The Temperature Proxy Controversy

Lawrence F. Gray

School of Mathematics

February 8, 2012

Overview

1 Introduction

2 A tiny bit of statistics

3 Controversy

4 Some details

A provocative statement

> *"Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: lies, damned lies, and statistics'."*

Chapters from my Autobiography, Mark Twain, 1906

Another provocative statement

> *" . . . it is likely that the 1990s have been the warmest decade and 1998 the warmest year of the millenium."*

Climate Change 2001: The Scientific Basis, Intergovernmental Panel on Climate Change. (IPCC2001)

Another provocative statement

> *" . . . it is likely that the 1990s have been the warmest decade and 1998 the warmest year of the millenium."*

Climate Change 2001: The Scientific Basis, Intergovernmental Panel on Climate Change. (IPCC2001)

- What is the scientific basis for this statement?

Another provocative statement

> *" . . . it is likely that the 1990s have been the warmest decade and 1998 the warmest year of the millenium."*

Climate Change 2001: The Scientific Basis, Intergovernmental Panel on Climate Change. (IPCC2001)

- What is the scientific basis for this statement?
- How do we know there wasn't a similar temperature spike in the 1190s or the 1530s?

The famous "Hockey Stick"

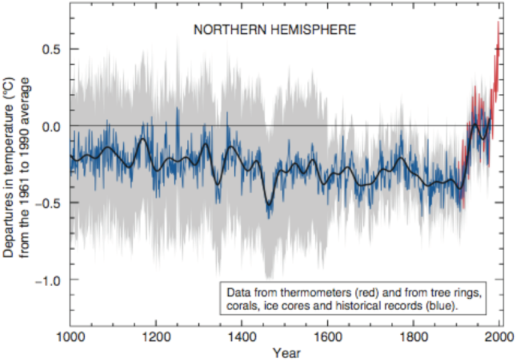


FIG 1. *Multiproxy reconstruction of Northern Hemisphere surface temperature variations over the past millennium (blue), along with 50-year average (black), a measure of the statistical uncertainty associated with the reconstruction (gray), and instrumental surface temperature data for the last 150 years (red), based on the work by* Mann et al. (*1999*). *This figure has sometimes been referred to as the hockey stick. Source:* IPCC (*2001*).

Some quick remarks

- The blue curve is a *weighted average* of temperature histories, conditioned on the data;
- The gray region only indicates *pointwise* confidence intervals;
- Known temperatures from the past 150 years ("the instrumental period") were used as *training data* for the statistical model;

Some quick remarks

- The blue curve is a *weighted average* of temperature histories, conditioned on the data;
- The gray region only indicates *pointwise* confidence intervals;
- Known temperatures from the past 150 years ("the instrumental period") were used as *training data* for the statistical model;

The first two features tend to *downplay* the likelihood that temperature spikes might have occurred prior to 1850.

Some quick remarks

- The blue curve is a *weighted average* of temperature histories, conditioned on the data;
- The gray region only indicates *pointwise* confidence intervals;
- Known temperatures from the past 150 years ("the instrumental period") were used as *training data* for the statistical model;

The first two features tend to *downplay* the likelihood that temperature spikes might have occurred prior to 1850. The third feature makes the model seem better than it is at finding such spikes.

Overview

1. Introduction
2. A tiny bit of statistics
3. Controversy
4. Some details

The general method

- Data from tree rings, ice cores, coral, lake sediments, etc., are *proxies* for the actual local temperatures. The data is gathered from many places and many time periods. The number of different proxies is $p \approx 1200$, but only about 100 are available for the entire millennium.

The general method

- Data from tree rings, ice cores, coral, lake sediments, etc., are *proxies* for the actual local temperatures. The data is gathered from many places and many time periods. The number of different proxies is $p \approx 1200$, but only about 100 are available for the entire millennium.
- Scientific theories lead to statistical models for the relationship between proxy measurements and temperatures (fancy linear regression).

