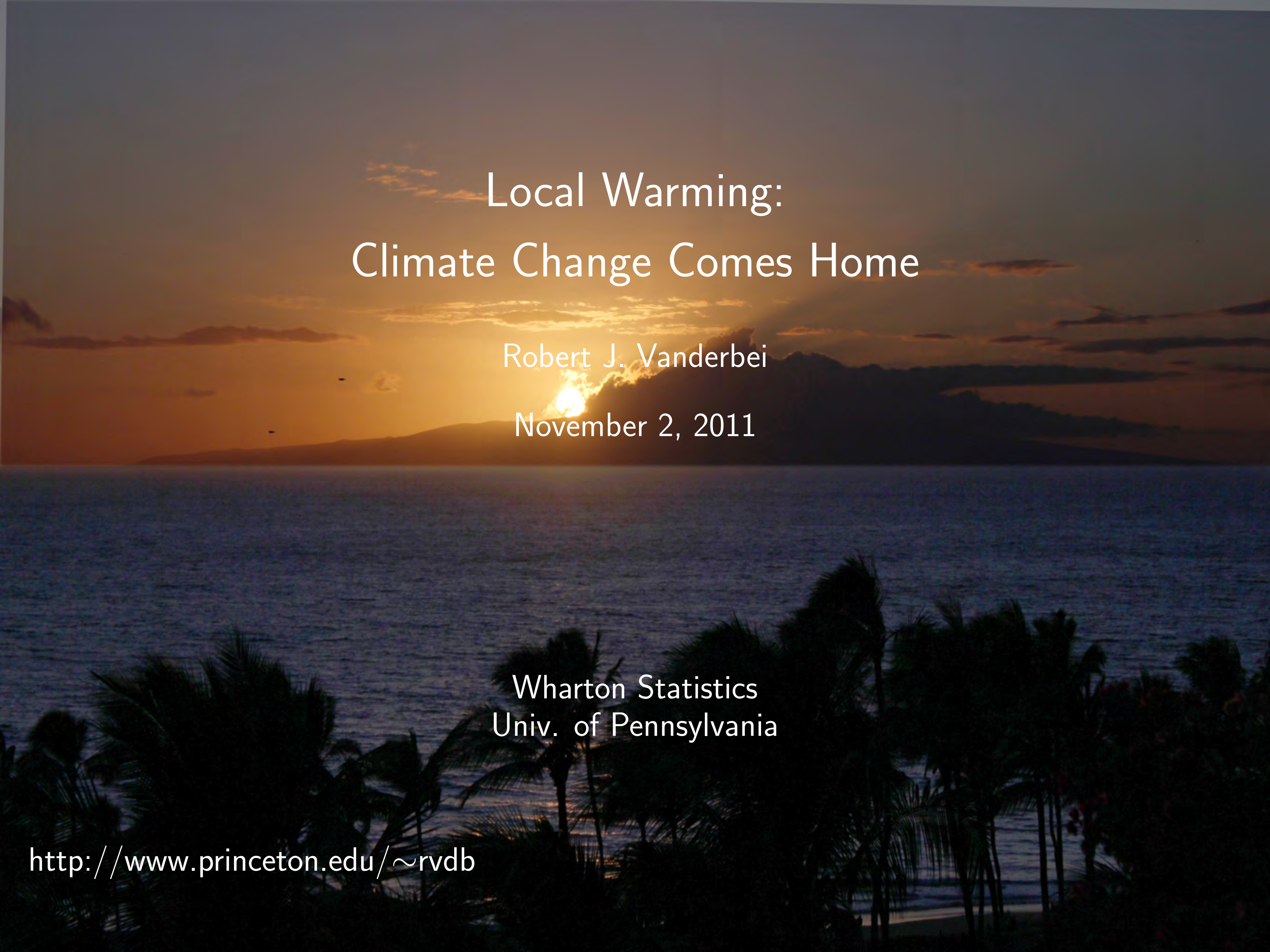


Local Warming:
Climate Change Comes Home

Robert J. Vanderbei

November 2, 2011

Wharton Statistics
Univ. of Pennsylvania

http://www.princeton.edu/∼rvdb

Introduction

There has been so much talk about global warming.

Is it real?

Is it anthropogenic?

Global warming starts at home.

So, let's address the question of *local warming*.

Has it been getting warmer in NJ?

The Data

Source: *National Oceanic and Atmospheric Administration (NOAA)*

Data format and downloading instructions:

ftp://ftp.ncdc.noaa.gov/pub/data/gsod/readme.txt

List of ∼9000 weather stations posted here:

ftp://ftp.ncdc.noaa.gov/pub/data/gsod/ish-history.txt

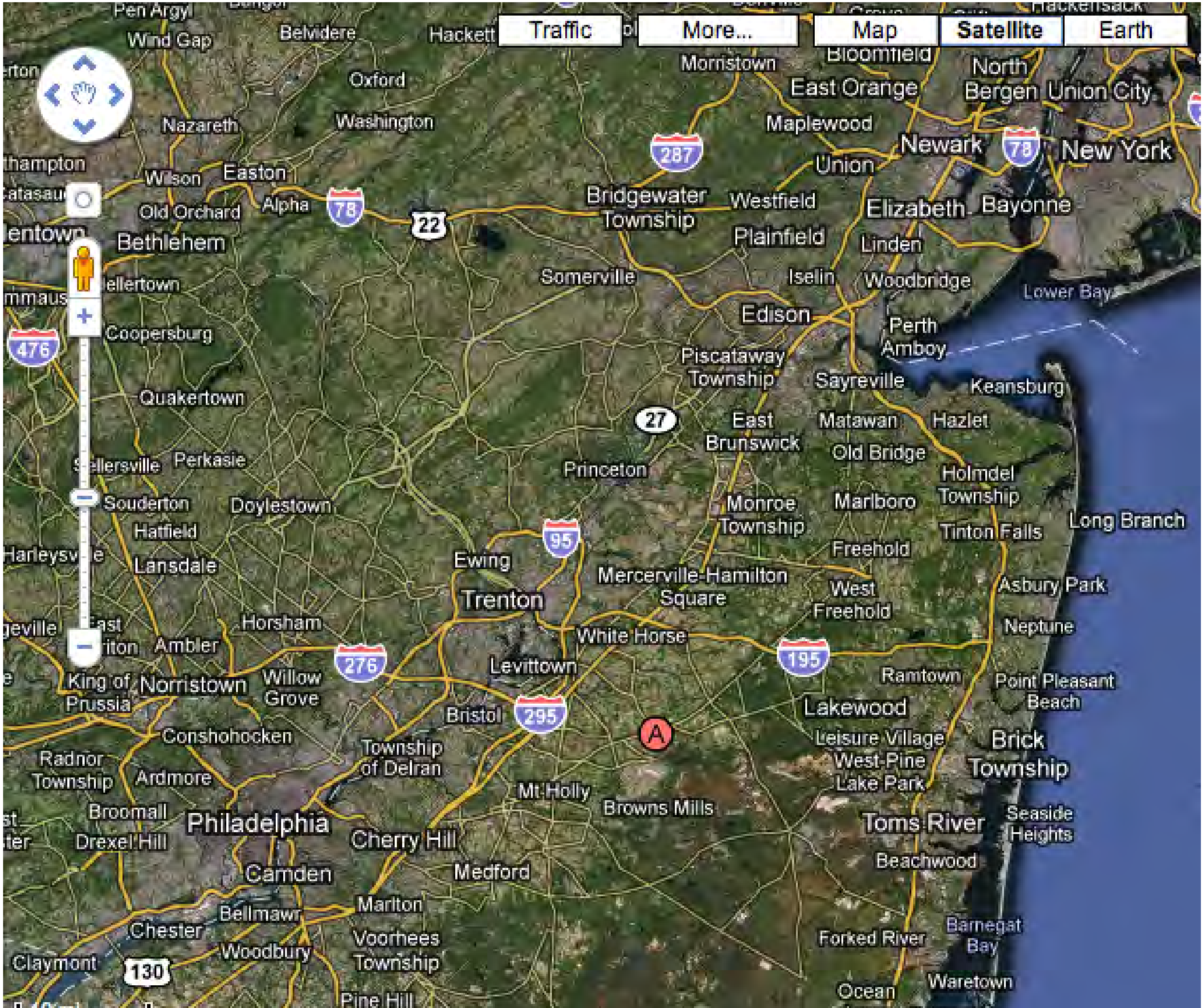
Shell script to grab 55 years of daily data for *McGuire AFB*:

http://www.princeton.edu/∼rvdb/LocalWarming/McGuireAFB/data/getData.sh

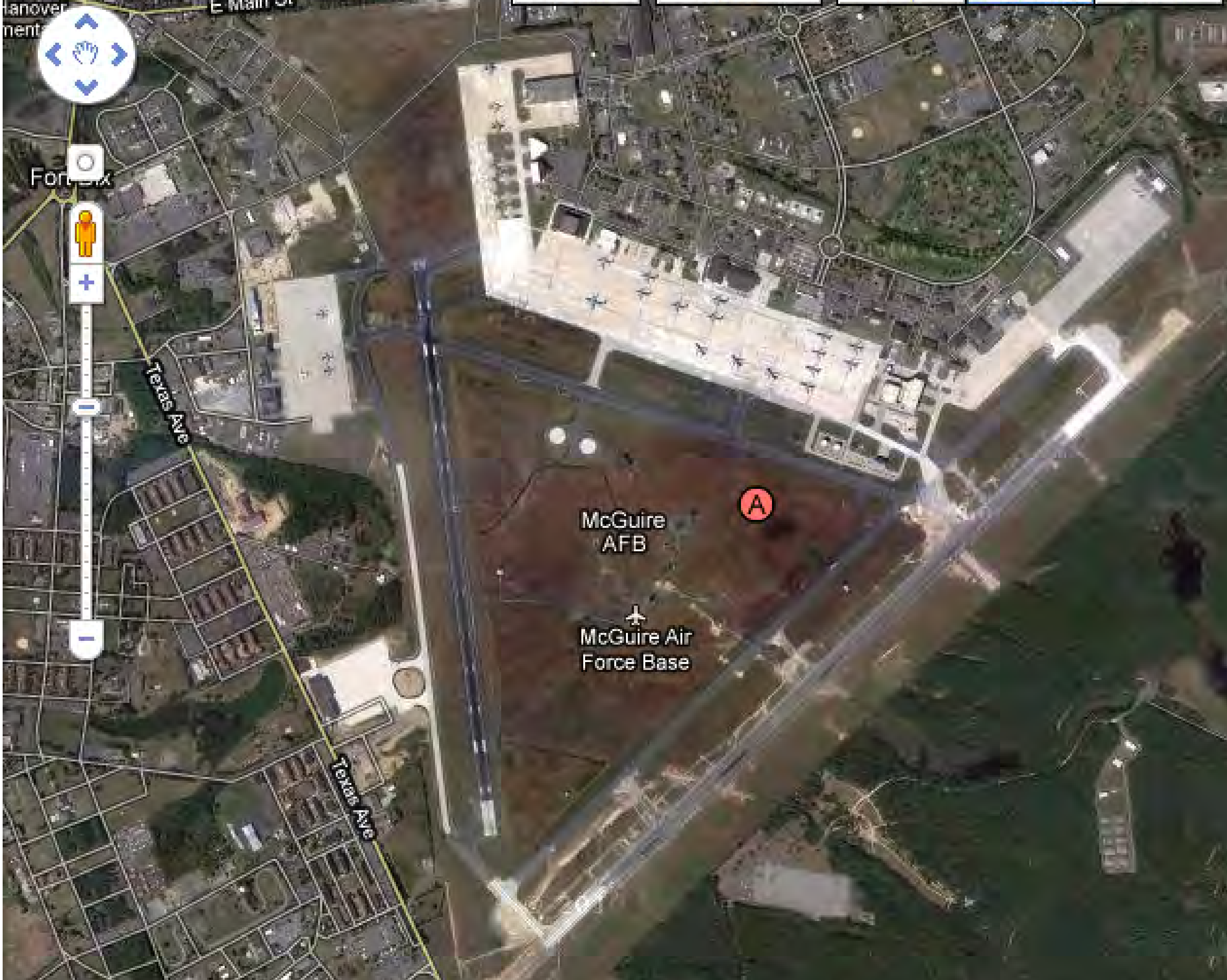
Resulting list of daily average temperatures from January 1, 1955, to August 13, 2010, is posted here...

