Developing a Bayesian Procedure to Detect Breakpoints in Time Series: Progress Report

Kathryn Haglich, Sarah Neitzel, Amy Pitts
Mentor: Jeff Liebner

Lafayette College, Unity College, Marist College

Wednesday, June 27, 2018

NSF Grant #1560222

Recap

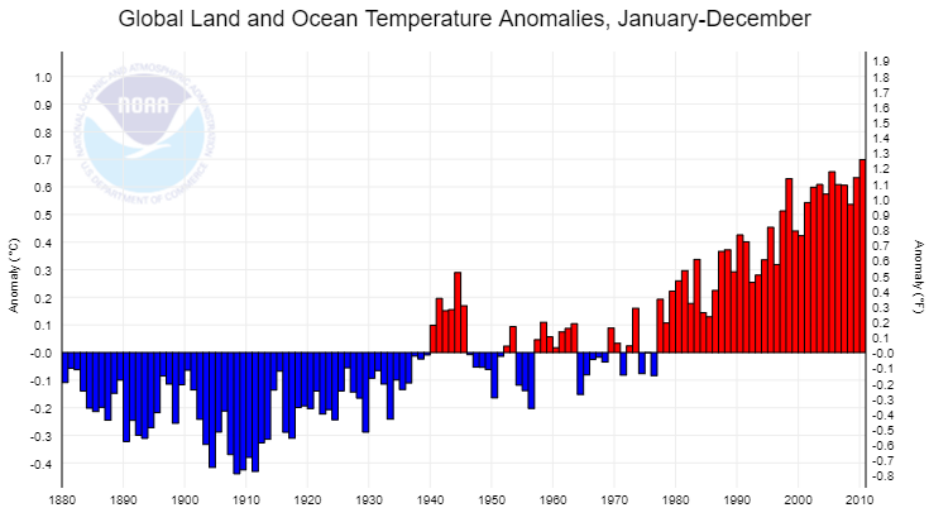


Figure 1: Climate Data

Goal:
to develop a better quantitative method for locating breakpoints in time series data

General Definitions

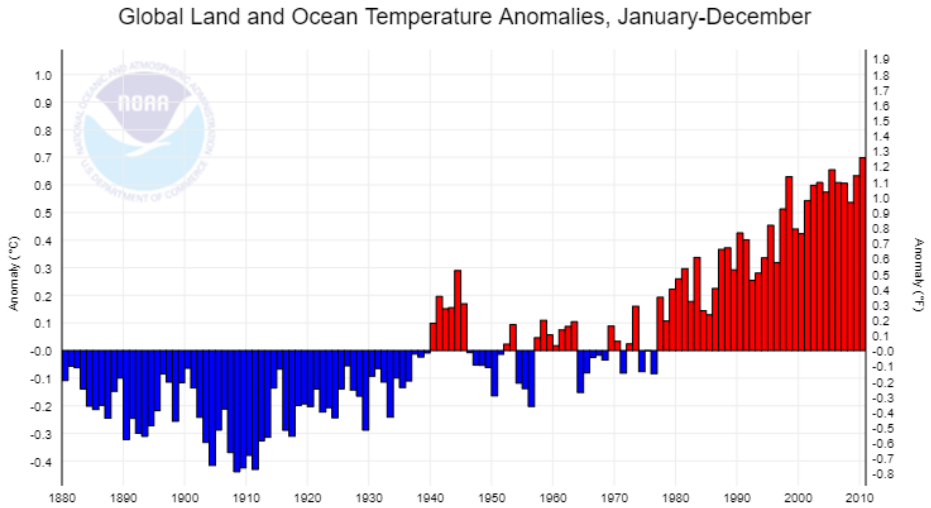


Figure 2: Climate Data

Breakpoints:
locations where there is a significant change in the data (i.e. changes in mean, variation, and or slope)

General Outline of Approach

Project Format:

1. Initial breakpoint problem (Bai-Perron Test)
2. Modification of Bayesian Adaptive Regression Splines (BARS) to propose a new set of breakpoints
3. Get $\theta = (K, \tau, \beta, \sigma)$ and find parameter distributions
   - $K =$ the number of breakpoints
   - $\tau =$ the location of breakpoints
   - $\beta =$ the regression coefficients
   - $\sigma =$ the standard deviations
4. Evaluate new proposed fit using Metropolis-Hastings
5. Repeat steps 2 through 4

Bai-Perron Test

Bai-Perron Test (1998):