The general method

- Data from tree rings, ice cores, coral, lake sediments, etc., are *proxies* for the actual local temperatures. The data is gathered from many places and many time periods. The number of different proxies is $p \approx 1200$, but only about 100 are available for the entire millennium.
- Scientific theories lead to statistical models for the relationship between proxy measurements and temperatures (fancy linear regression).
- Model parameters are tuned using known temperatures from the past $n = 150$ years

It's an art

- Tuned model "backcasts" temps to time period 1000-1850.

It's an art

- Tuned model "backcasts" temps to time period 1000-1850.
- This is not a standard statistical problem, and there are many competing approaches, most of them attempting to deal with the fact that $p \gg n$.

It's an art

- Tuned model "backcasts" temps to time period 1000-1850.
- This is not a standard statistical problem, and there are many competing approaches, most of them attempting to deal with the fact that $p \gg n$.
- Two approaches are: (i) Select a small number of "principal components"; (ii) The "Lasso", which is a regression that penalizes the inclusion of more covariates.

Selection, validation, comparison

Selection, validation, comparison

- A common way to validate a method or model, or to compare two of them, is *block holdout RMSE*: block(s) of the training data are withheld during the tuning stage, and then the tuned model is used to predict the held out temperatures and the RMSE is calculated.

Selection, validation, comparison

- A common way to validate a method or model, or to compare two of them, is *block holdout RMSE*: block(s) of the training data are withheld during the tuning stage, and then the tuned model is used to predict the held out temperatures and the RMSE is calculated.
- In selecting and validating the model that produced the Hockey Stick, two blocks were held out: the first 50 years and the last 50 years, so the tuning was based on the middle 50 years.

The three blocks

This was done to favor models that are good at extrapolation.

The three blocks

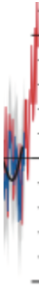
This was done to favor models that are good at extrapolation. But let's think about this for a minute . . .

The three blocks

This was done to favor models that are good at extrapolation. But let's think about this for a minute . . .

1850-1900

1900-1950



1950-2000

Overview

1 Introduction

2 A tiny bit of statistics

3 Controversy

4 Some details

Is there a Hockey Stick Bias?

- In Mann et al [1998], a slightly non-standard method was used involving a single principal component, and it was later shown that this method has a "Hockey Stick Bias".

Is there a Hockey Stick Bias?

- In Mann et al [1998], a slightly non-standard method was used involving a single principal component, and it was later shown that this method has a “Hockey Stick Bias”.
- In a subsequent paper, Mann et al [2004], the block holdout method was used to decide on the number of principal components, and the Hockey Stick (magically) reappeared.

Is there a Hockey Stick Bias?

- In Mann et al [1998], a slightly non-standard method was used involving a single principal component, and it was later shown that this method has a "Hockey Stick Bias".
- In a subsequent paper, Mann et al [2004], the block holdout method was used to decide on the number of principal components, and the Hockey Stick (magically) reappeared.
- "Mainstream statisticians" were typically not involved (beyond making a few criticisms).

Enter: McShane and Wyner [2010]

- The most famous line in the paper is probably “We find that the proxies do not predict temperature significantly better than random series generated independently of temperature.”

Enter: McShane and Wyner [2010]

- The most famous line in the paper is probably "We find that the proxies do not predict temperature significantly better than random series generated independently of temperature."
- Doubt is cast on the ability of the proxies to forecast "high levels of and sharp run-up in temperature . . . if in fact they occurred several hundred years ago."

Enter: McShane and Wyner [2010]

- The most famous line in the paper is probably “We find that the proxies do not predict temperature significantly better than random series generated independently of temperature.”
- Doubt is cast on the ability of the proxies to forecast “high levels of and sharp run-up in temperature . . . if in fact they occurred several hundred years ago.”
- Other equally valid models produce “extremely different historical backcasts”.