http://www.princeton.edu/∼rvdb/LocalWarming/McGuireAFB/data/McGuireAFB.dat

McGuire AFB


Traffic
More...
Map
Satellite
Earth
Pen Argyl
Wind Gap
Belvidere
Hackett
Morristown
Bloomfield
North Bergen
Union City
East Orange
Oxford
Nazareth
Washington
Maplewood
Newark
New York
Union
Easton
Wilson
Bethlehem
Old Orchard
Alpha
Bridgewater Township
Westfield
Elizabeth
Bayonne
Plainfield
Linden
Somerville
Iselin
Woodbridge
Lower Bay
Edison
Perth Amboy
Coopersburg
Piscataway Township
Sayreville
Keansburg
Quakertown
East Brunswick
Matawan
Hazlet
Old Bridge
Princeton
Perkasie
Holmdel Township
Souderton
Doylestown
Monroe Township
Marlboro
Hatfield
Tinton Falls
Long Branch
Freehold
Ewing
Lansdale
Mercerville-Hamilton Square
West Freehold
Asbury Park
Trenton
Neptune
Horsham
White Horse
Ambler
Ramtown
Point Pleasant Beach
Levittown
Willow Grove
Norristown
King of Prussia
Lakewood
Bristol
Brick Township
Leisure Village West-Pine Lake Park
Conshohocken
Township of Delran
Radnor Township
Ardmore
Mt Holly
Browns Mills
Seaside Heights
Broomall
Philadelphia
Toms River
Drexel Hill
Cherry Hill
Beachwood
Camden
Medford
Marlton
Bellmawr
Chester
Barnegat Bay
Voorhees Township
Forked River
Woodbury
Claymont
Waretown
Pine Hill
Ocean Acres
287
78
22
476
27
95
276
295
195
130
A

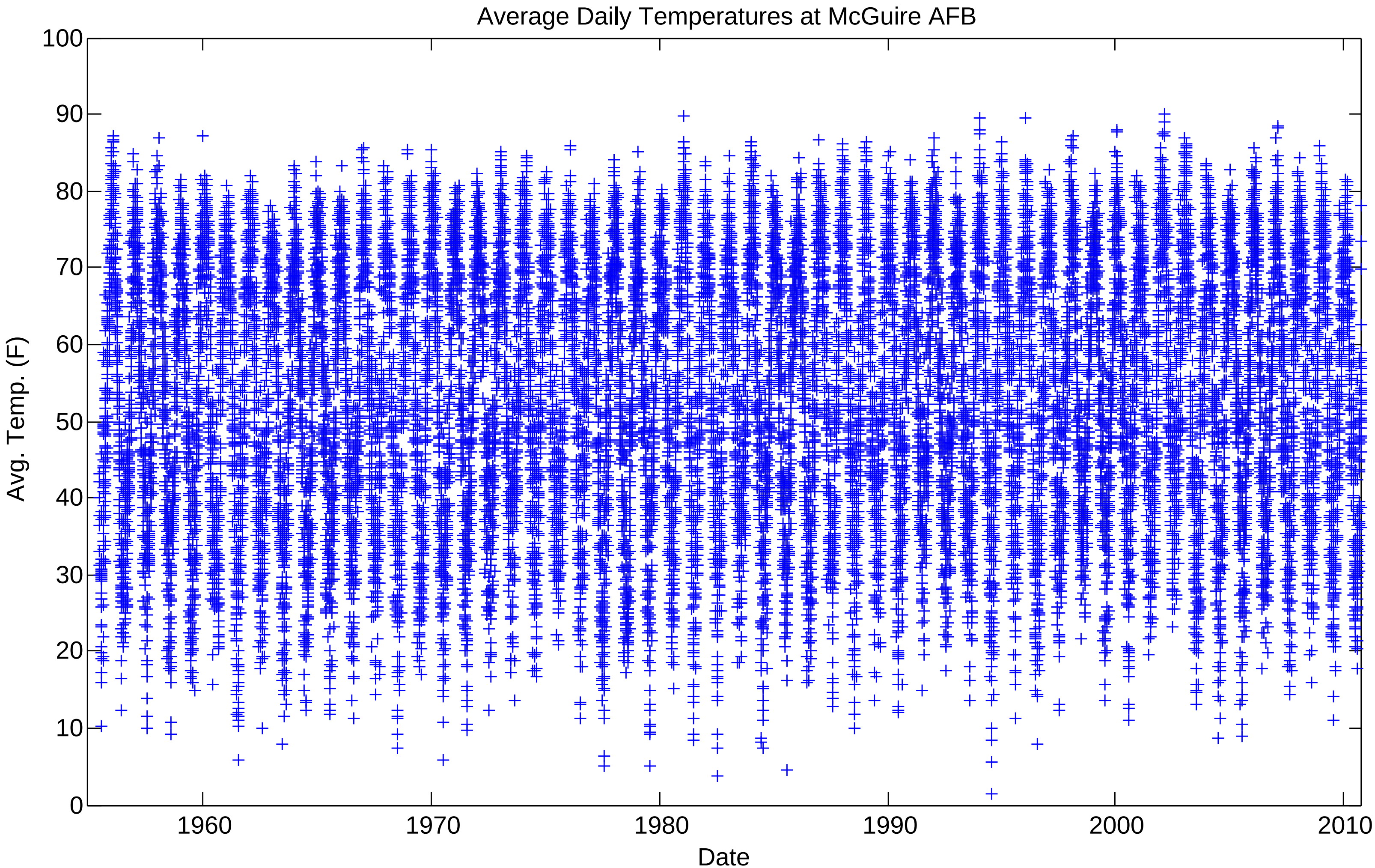

Cookstown
Bunting Bridge Rd
McGuire AFB
McGuire Access Hwy
Sykesville Rd
Hockamick Rd
Crowshaw Park
E Main St
Wrightstown
Pond Rd
Fort Dix
Pemberton Wrightstown Rd
68
Texas Ave
Browns Mills Cookstown Rd
McGuire AFB
A
McGuire Air Force Base
Texas Ave
Range Rd
Cookstown-Browns Mills Rd
Range Rd



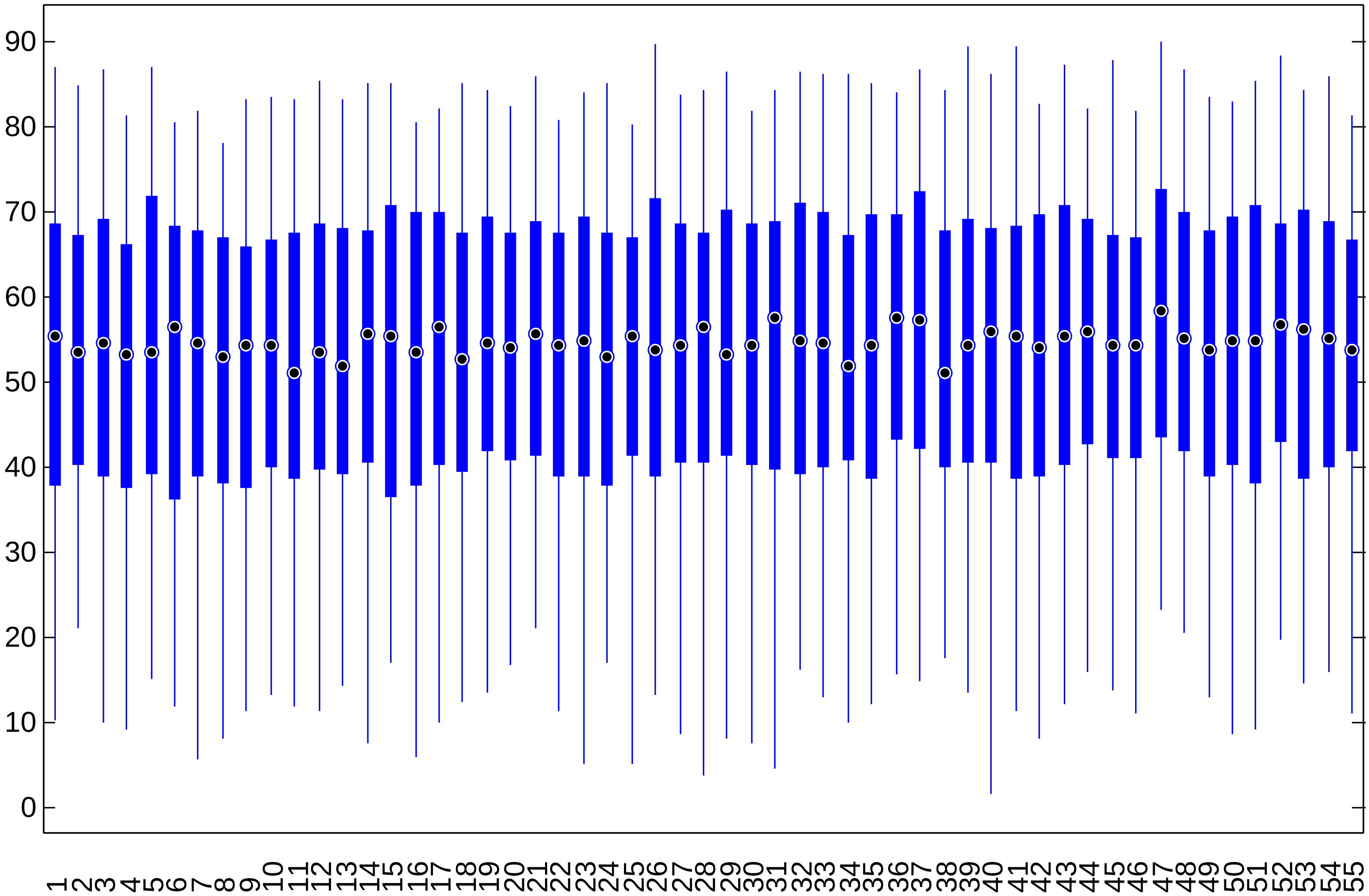
Fort Dix
Texas Ave
Texas Ave
McGuire AFB
McGuire Air Force Base
A

