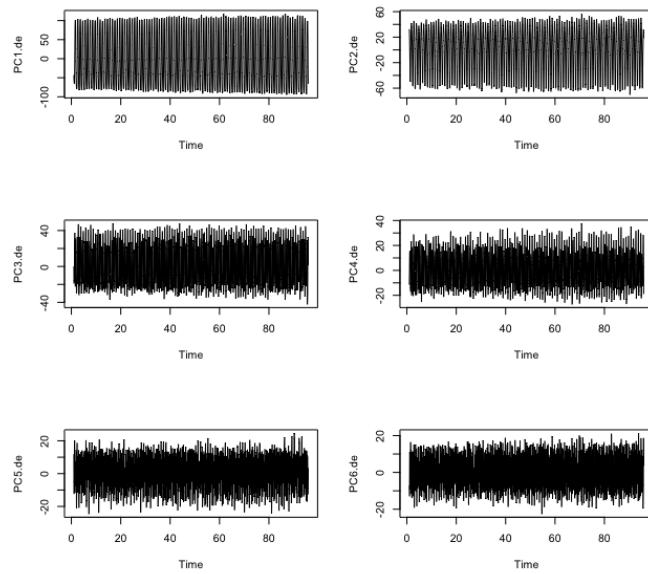


Figure 9: Time Series Plots

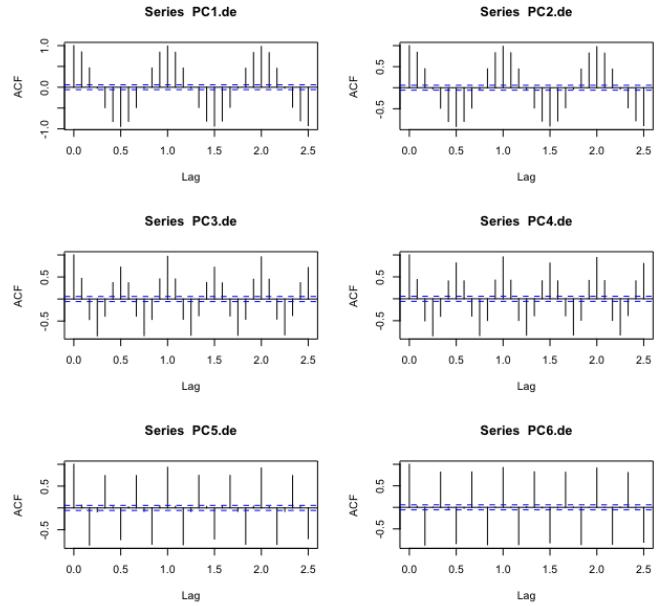


Figure 10: Auto-Covariance Function

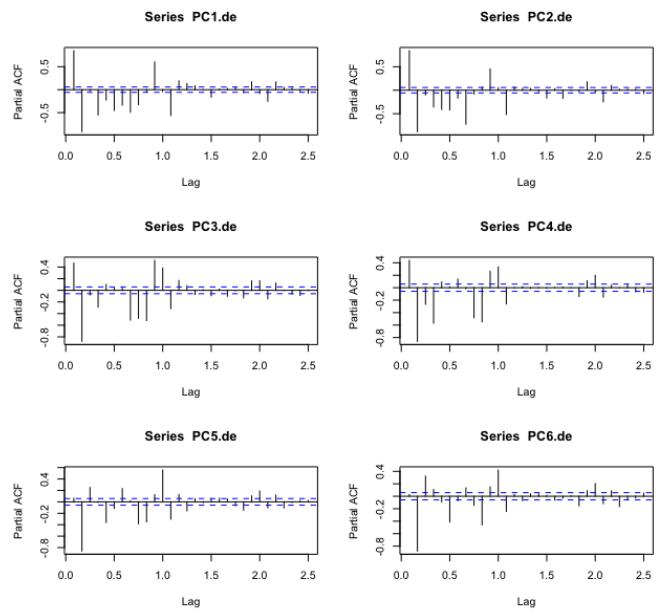


Figure 11: Partial ACF