McShane and Wyner [2010] (continued)

- They create their own Bayesian model. It produces a reconstruction of the past that is similar to the Hockey Stick, but it has a much wider error band, which reflects "the weak signal and large uncertainty . . . in this setting."

McShane and Wyner [2010] (continued)

- They create their own Bayesian model. It produces a reconstruction of the past that is similar to the Hockey Stick, but it has a much wider error band, which reflects "the weak signal and large uncertainty . . . in this setting."
- The Bayesian model produces a "36% posterior probability that 1998 was the warmest in the past thousand years."

McShane and Wyner [2010] (continued)

- They create their own Bayesian model. It produces a reconstruction of the past that is similar to the Hockey Stick, but it has a much wider error band, which reflects "the weak signal and large uncertainty . . . in this setting."
- The Bayesian model produces a "36% posterior probability that 1998 was the warmest in the past thousand years."
- An entire issue of the *Annals of Applied Statistics* was devoted to this paper, 13 responses, and a rejoinder.

Overview

1 Introduction

2 A tiny bit of statistics

3 Controversy

4 Some details

Assumptions ("Best Case Scenario")

- The data set used by Mann et al is reliable.

Assumptions ("Best Case Scenario")

- The data set used by Mann et al is reliable.
- The basic modeling assumptions used by Mann et al and other climate scientists are acceptable. These are that the relationships between the proxies and the temperatures are approximately linear and stationary.

Assumptions ("Best Case Scenario")

- The data set used by Mann et al is reliable.
- The basic modeling assumptions used by Mann et al and other climate scientists are acceptable. These are that the relationships between the proxies and the temperatures are approximately linear and stationary.
- Block holdout RMSE can be used to evaluate models, but not necessarily just the first and last thirds of 1850-2000.

The Mainstream Statisticians' Touch

- They introduce several interesting "null models" to provide benchmarks for the predictive strength of the proxies

The Mainstream Statisticians' Touch

- They introduce several interesting "null models" to provide benchmarks for the predictive strength of the proxies
- To evaluate models, McShane and Wyner calculate the holdout RMSE for all 120 possible 30-year blocks of data from 1850-2000.

The Mainstream Statisticians' Touch

- They introduce several interesting "null models" to provide benchmarks for the predictive strength of the proxies
- To evaluate models, McShane and Wyner calculate the holdout RMSE for all 120 possible 30-year blocks of data from 1850-2000.
- For regression, they use the so-called "Lasso" method, considered to be "a reasonable procedure that has proven powerful, fast, and popular, and performs comparably well in a $p \gg n$ context."

The null models

The null models in this paper are of two types:

- Those that are based strictly on the temperatures from 1850-2000, the simplest one being to take the mean temperature from that time period;

The null models

The null models in this paper are of two types:

- Those that are based strictly on the temperatures from 1850-2000, the simplest one being to take the mean temperature from that time period;
- Those that replace the proxy data by "pseudo-proxies". These are randomly generated time series, independent of the historical record. One example is to replace each proxy time series with a Brownian motion.

The most interesting null model

One pseudo-proxy time series is called the Empirical $AR1$ Model.

The most interesting null model

One pseudo-proxy time series is called the Empirical $AR1$ Model. An $AR1$ time series is of the form $X_t = \phi X_{t-1} + \epsilon_t$, where $0 \leq \phi \leq 1$ is a parameter and the ϵ_t are independent standard Gaussian random variables.

The most interesting null model

One pseudo-proxy time series is called the Empirical $AR1$ Model. An $AR1$ time series is of the form $X_t = \phi X_{t-1} + \epsilon_t$, where $0 \leq \phi \leq 1$ is a parameter and the ϵ_t are independent standard Gaussian random variables. For each proxy, a parameter $\phi_i, i = 1, \ldots, 1200$, is "estimated" for each proxy series.