Daily Temperatures

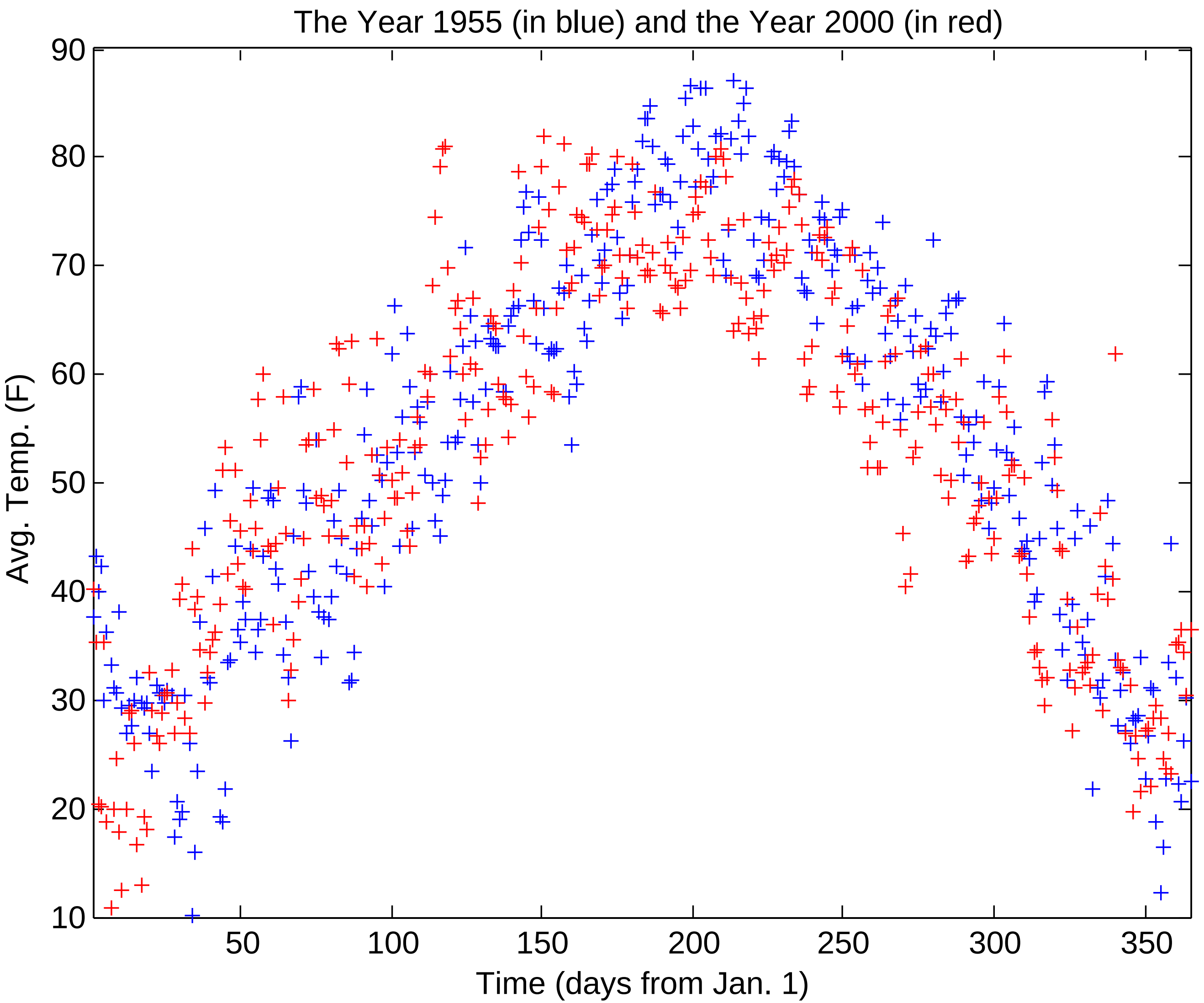
McGuire AFB | Data From NOAA | 55+ Years (20,309 days)