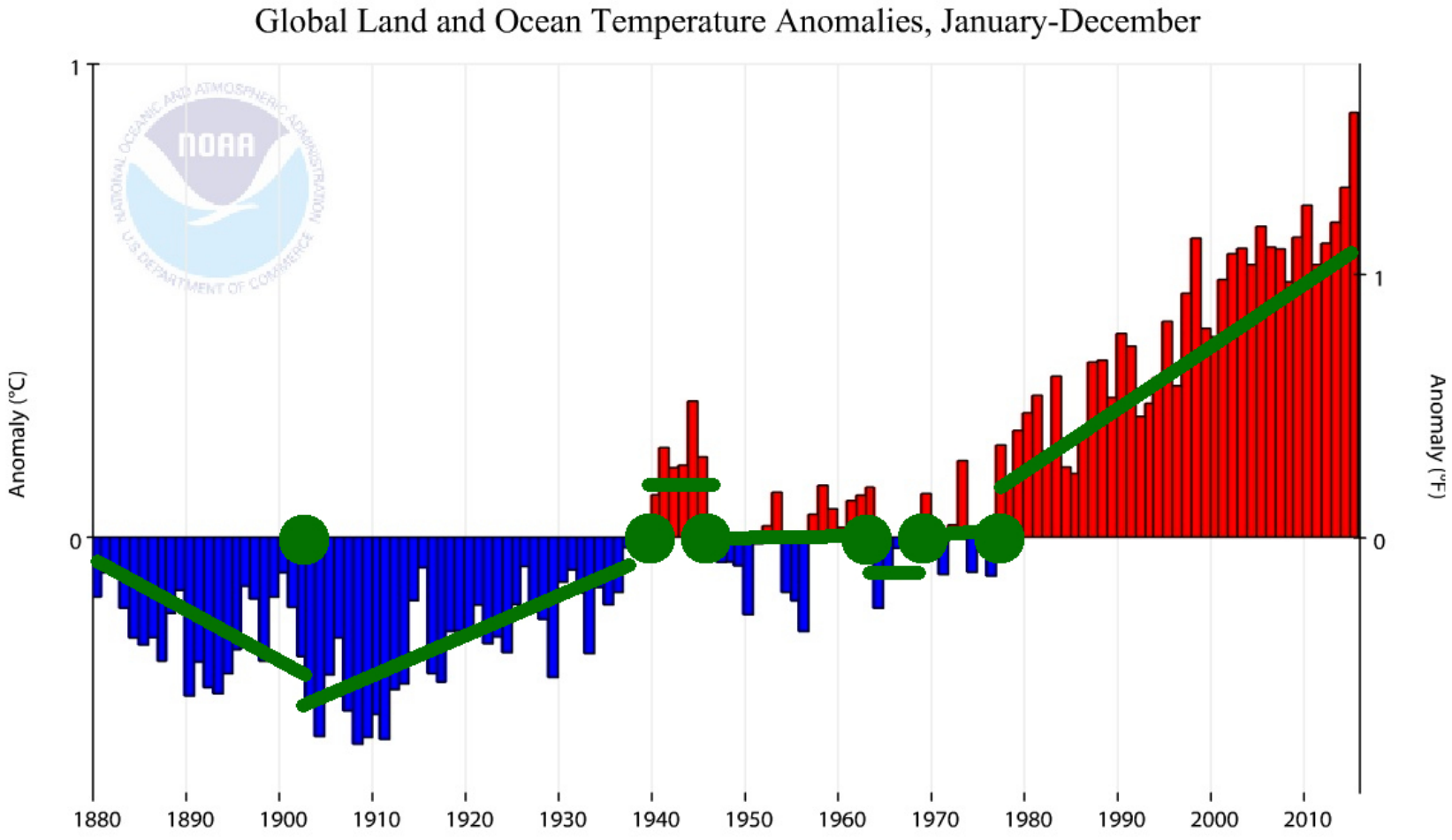
- Frequentist approach to identify significant changes in the behavior of a model describing a data set
- Searches through every single possible breakpoint location and determine the best model based off of the residual sum of squares (sum of squared distances from original data points to fitted model)
- The outcome is a single model conditioned on the number of breaks specified by the user

Bayesian Adaptive Regression Splines

Bayesian Adaptive Regression Splines (BARS):

- A stochastic process that proposes new set of breakpoints for the splines.
- This process involves adding, subtracting, and moving breakpoints to obtain a draw from the possible fits to the model.
- The optimal fit is obtained by averaging the different fits.

Bai-Perron Test on Global Temperatures



The Bayesian Approach

- Frequentist statistics is based on data being drawn from a distribution with fixed parameters
- Bayesian statistics is based on *parameters* being drawn from a distribution
- The distribution is determined from both the current data set as well as prior information

The Bayesian Approach

Bayes' Theorem:

$$g(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n|\theta)\pi(\theta)}{\int f(x_1,\ldots,x_n|\theta)\pi(\theta)d\theta} \propto f(x_1,\ldots,x_n|\theta)\pi(\theta)$$

$g(\theta|x_1,\ldots,x_n) \rightarrow$ posterior
$f(x_1,\ldots,x_n|\theta) \rightarrow$ likelihood
$\pi(\theta) \rightarrow$ prior

The Bayesian Approach

Bayes' Theorem:

$$g(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n|\theta)\pi(\theta)}{\int f(x_1,\ldots,x_n|\theta)\pi(\theta)d\theta} \propto f(x_1,\ldots,x_n|\theta)\pi(\theta)$$

$g(\theta|x_1,\ldots,x_n) \rightarrow$ posterior
$f(x_1,\ldots,x_n|\theta) \rightarrow$ likelihood
$\pi(\theta) \rightarrow$ prior

Our parameters:
$\theta = K, \tau_1, \ldots \tau_K, \beta, \sigma$

The Bayesian Approach

Our Distributions

β: Multivariate Normal

σ: Inverse Gamma

K and $\tau_1, ..., \tau_K$: something weird we'll figure out later... it's complicated...

Our Distributions

β: Multivariate Normal

σ: Inverse Gamma

K and $\tau_1, ..., \tau_K$: something weird we'll figure out later... it's complicated...

Distribution of everything together: a giant mess???