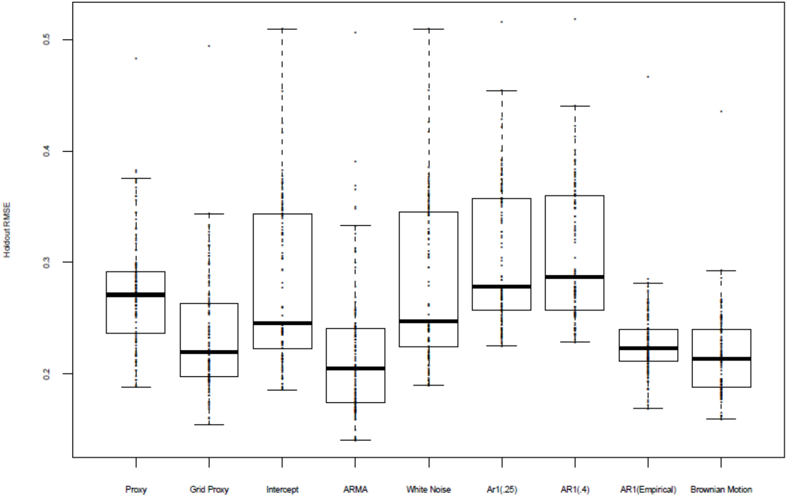
The most interesting null model

One pseudo-proxy time series is called the Empirical $AR1$ Model. An $AR1$ time series is of the form $X_t = \phi X_{t-1} + \epsilon_t$, where $0 \leq \phi \leq 1$ is a parameter and the ϵ_t are independent standard Gaussian random variables. For each proxy, a parameter $\phi_i, i = 1, \ldots, 1200$, is "estimated" for each proxy series. Then the 1200 pseudo-proxy time series are generated independently, using these values of ϕ_i.

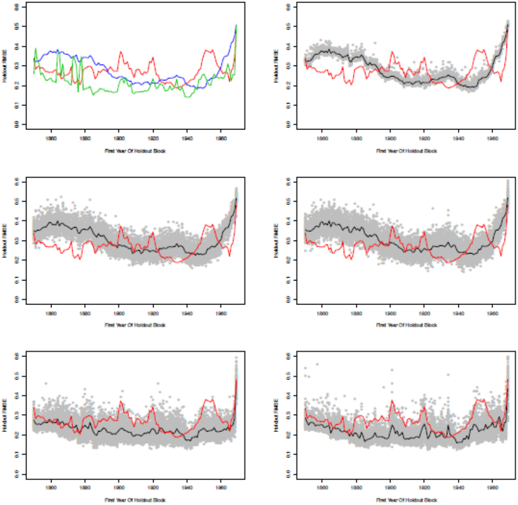
The most interesting null model

One pseudo-proxy time series is called the Empirical $AR1$ Model. An $AR1$ time series is of the form $X_t = \phi X_{t-1} + \epsilon_t$, where $0 \leq \phi \leq 1$ is a parameter and the ϵ_t are independent standard Gaussian random variables. For each proxy, a parameter $\phi_i, i = 1, \ldots, 1200$, is "estimated" for each proxy series. Then the 1200 pseudo-proxy time series are generated independently, using these values of ϕ_i. Also considered are the cases $\phi_i \equiv 1$ (Brownian motion), $\phi_i \equiv 0$ (White noise), and $\phi_i \equiv .25, .4$.

The results


Holdout RMSE
0.2
0.3
0.4
0.5
Proxy
Grid Proxy
Intercept
ARMA
White Noise
Ar1(.25)
AR1(.4)
AR1(Empirical)
Brownian Motion