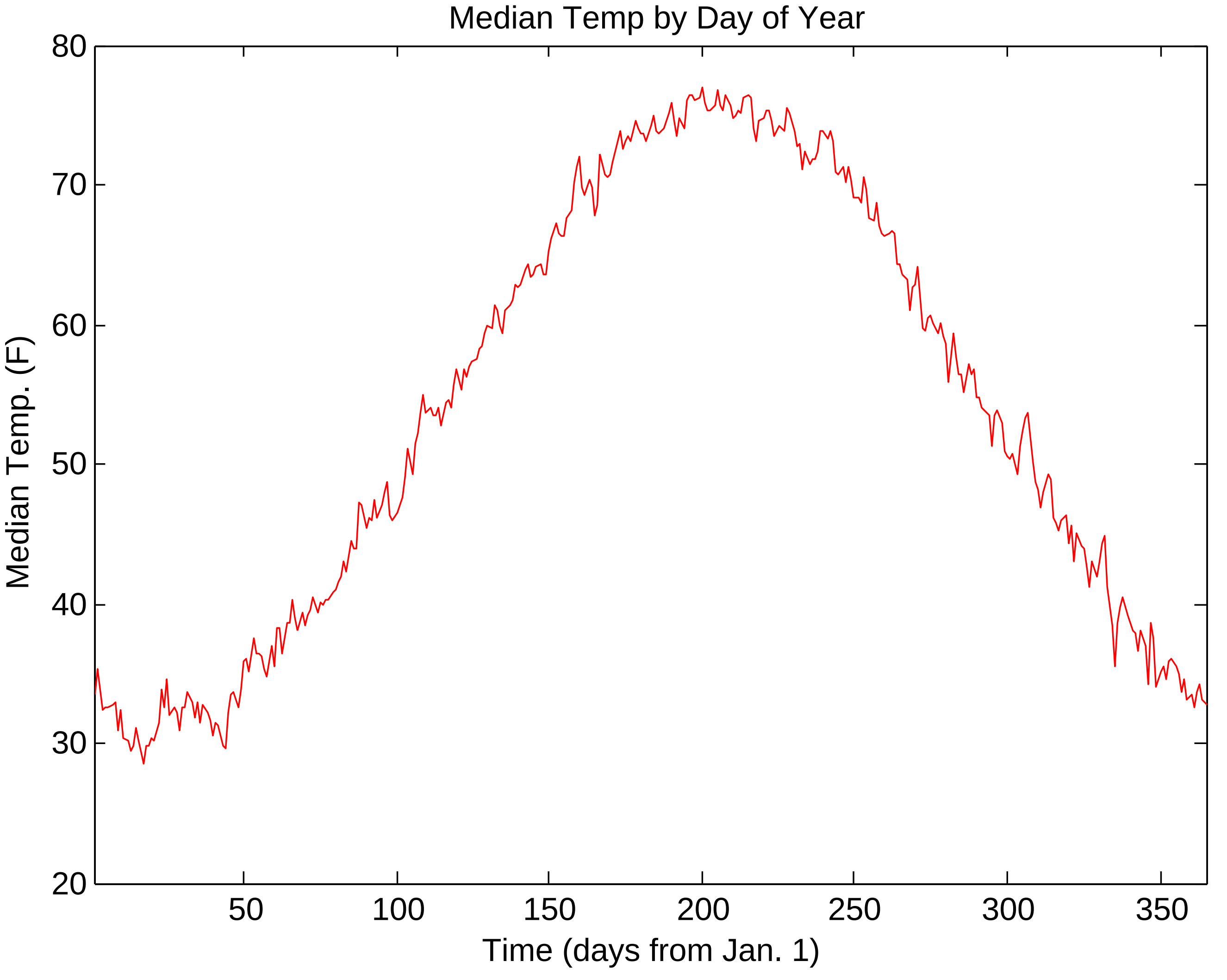
Box Plots of Same



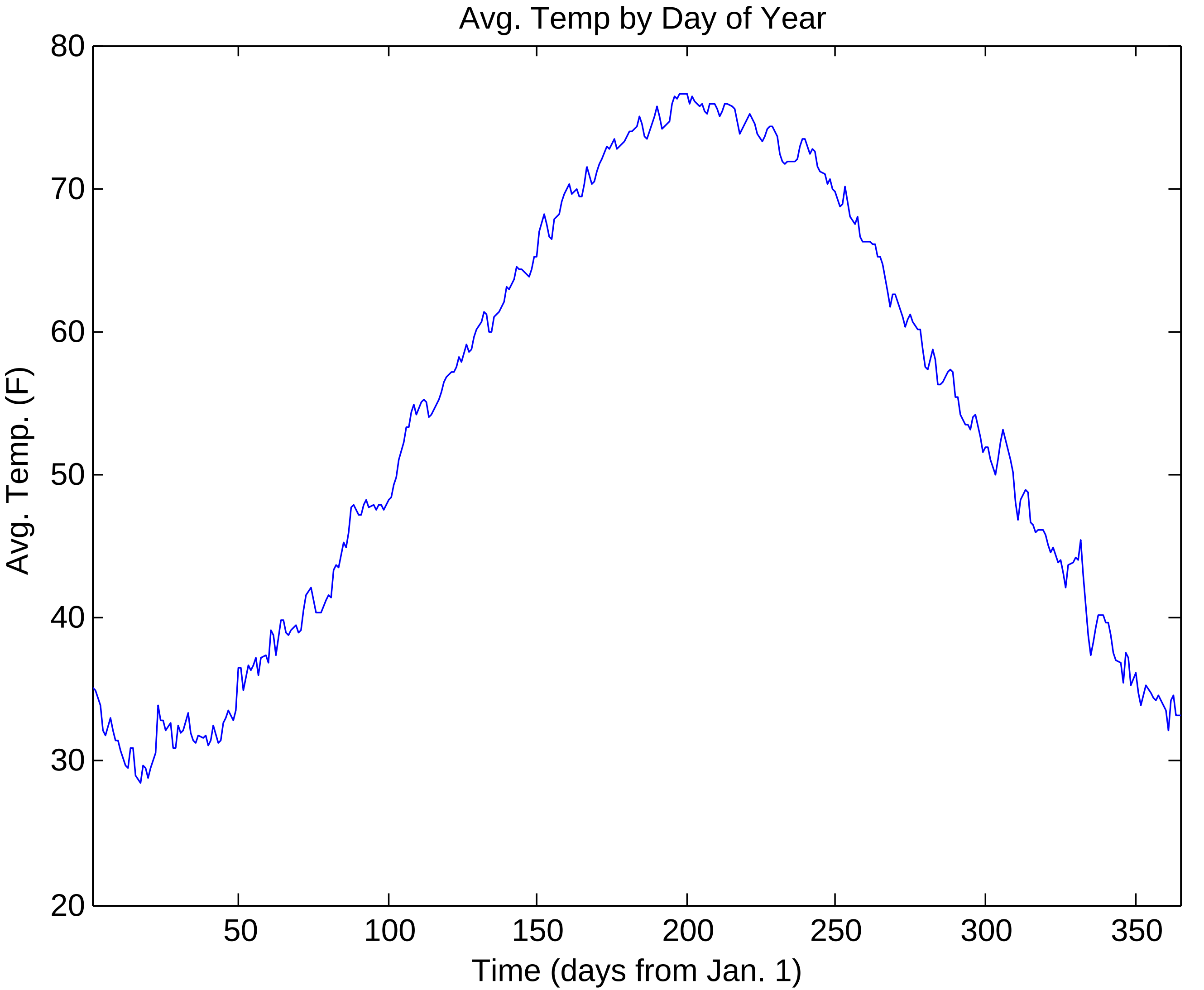
Two Years Overlayed



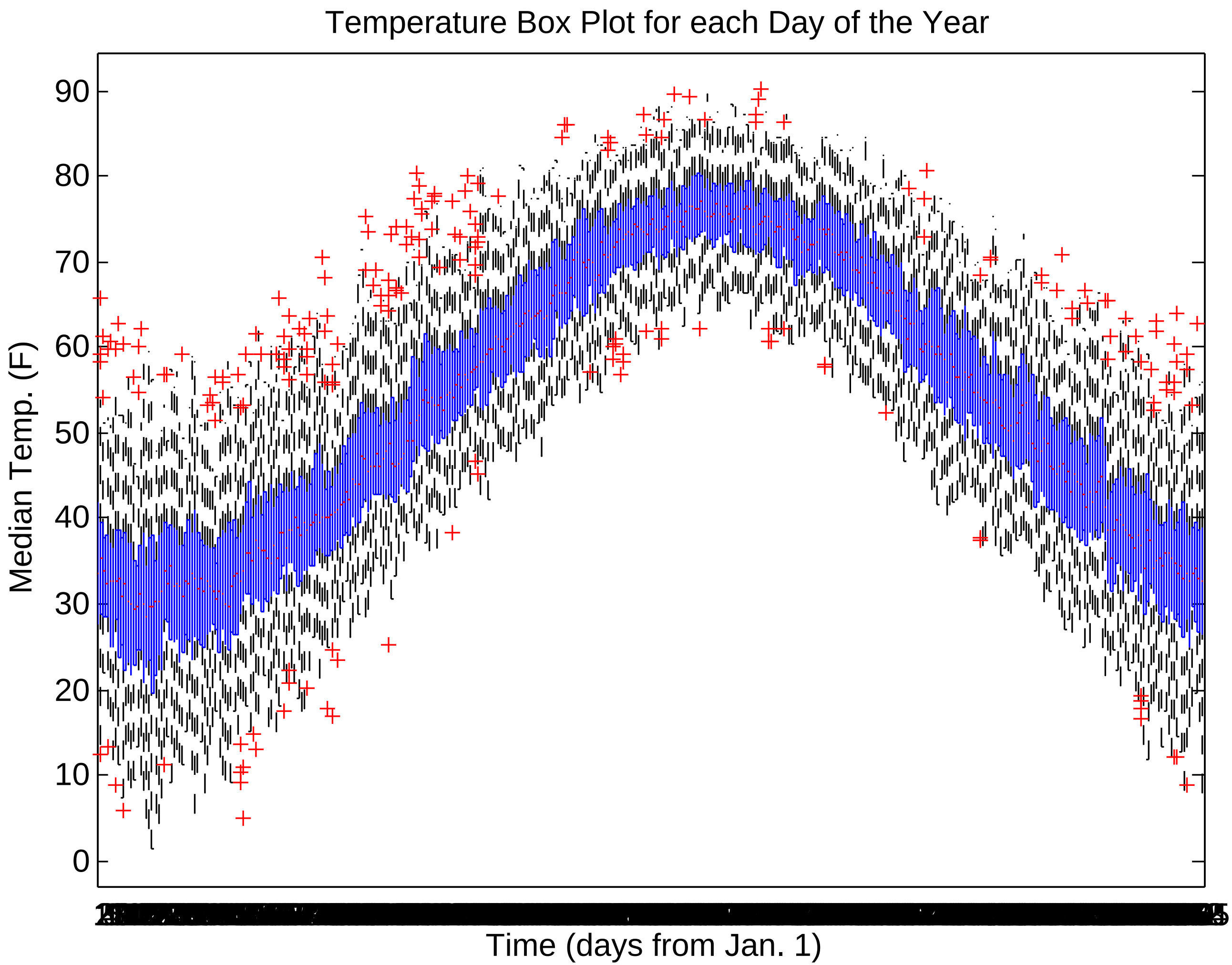
Median Temperature by Day of Year



Me an Temperature by Day of Year



BoxPlot of Temperature by Day of Year

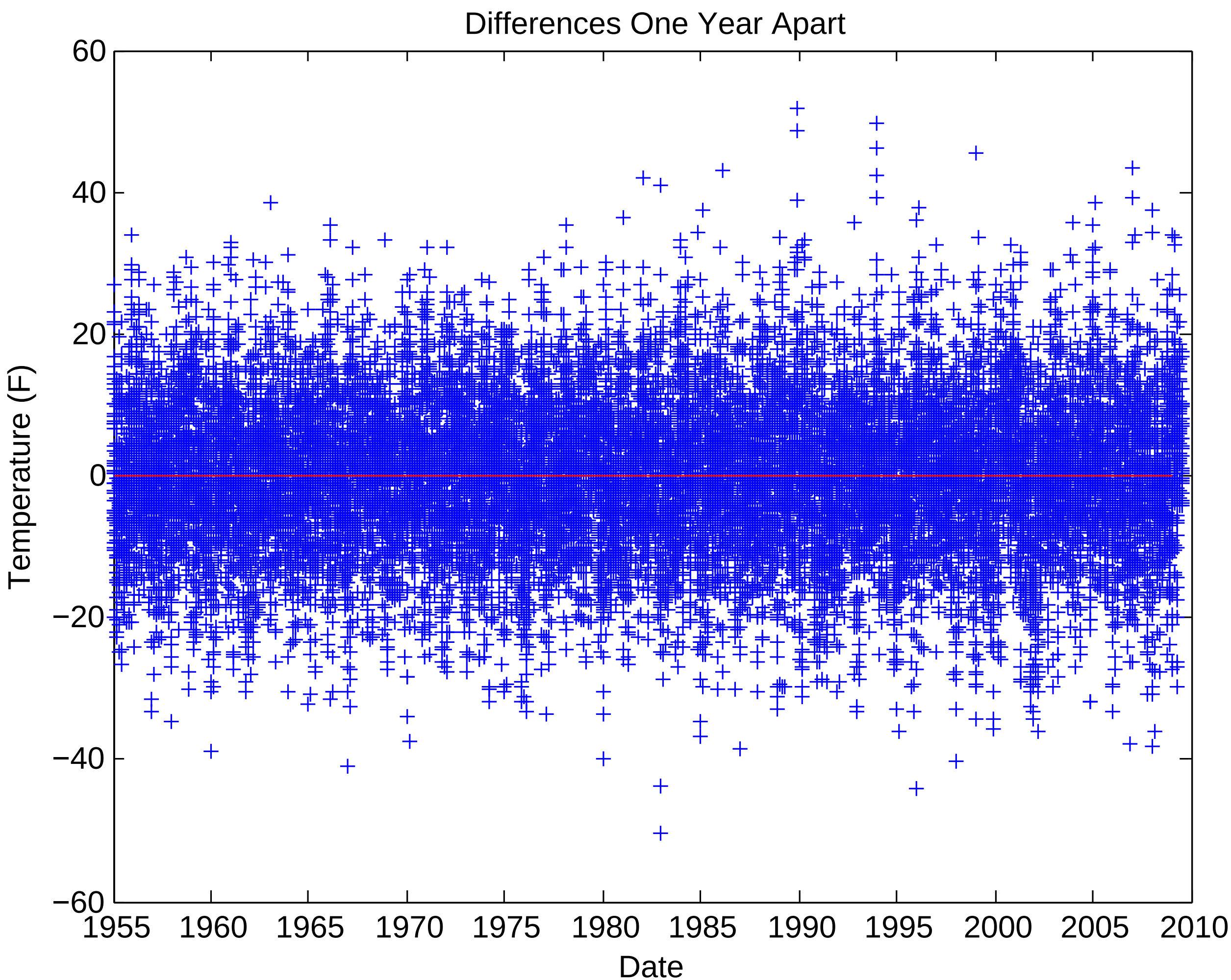


Seasonal Variation Dominates

How Can We Remove The Seasonal Variation?

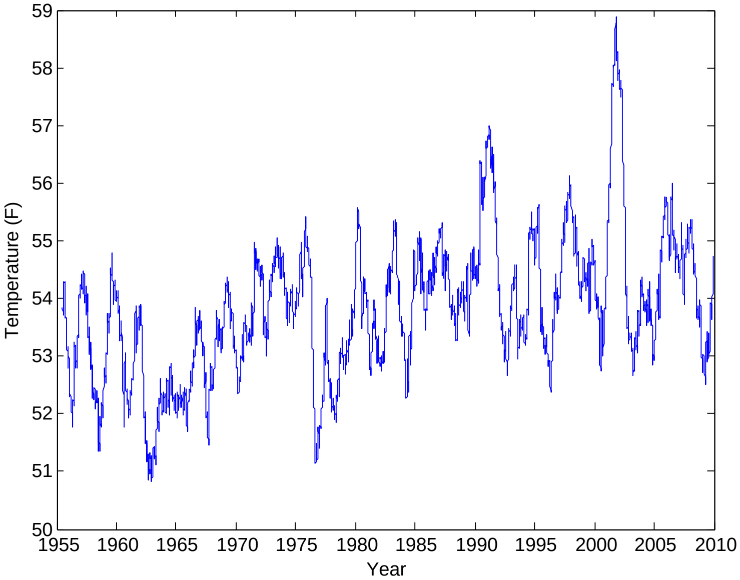
1. Take year-to-year differences.
2. Average.