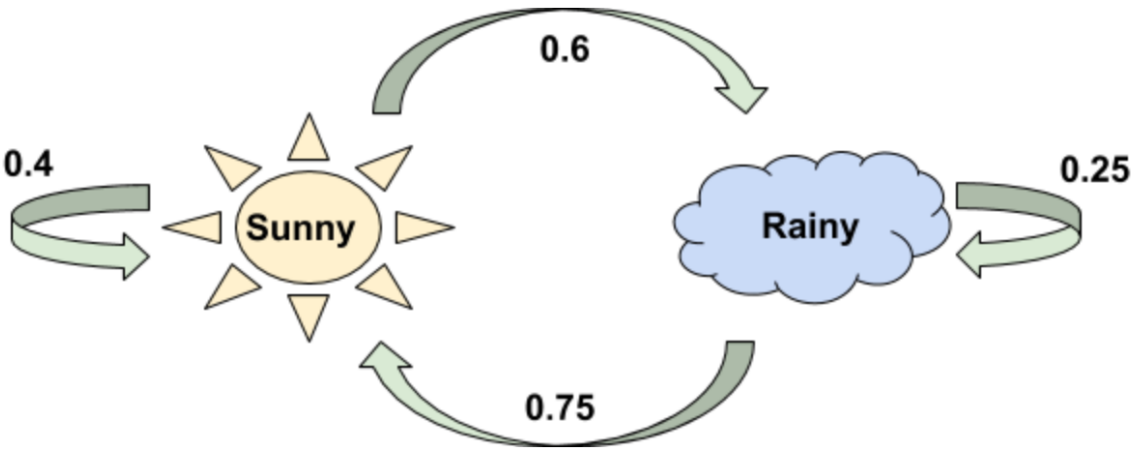
Metropolis Hastings

Metropolis Hastings - a mechanism consisting of a Markov Chain Monte Carlo that is used to sample a distribution when sampling is difficult

Our Overall Goal: is to sample the distribution $\theta = \{K, \tau_1, \ldots, \tau_k\}$ given $g(\theta|x_1, \ldots, x_n)$ where K, and τ are the number and location of breakpoints, given our data.

Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC)- a repeated stochastic process of state changes



Metropolis Hastings

Let q be a function, $q(\theta_{new}|\theta_{old})$ where $\theta = \{K, \tau_1, \ldots, \tau_k\}$, such that $q(\theta_{new}|\theta_{old}) = q(\theta_{old}|\theta_{new})$.

1. Draw $\theta_{new} \sim q(\cdot|\theta_{old})$, where $\theta_{old} = \theta^{(i-1)}$, the values of the parameters from the previous iteration of the MCMC;
2. Compute the ratio $r = g(\theta_{new})/g(\theta_{old})$ where g is the posterior distribution of θ
3. If $r \geq 1$, set $\theta^{(i)} = \theta_{new}$;

   If $r < 1$, set $\theta^{(i)} = \begin{cases} \theta_{new} \text{ with probability } r \\ \theta_{old} \text{ with probability } 1 - r \end{cases}$

General Approach Outline

First: Choose a type of step and then create a proposed breakpoint set from chosen step

Step Options

- Make (Random addition of a breakpoint)
- Murder (Random deletion of a breakpoint)
- Move (Combination of Murder and then Make)

Iteration Run Through: Initializing

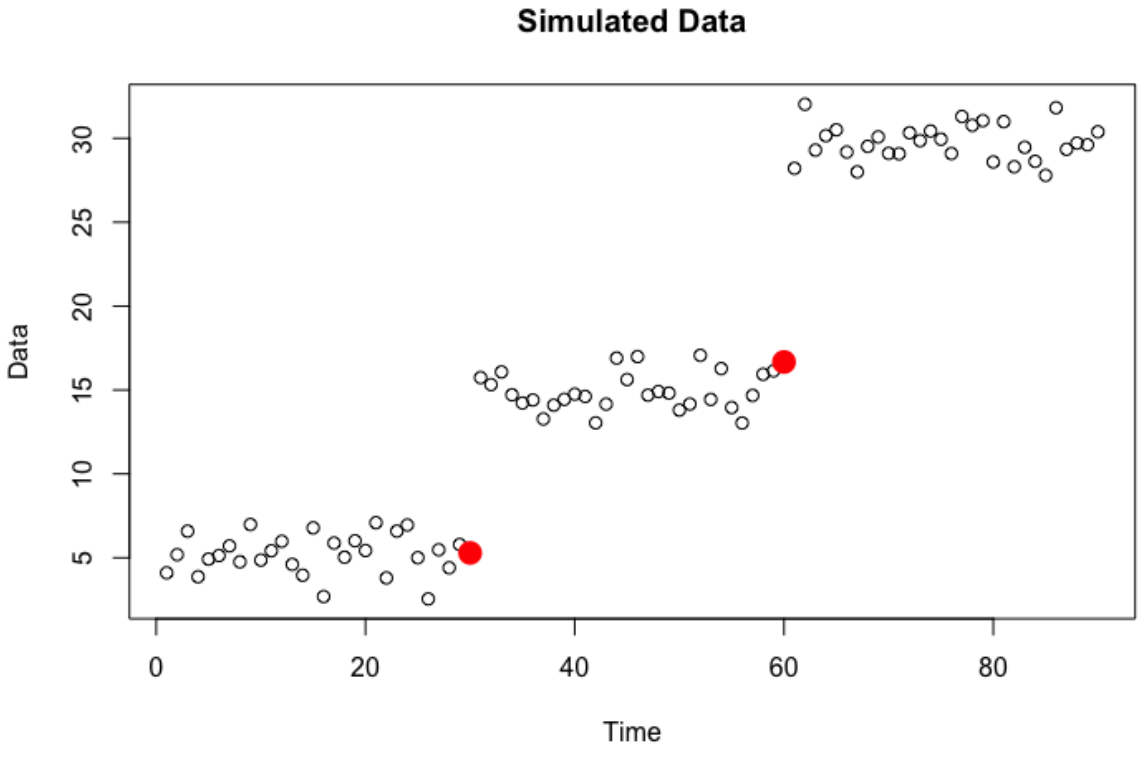


Figure 3: The red dots represent the breakpoints obtained by using the Bai-Perron test.