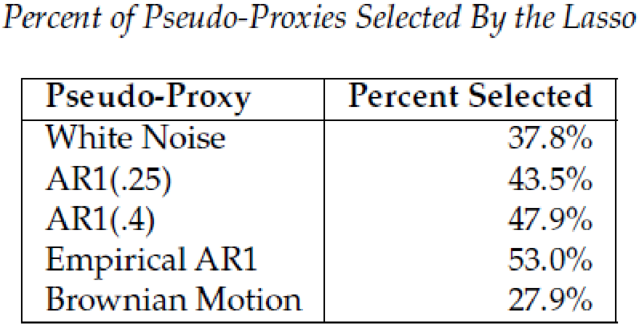
The results


0.0
0.1
0.2
0.3
0.4
0.5
0.6
Holdout RMSE
1860
1880
1900
1920
1940
1960
First Year Of Holdout Block

A different way of using pseudo-proxies

An alternative method for determining the predictive value of the proxies is to create a data set consisting of the 1200 proxy time series and 1200 independently generated pseudo-proxy time series. Then perform a regression on the expanded data set, using the Lasso. In general, the Lasso selects a relatively small subset of the covariates, based on their apparent predictive value during the training phase. So it is interesting to see whether the Lasso selects true proxies more often than pseudo-proxies.

The results



Percent of Pseudo-Proxies Selected By the Lasso

Pseudo-Proxy	Percent Selected
White Noise	37.8%
AR1(.25)	43.5%
AR1(.4)	47.9%
Empirical AR1	53.0%
Brownian Motion	27.9%

Comparing several proxy models

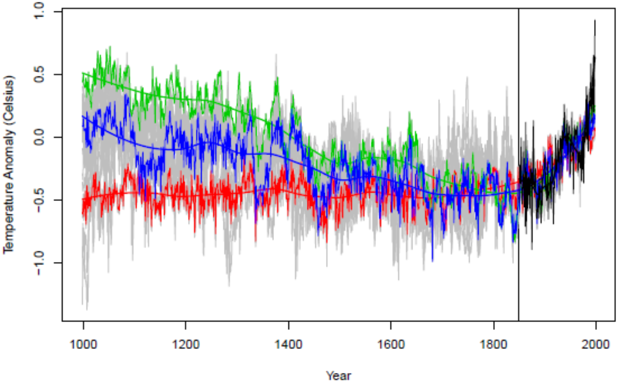
McShane and Wyner also looked at the backcasts produced by several different regression models, based only on the roughly 100 proxies for which data is available over the entire millennium.

Comparing several proxy models

McShane and Wyner also looked at the backcasts produced by several different regression models, based only on the roughly 100 proxies for which data is available over the entire millennium. There were 27 models in all, including the Lasso, “stepwise regression”, applied to the full proxy series and to principal components of the proxies, ordinary regression applied to principal components, and two-stage models involving local temperatures.

A variety of backcasts

Three backcasts, using models with very similar (relatively good) RMSE:



PC1 (red), PC10 (green), Two-stage (blue)

The Bayesian Model

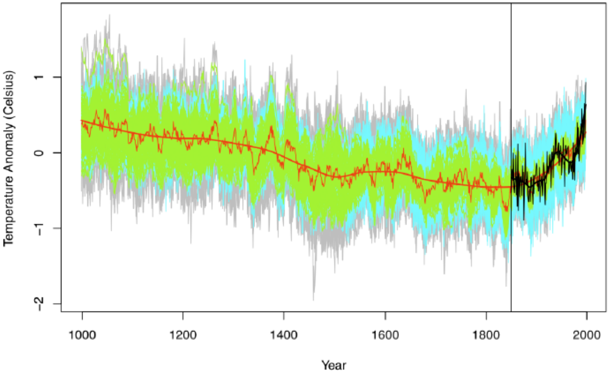


FIG 16. *Backcast from Bayesian Model of Section 5. CRU Northern Hemisphere annual mean land temperature is given by the thin black line and a smoothed version is given by the thick black line. The forecast is given by the thin red line and a smoothed version is given by the thick red line. The model is fit on 1850-1998 AD and backcasts 998-1849 AD. The cyan region indicates uncertainty due to ϵ_t, the green region indicates uncertainty due to $\vec{\beta}$, and the gray region indicates total uncertainty.*