Differences One Year Apart

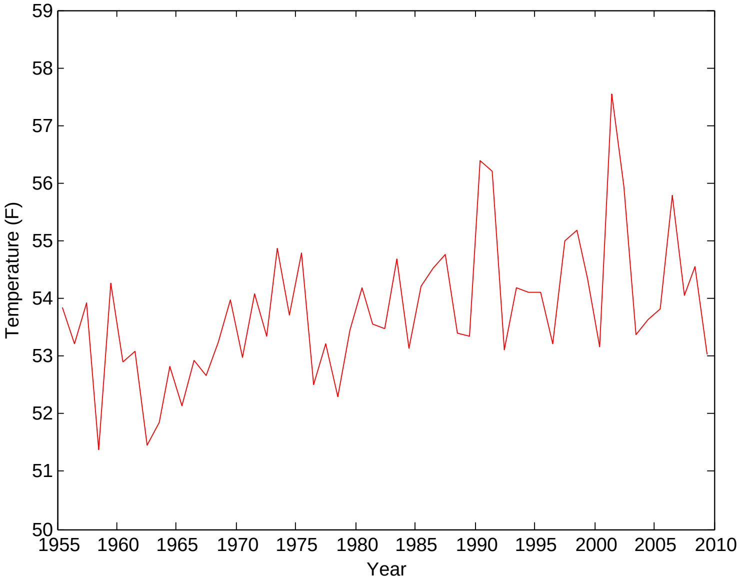


Mean difference: 2.89 °F per century. Std deviation: ± 7.40 °F per century. *Ouch!*

One-Year (365 Day) Averages

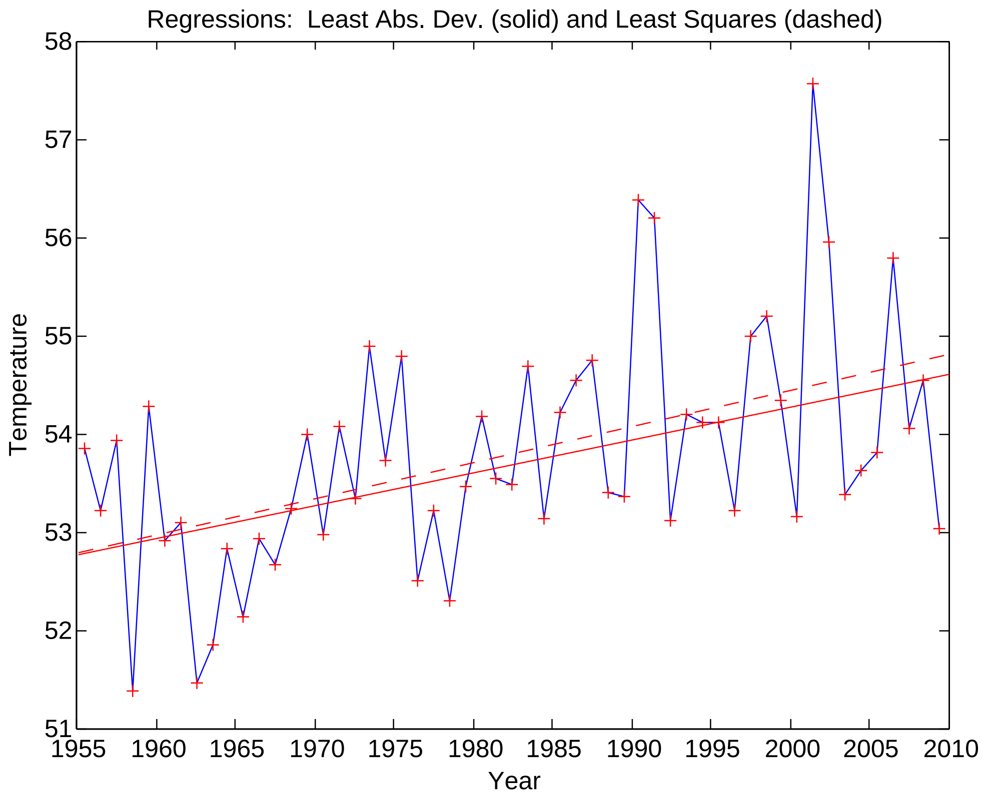


Rolling Average



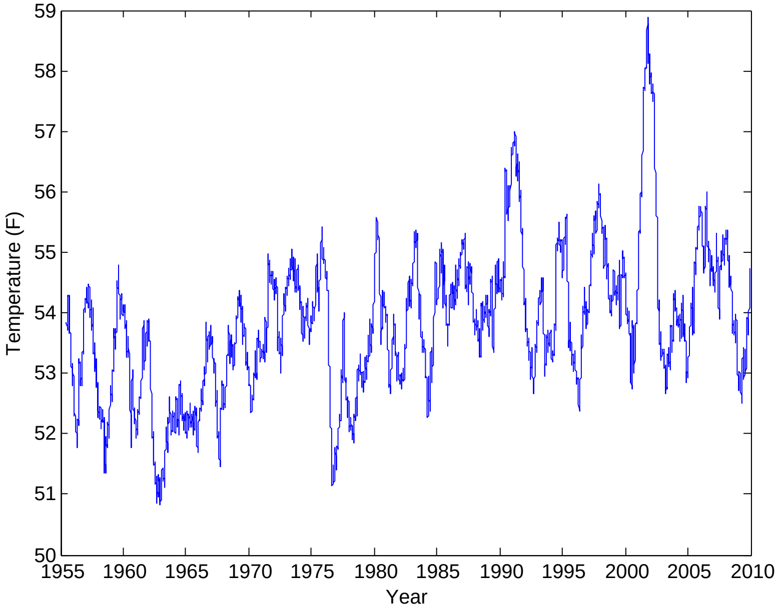
Year-by-Year

Year-By-Year With Regression Line

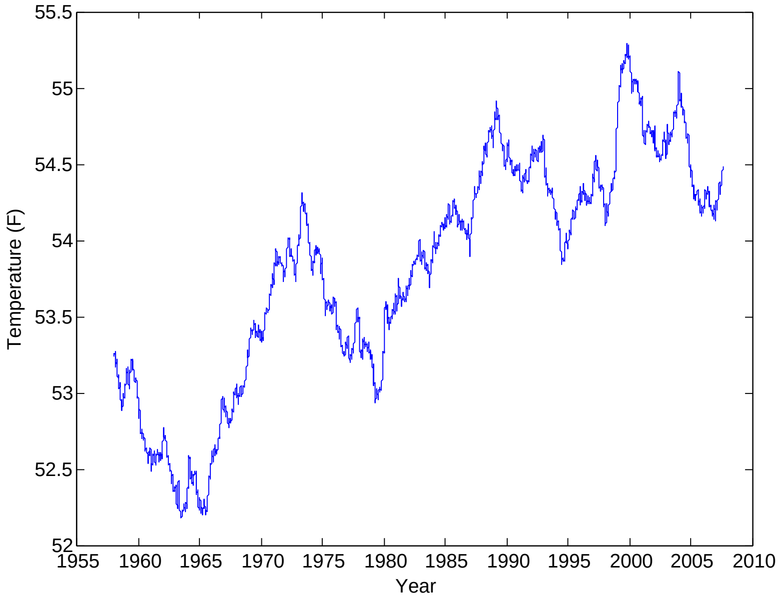


	Least-Abs-Dev	Least-Squares
Average Temperature in 1955 (°F):	52.77	52.80
Rate of Temperature Change (°F/century):	3.32	3.65

One-Year vs. Six-Year Rolling Averages



One-Year Averages

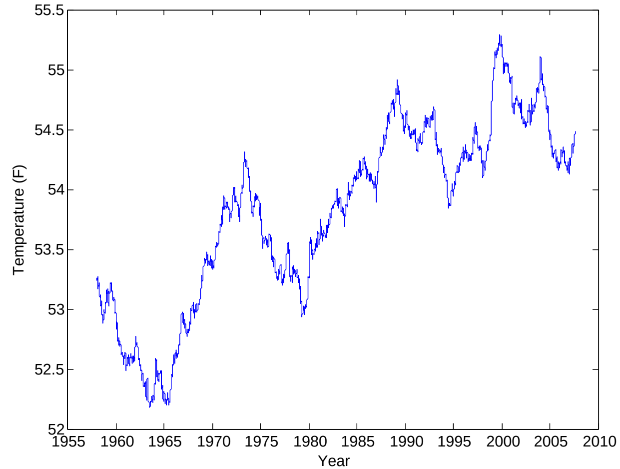


Six-Year Averages

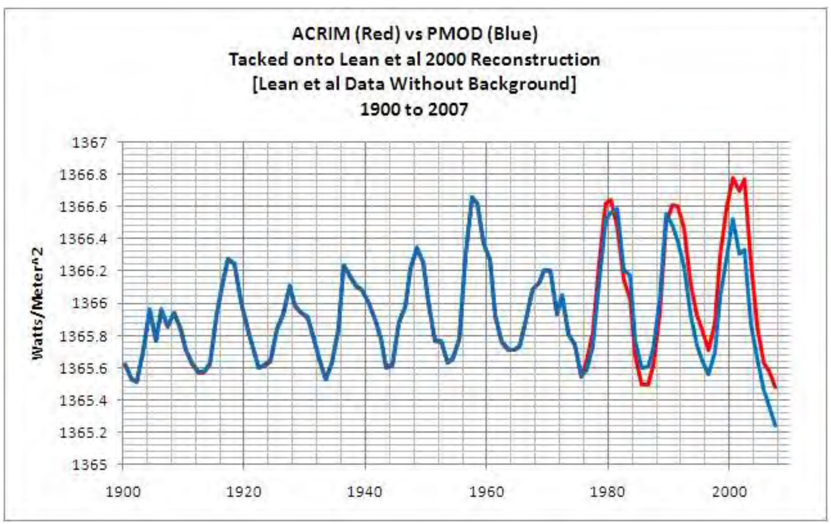
There seems to be a periodic variation!

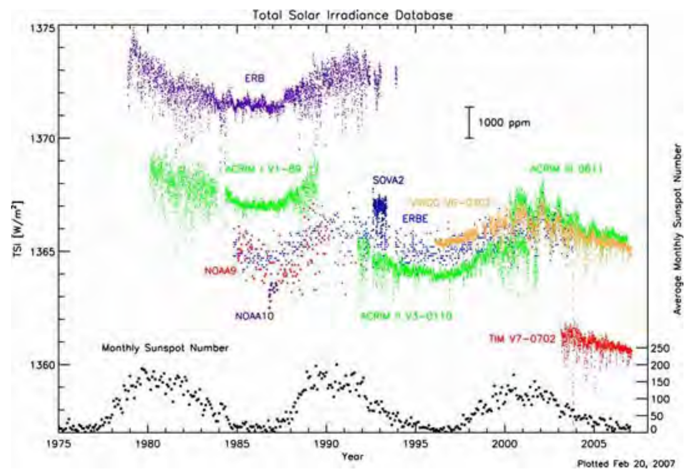
Period ≈ 11 years.