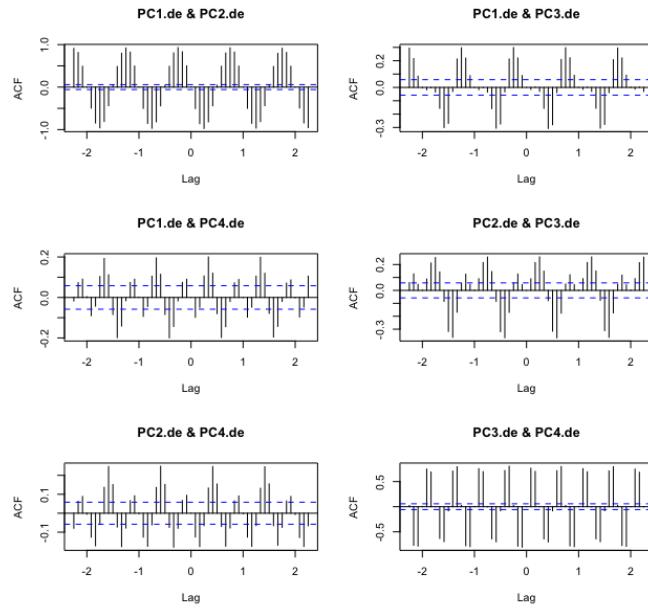


Figure 12: Cross-Covariance Function

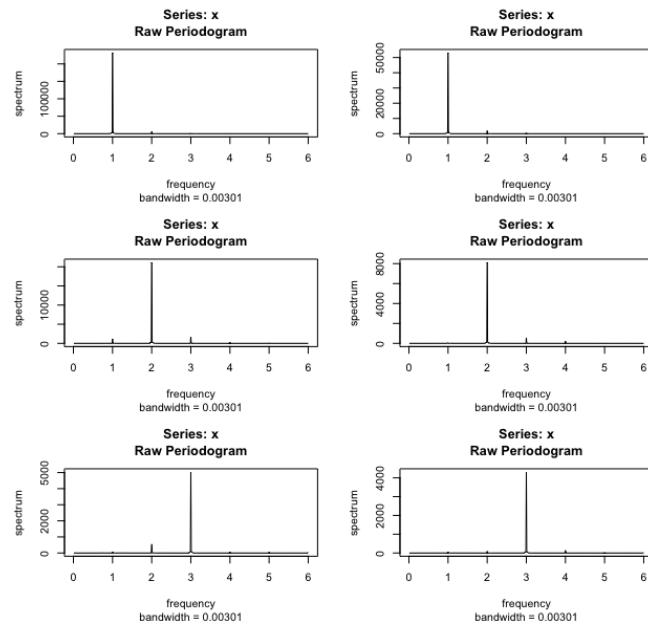


Figure 13: PCA Periodogram

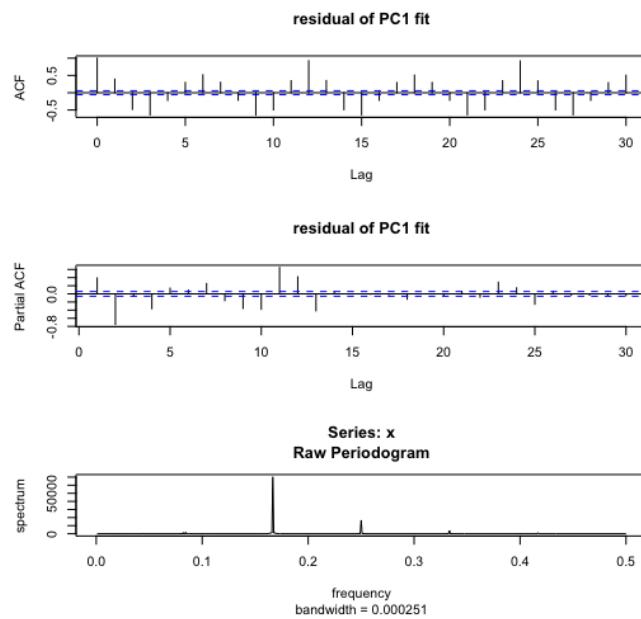


Figure 14: Analysis on Residuals PC1 Fit