Iteration Run Through: Proposed Point

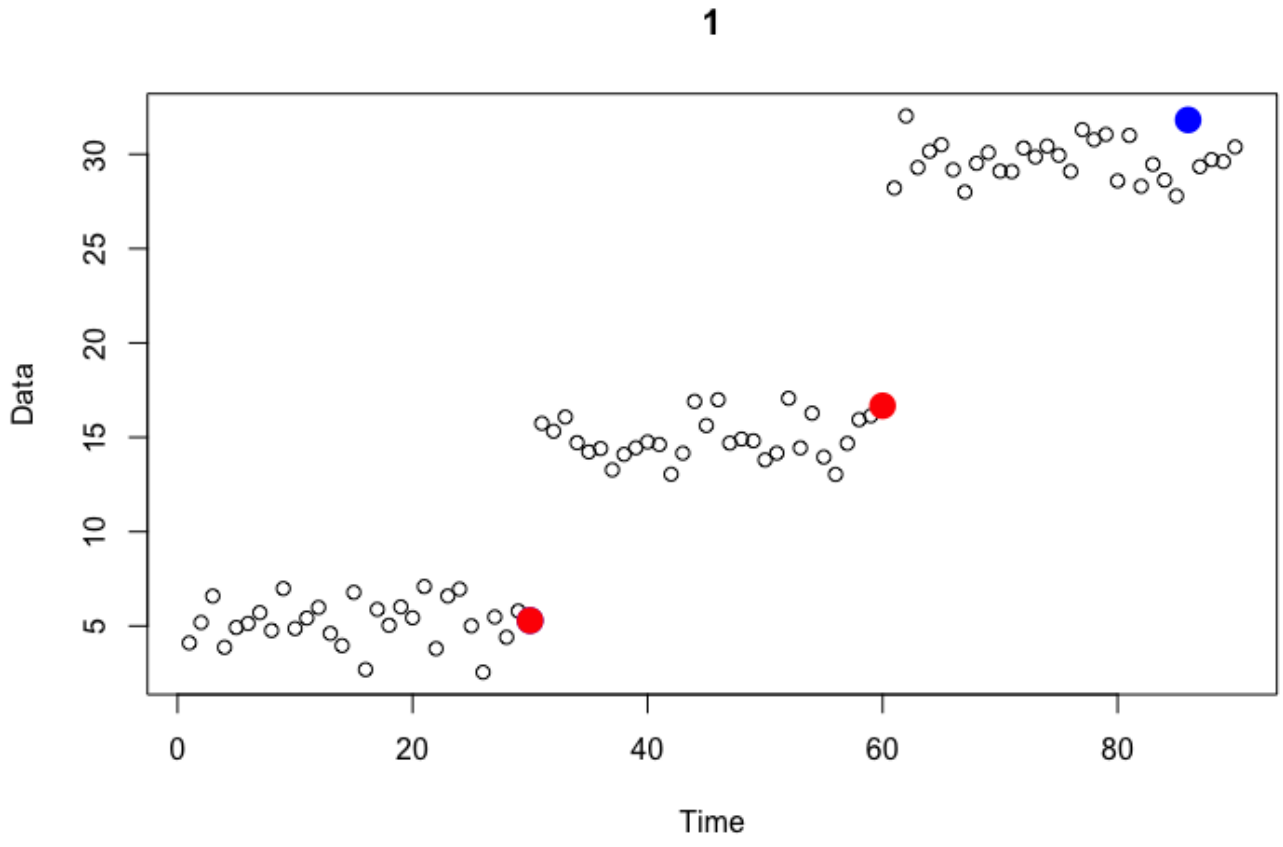


Figure 4: First Proposed Breakpoint. The red dots represent the original breakpoints and the blue dot represent the proposed breakpoint

Iteration Run Through: Metropolis Hastings

Second: Compare the new breakpoint set to the old using Metropolis Hastings where the ratio r is approximated by the difference in the values of the Bayesian Information Criterion (BIC).

Bayesian Information Criterion (BIC)

$BIC = log(n) \times (k+1)(p+1) - 2W$

- W being the log likelihood of breakpoint set
- p being the number of parameters
- k number of breakpoints
- n number of data points