Comparison With Solar Cycle



Six-Year Averages





Solar Irradiance Graphs

A Model Using All The Data

Let T_d denote the average temperature in degrees Fahrenheit on day $d \in D$ where D is the set of days from January 1, 1955, to August 13, 2010 (a whopping 20,309 days!).

$$\begin{aligned} T_d = \; & x_0 + x_1 d && \textit{linear trend} \\ & + x_2 \cos(2\pi d/365.25) + x_3 \sin(2\pi d/365.25) && \textit{seasonal cycle} \\ & + x_4 \cos(x_6 2\pi d/(10.7 \times 365.25)) + x_5 \sin(x_6 2\pi d/(10.7 \times 365.25)) && \textit{solar cycle} \\ & + \varepsilon_d. && \textit{error term} \end{aligned}$$

The parameters $x_0, x_1, \ldots, x_6$ are unknown regression coefficients.

Either

$$\min \sum_{d \in D} |\varepsilon_d| \qquad \textit{Least Absolute Deviations (LAD)}$$

or

$$\min \sum_{d \in D} \varepsilon_d^2 \qquad \textit{Least Squares}$$

Linearizing the Solar Cycle

If the unknown parameter x_6 is *fixed at* 1, forcing the solar-cycle to have a period of exactly 10.7 years, then the problem can be reduced to a *linear programming problem*.

If, on the other hand, we allow x_6 to vary, then the problem is *nonlinear* and even *nonconvex* and therefore harder in principle. Nonetheless, if we initialize x_6 to one, then the problem might, and in fact does, prove to be tractable.

Note: The *least-absolute-deviations* (LAD) model automatically ignores "outliers".

AMPL Code For LAD Model

```
set DATES ordered;
param avg {DATES};
param day {DATES};
param pi := 4*atan(1);

var a {j in 0..6};
var dev {DATES} >= 0, := 1;

minimize sumdev: sum {d in DATES} dev[d];
subject to def_pos_dev {d in DATES}:
    x[0] + x[1]*day[d] + x[2]*cos( 2*pi*day[d]/365.25)
                       + x[3]*sin( 2*pi*day[d]/365.25)
                       + x[4]*cos( x[6]*2*pi*day[d]/(10.7*365.25))
                       + x[5]*sin( x[6]*2*pi*day[d]/(10.7*365.25))
        - avg[d]
    <= dev[d];
subject to def_neg_dev {d in DATES}:
    -dev[d] <=
    x[0] + x[1]*day[d] + x[2]*cos( 2*pi*day[d]/365.25)
                       + x[3]*sin( 2*pi*day[d]/365.25)
                       + x[4]*cos( x[6]*2*pi*day[d]/(10.7*365.25))
                       + x[5]*sin( x[6]*2*pi*day[d]/(10.7*365.25))
        - avg[d];
```

AMPL Data and Variable Initialization

```
data;

set DATES := include "data/Dates.dat";
param: avg := include "data/McGuireAFB.dat";
let {d in DATES} day[d] := ord(d,DATES);

let x[0] := 60;
let x[1] := 0;
let x[2] := 20;
let x[3] := 20;
let x[4] := 0.01;
let x[5] := 0.01;
let x[6] := 1;
```

The nice thing about AMPL and LOQO is that anyone can use these programs via the NEOS server at Argonne National Labs...

http://www-neos.mcs.anl.gov/

The Results

The linear version of the problem solves in a small number of iterations and only takes a minute or so on my MacBook Pro laptop computer. The nonlinear version takes more iterations and more time but eventually converges to a solution that is almost identical to the solution of the linear version. The optimal values of the parameters are

$$\begin{aligned}
x_0 &= \quad 52.6\ ^\circ\text{F} \\
x_1 &= \quad 9.95 \times 10^{-5}\ ^\circ\text{F/day} \\
x_2 &= -20.4\ ^\circ\text{F} \\
x_3 &= \quad -8.31\ ^\circ\text{F} \\
x_4 &= \quad -0.197\ ^\circ\text{F} \\
x_5 &= \quad 0.211\ ^\circ\text{F} \\
x_6 &= \quad 0.992
\end{aligned}$$

From x_0, we see that the nominal temperature at McGuire AFB was 52.56 °F (on January 1, 1955).

We also see, from x_1, that there is a positive trend of 0.000099 °F/day. That translates to 3.63 °F per century—in excellent agreement with results from global climate change models.

Using *bootstrap*, a 95% *confidence interval* for x_1 is $[2.88\ ^\circ\text{F},\ 4.38\ ^\circ\text{F}]$/century.

Magnitude of the Sinusoidal Fluctuations

From x_2 and x_3, we can compute the amplitude of annual seasonal changes in temperatures...

$$\sqrt{x_2^2 + x_3^2} = 22.02 \text{ °F}.$$

In other words, on the hottest summer day we should expect the temperature to be 22.02 degrees warmer than the nominal value of 52.56 degrees; that is, 77.58 degrees. Of course, this is a daily average—daytime highs will be higher and nighttime lows should be about the same amount lower.

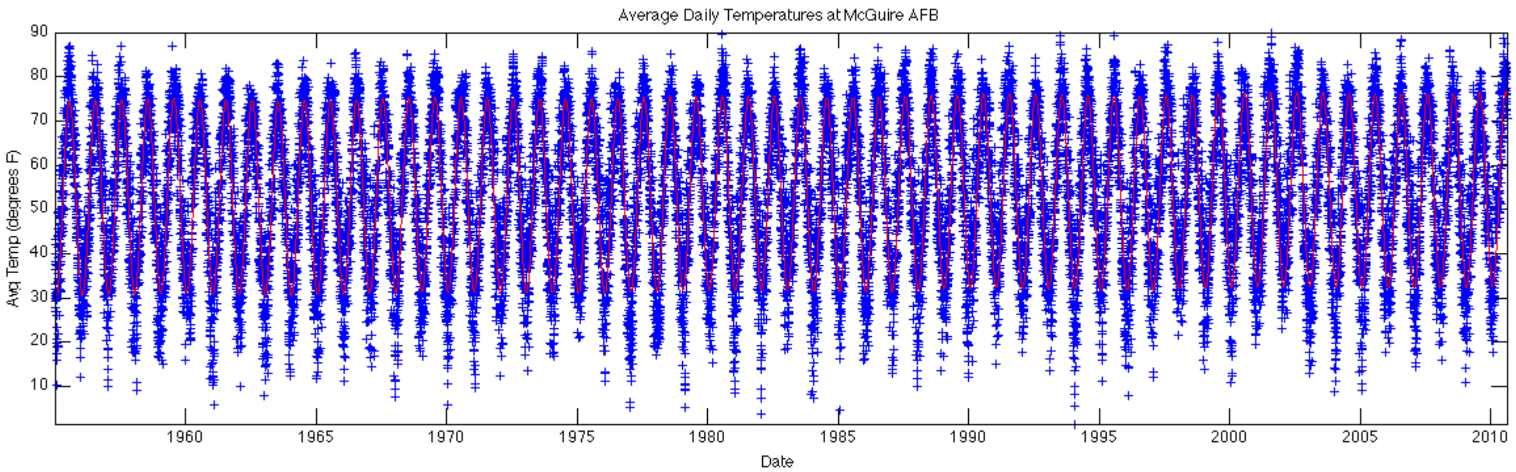
Similarly, from x_4 and x_5, we can compute the amplitude of the temperature changes brought about by the solar-cycle...

$$\sqrt{x_4^2 + x_5^2} = 0.2887 \text{ °F}.$$

The effect of the *solar cycle* is real but relatively small.

The fact that x_6 came out slightly less than one indicates that the solar cycle is slightly longer than the nominal 10.7 years. It's closer to $10.7/x_6 = 10.78$ years.

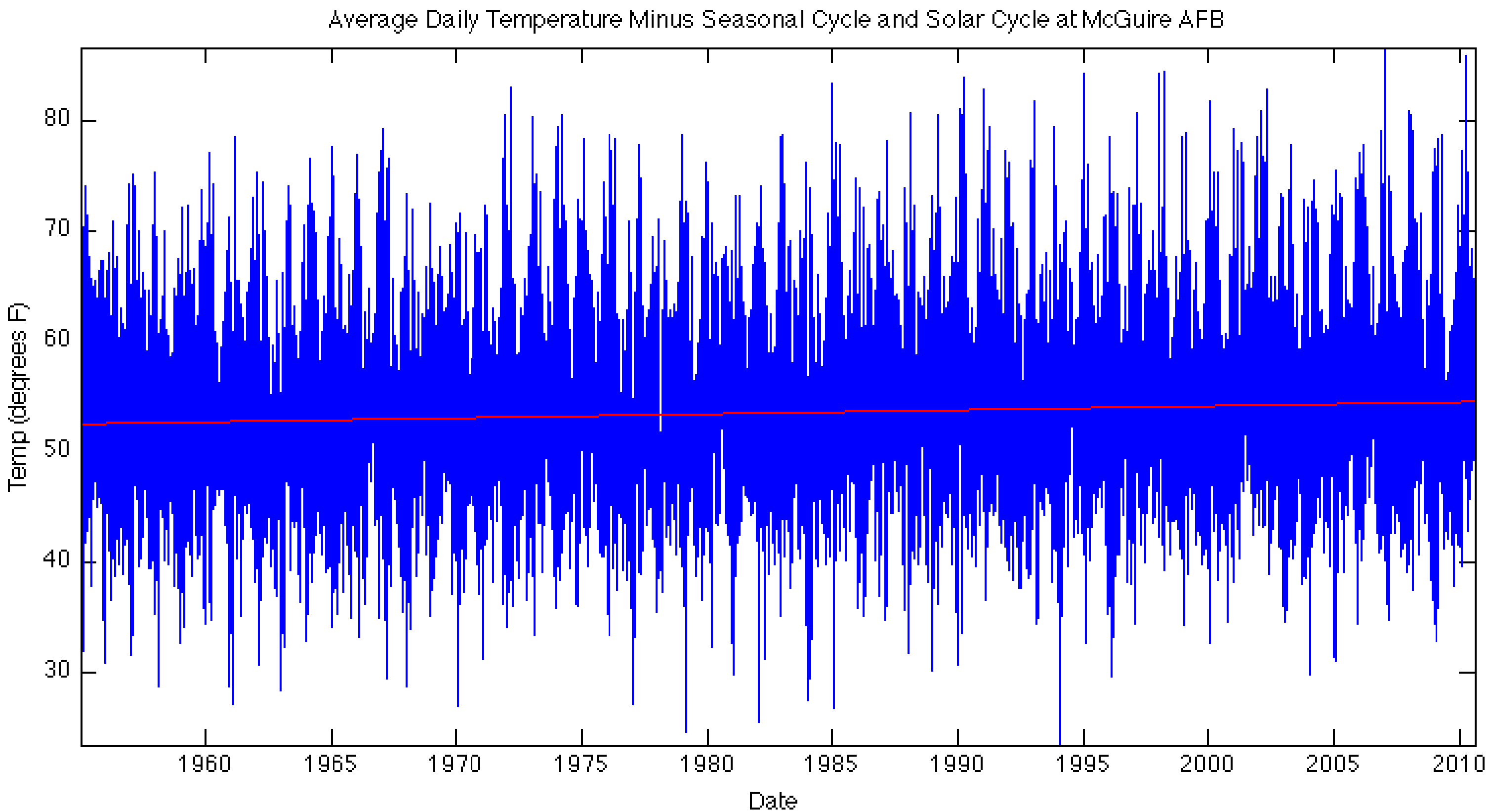
Plot Showing Actual Data and Regression Curve



Blue: Average daily temperatures at McGuire AFB from 1955 to 2010.
Red: Output from least absolute deviation regression model.

Seasonal fluctuations completely dominate other effects.

Subtracting Out Seasonal Effects

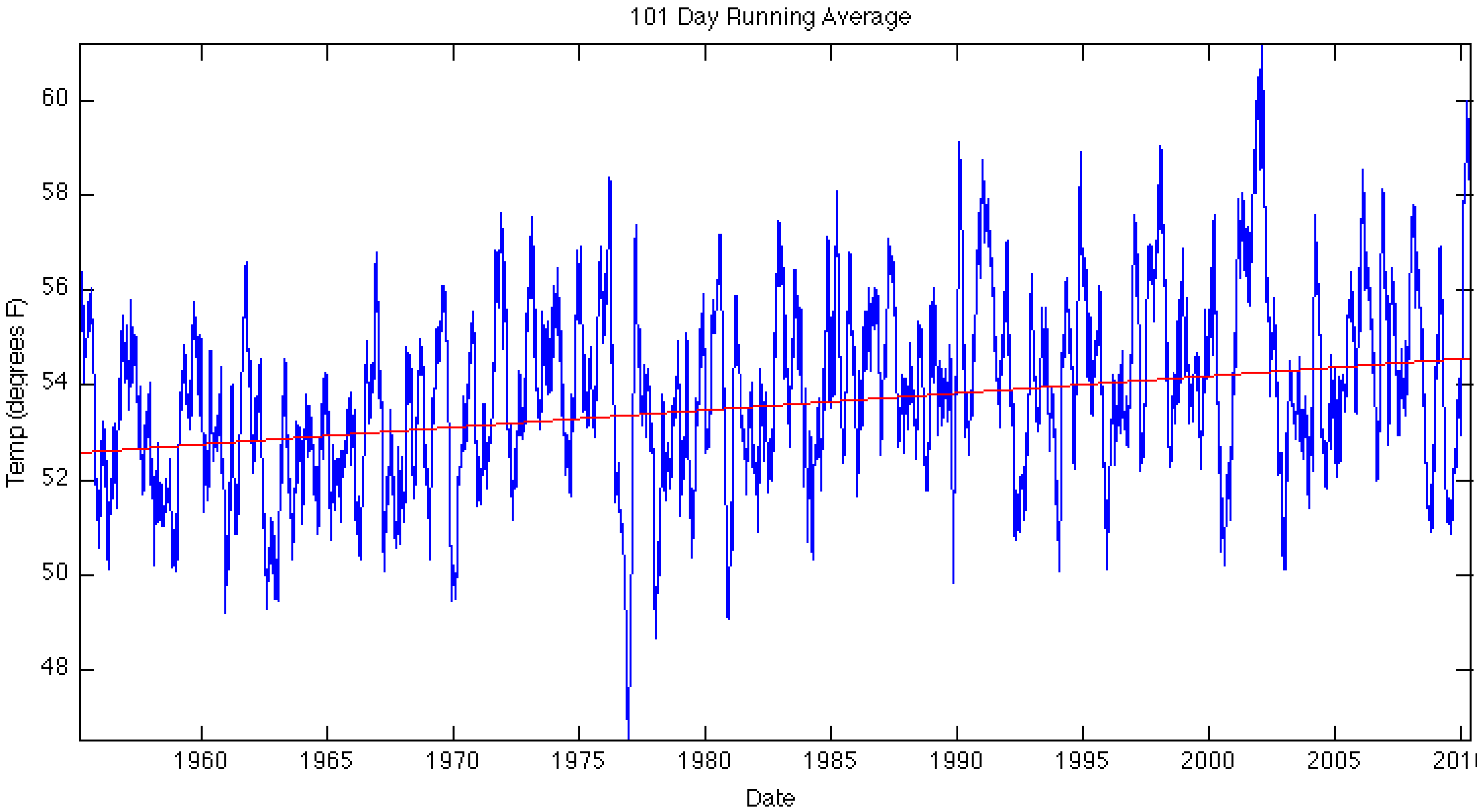


As before but with sinusoidal seasonal variation removed and sinusoidal solar-cycle variation removed as well.

Even this plot is noisy simply because there are many days in a year and some days are unseasonably warm while others are unseasonably cool.

Smoothed Seasonally Subtracted Plot

To smooth out high frequency fluctuations, we use 101 day rolling averages of the data.



In this plot, the long term trend in temperature is clearly seen. In NJ we have *local warming*.

Autoregression

Modify the model as follows:

$$
\begin{aligned}
T_d = \;& x_0 + x_1 d && \textit{linear trend} \\
& + x_2 \cos(2\pi d/365.25) + x_3 \sin(2\pi d/365.25) && \textit{seasonal cycle} \\
& + x_4 \cos(x_6 2\pi d/(10.7 \times 365.25)) + x_5 \sin(x_6 2\pi d/(10.7 \times 365.25)) && \textit{solar cycle} \\
& + \sum_{j=1}^{30} \lambda_j T_{d-j} && \textit{autoregressive terms} \\
& + \varepsilon_d && \textit{error term}
\end{aligned}
$$

with the constraint

$$\sum_{j=1}^{30} \lambda_j = 0.$$

The new parameters $\lambda_1, \lambda_2, \ldots, \lambda_{30}$ capture correlation from one day to the next.

Results

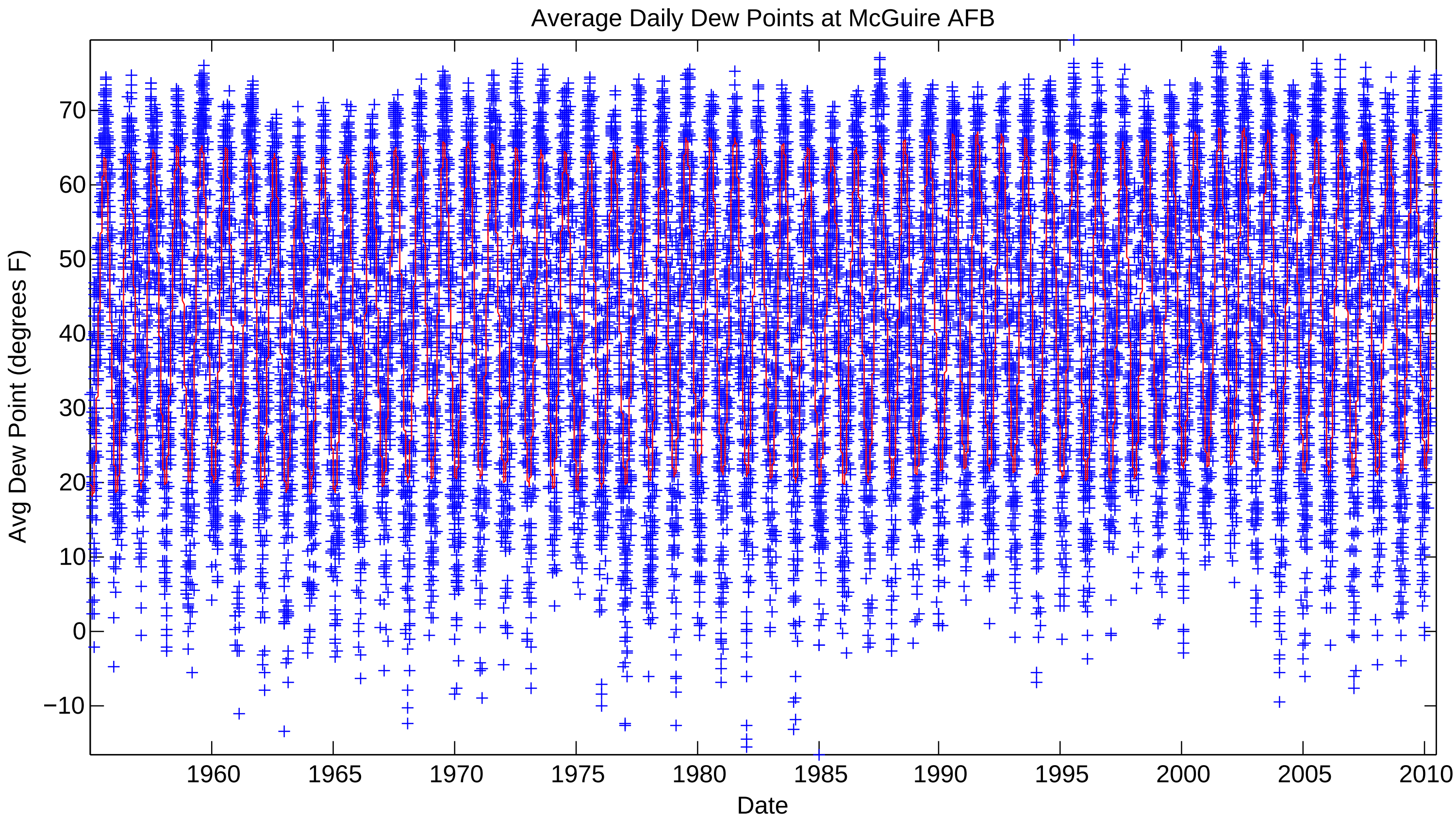
Warming rate: 3.63 °F per century—same as before.

lambda [*] :=

1	0.782569	9	-0.015877	17	-0.0158704	25	-0.039922
2	-0.300232	10	-0.018313	18	-0.0241674	26	-0.011331
3	0.085447	11	-0.031793	19	-0.0228709	27	-0.024117
4	-0.052355	12	-0.015495	20	-0.0102066	28	-0.023411
5	0.014926	13	-0.018906	21	-0.0130837	29	-0.006260
6	-0.026622	14	-0.023404	22	-0.0366945	30	-0.067680
7	-0.017368	15	0.008520	23	-0.0051124		
8	-0.016240	16	-0.032370	24	-0.0217553		