Iteration Run Through: Linear Fit

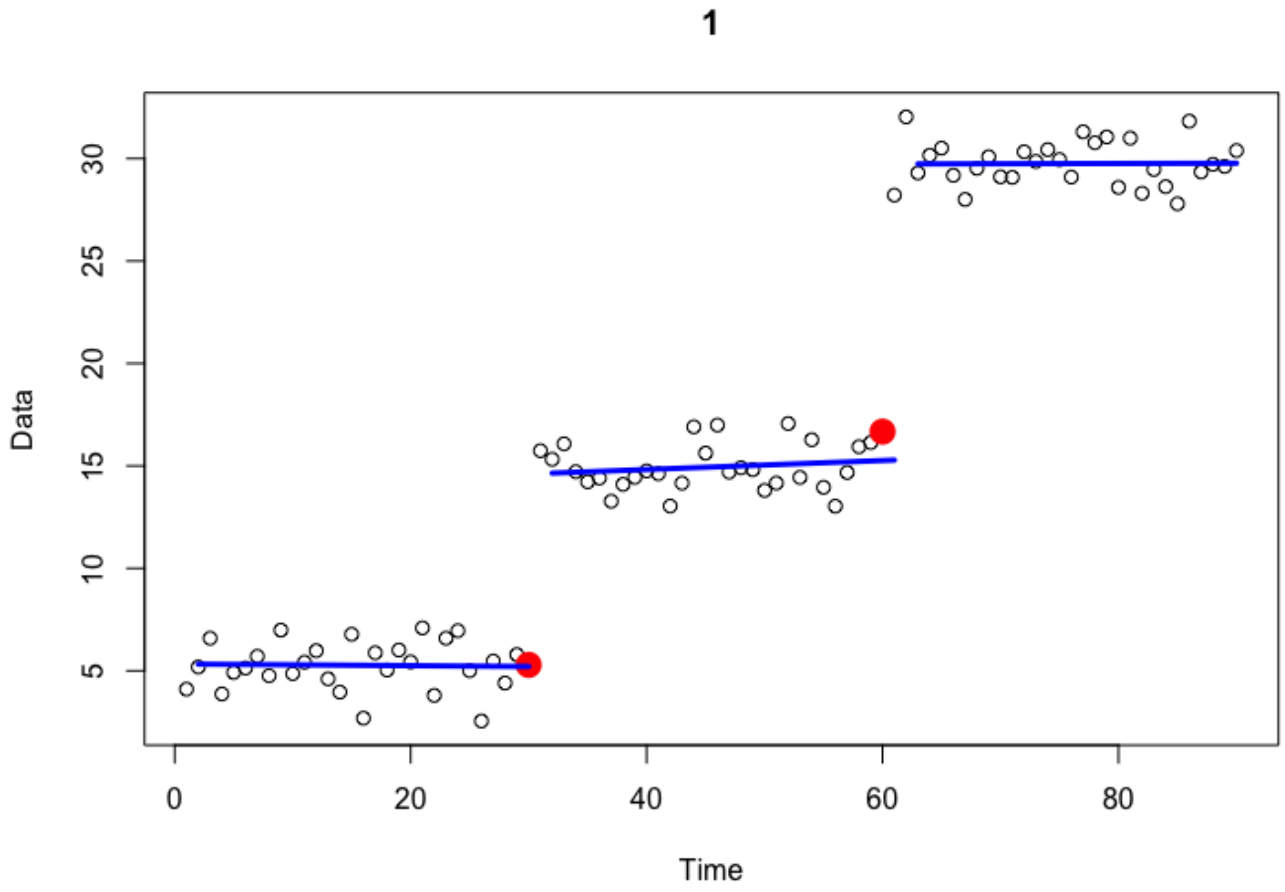
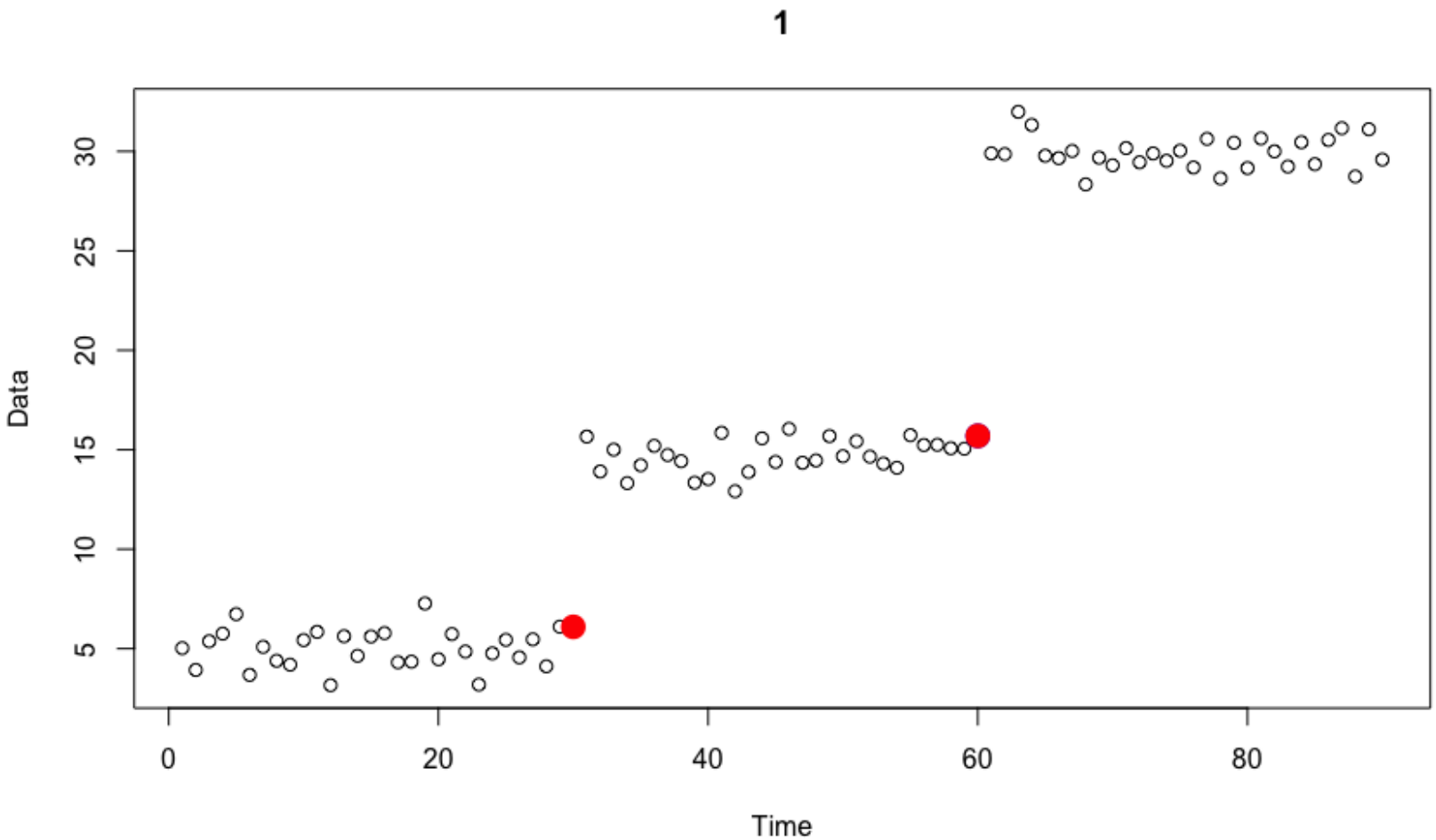
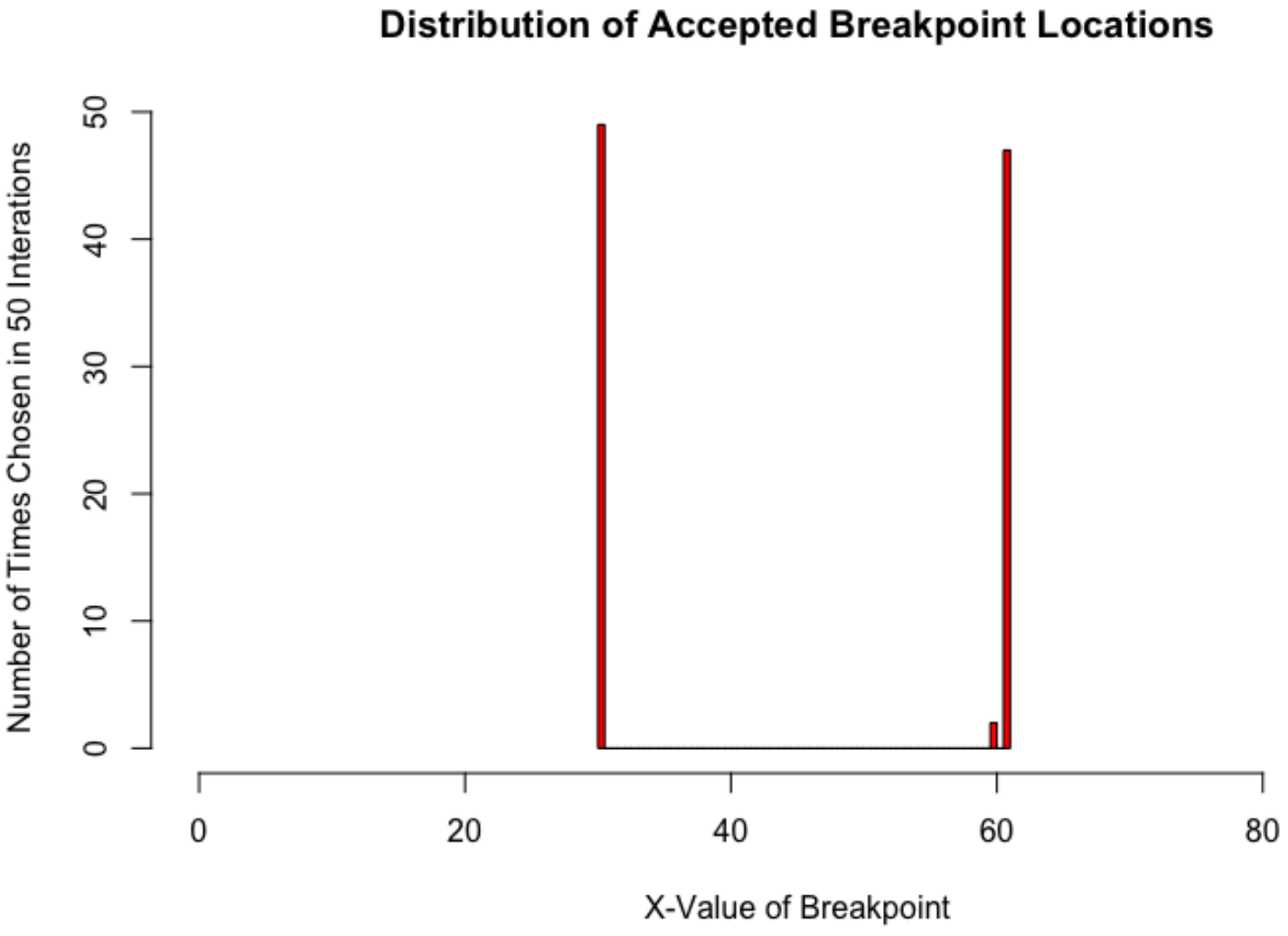


Figure 5: Linear regression lines between the kept breakpoint intervals.

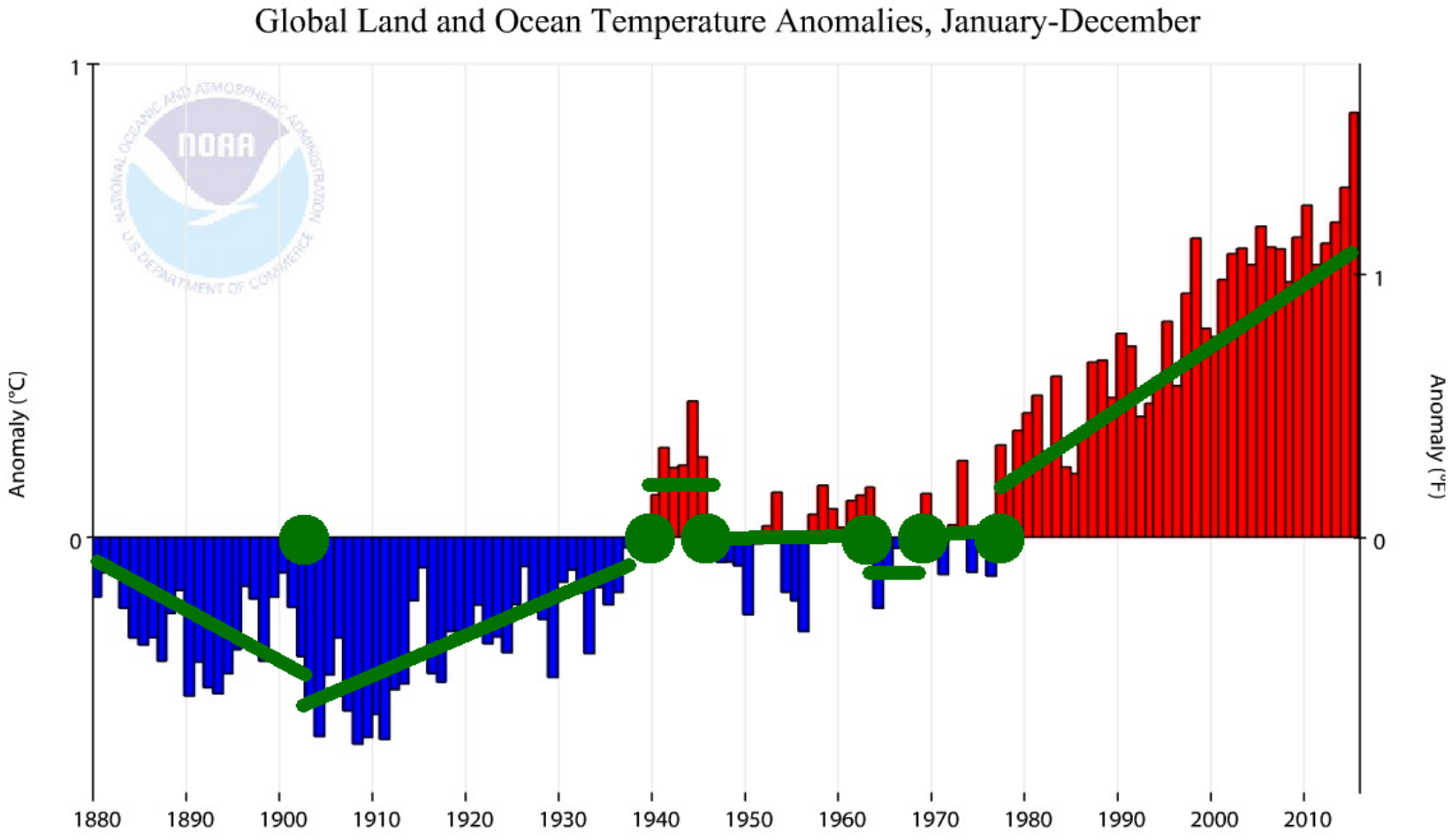
Iteration Run Through



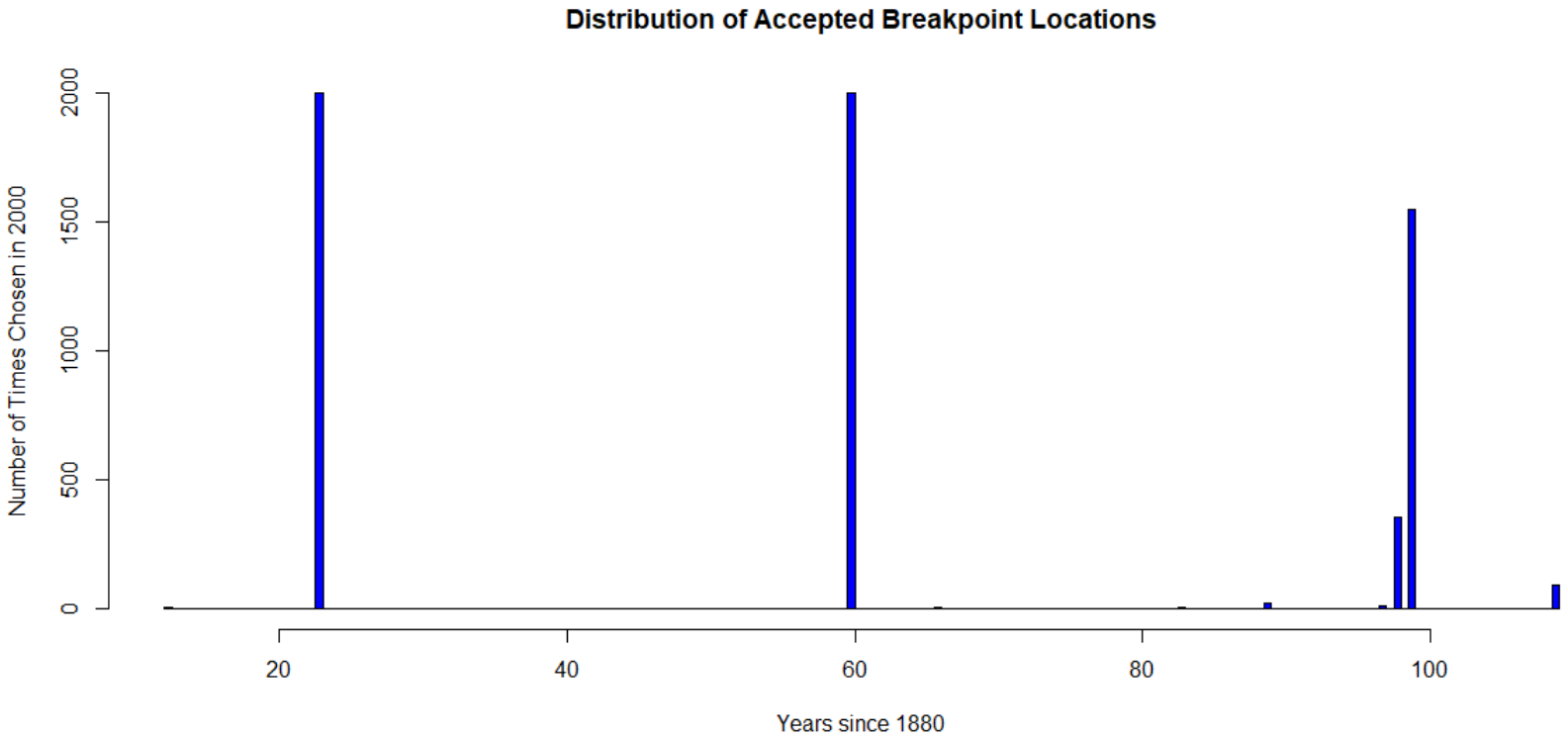
Iteration Run Through



Back to Climatology



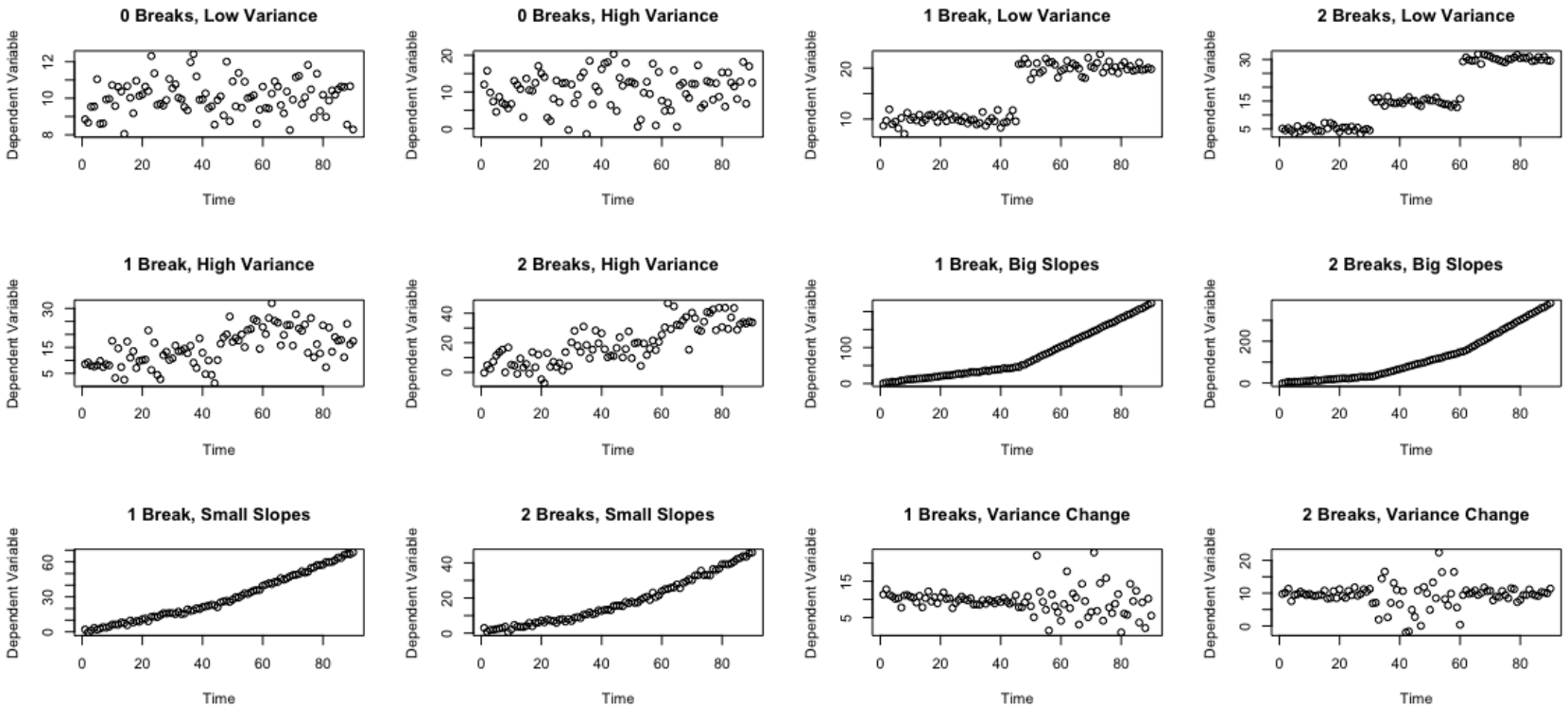
Back to Climatology



Moving Forward

- Simulations
  - To test different make, murder, move steps combinations which will effectively change q
  - Run many iterations on all different simulated data
- Prove BIC is a good estimator for the posterior
- Apply this process to Autoregressive (AR) models

Test Data


0 Breaks, Low Variance
0 Breaks, High Variance
1 Break, Low Variance
2 Breaks, Low Variance
1 Break, High Variance
2 Breaks, High Variance
1 Break, Big Slopes
2 Breaks, Big Slopes
1 Break, Small Slopes
2 Breaks, Small Slopes
1 Breaks, Variance Change
2 Breaks, Variance Change
Dependent Variable
Time