Dew Point

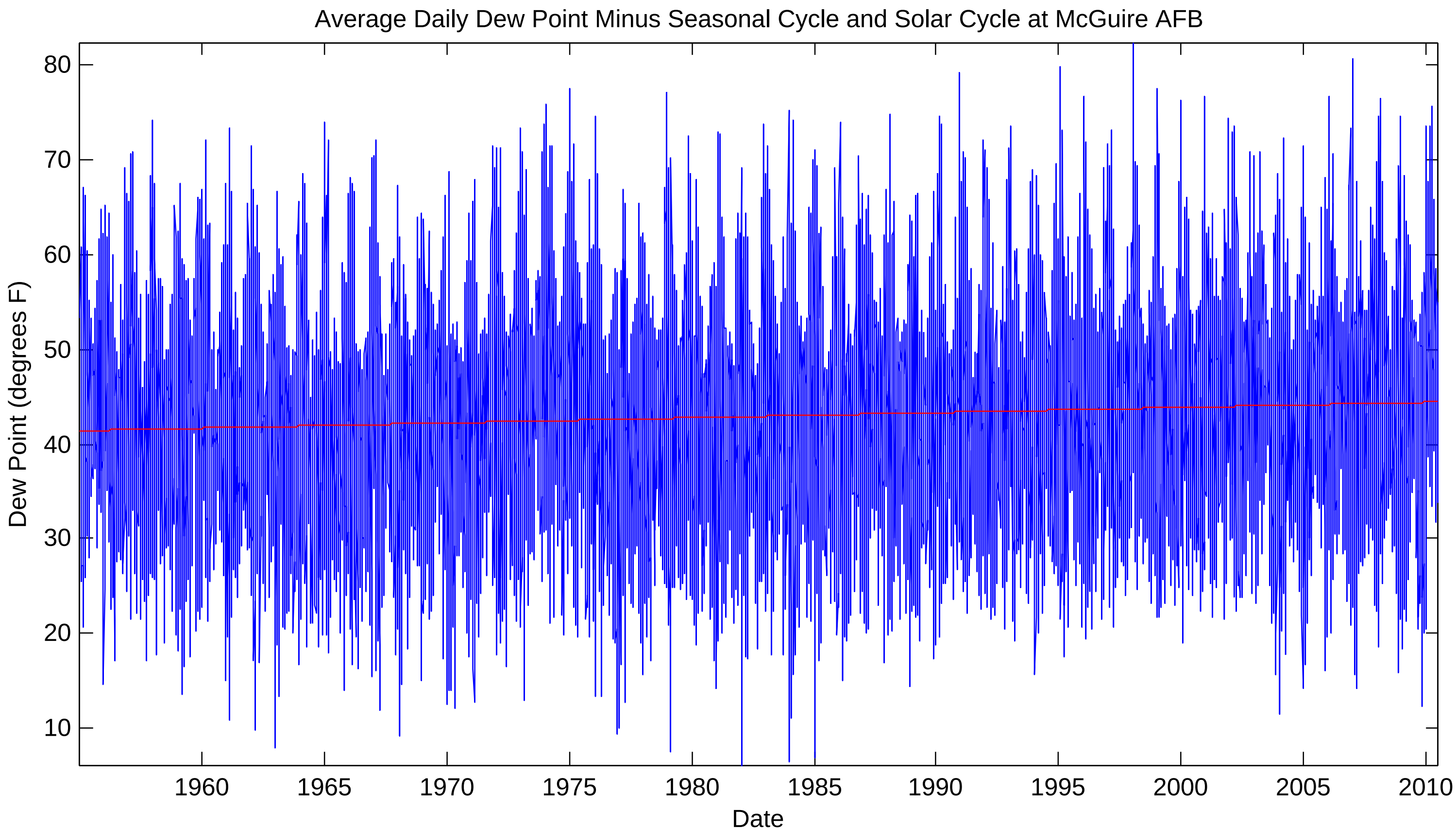
Plot Showing Actual Data and Regression Curve



Blue: Average Dew Points at McGuire AFB from 1955 to 2010.
Red: Output from least absolute deviation regression model.

As with temperature, seasonal fluctuations completely dominate other effects.

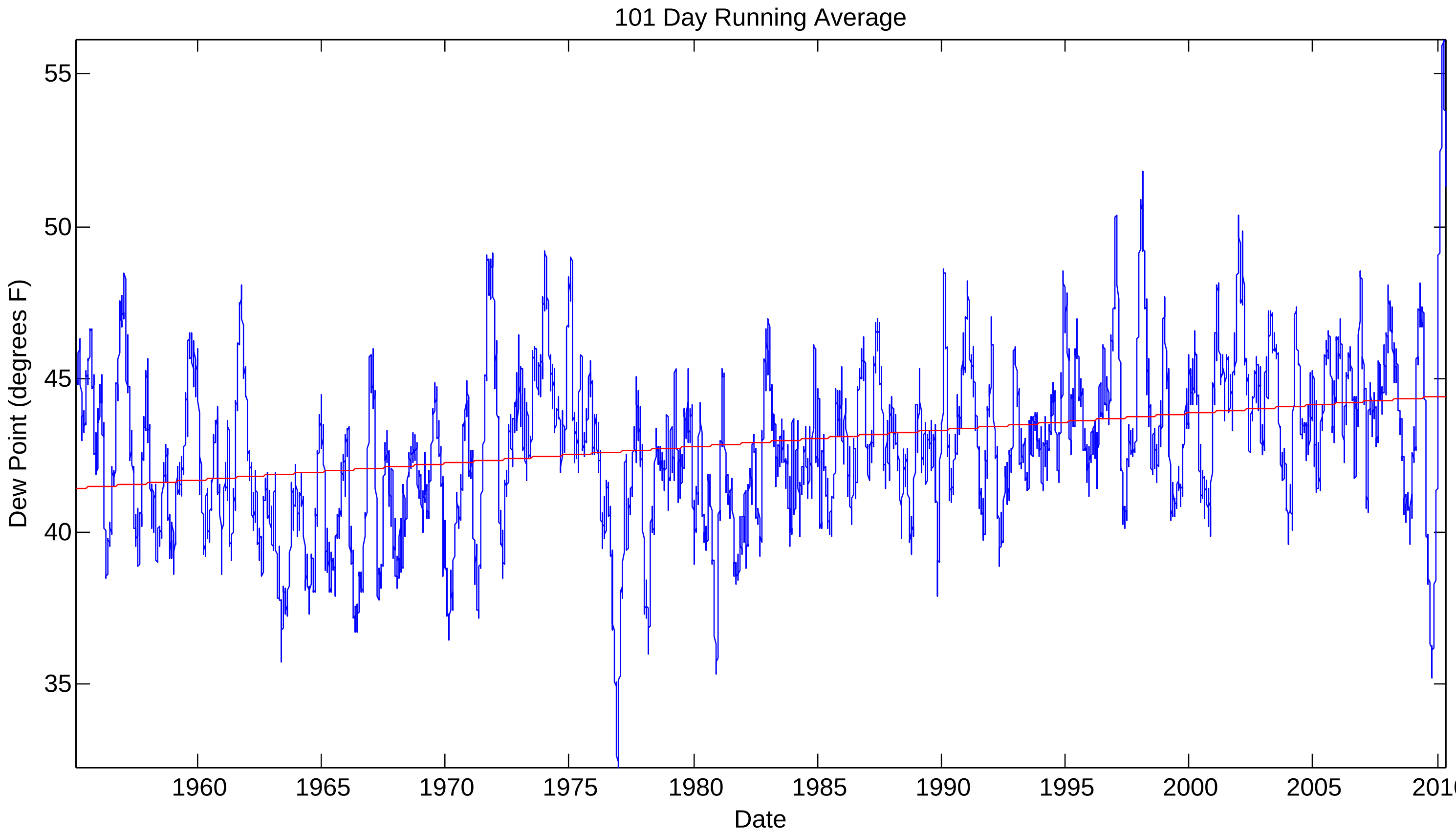
Subtracting Out Seasonal Effects



As previous slide but with sinusoidal seasonal variation removed and sinusoidal solar-cycle variation removed as well.

Even this plot is noisy simply because there are many days in a year and some days are unseasonably damp while others are unseasonably dry.

Smoothed Seasonally Subtracted Plot



Dew point is going up at a rate of 5.51 °F per century—faster than the rate at which temperatures are increasing (3.63 °F per century).

In NJ we have *local damping!*

Why Least Absolute Deviations?

Means, Medians, and Optimization

Let $b_1, b_2, \ldots, b_n$ denote a set of measurements.

Solving

$$\text{argmin}_x \sum_i (x - b_i)^2$$

computes the *mean*.

Solving

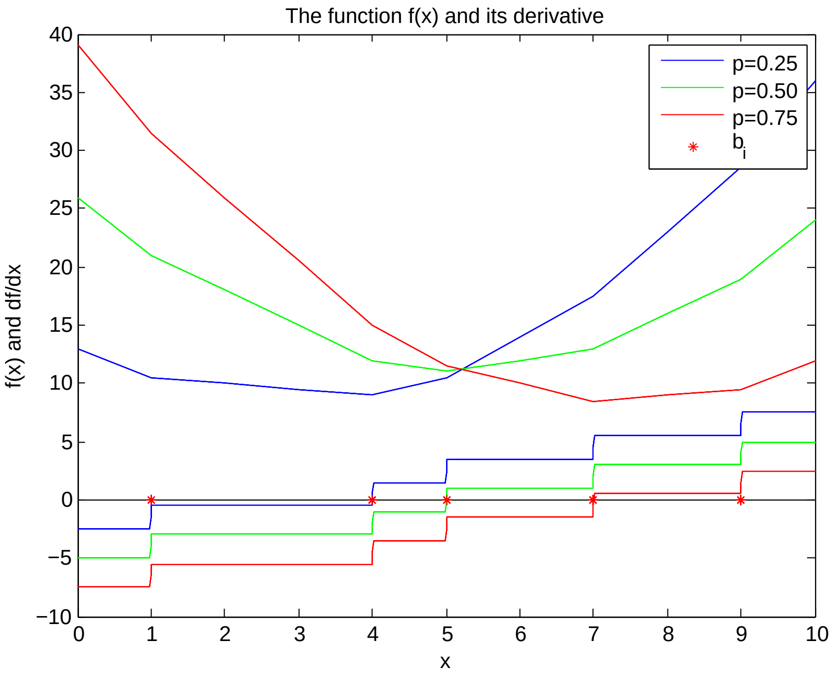
$$\text{argmin}_x \sum_i |x - b_i|$$

computes the *median*.

Medians correspond to *nonparametric statistics*. Nonparametric confidence intervals are given by percentiles. The p-th percentile is computed by solving the following optimization problem:

$$\text{argmin}_x \sum_i \left(|x - b_i| + (1 - 2p)(x - b_i) \right) .$$

Quantiles = Percentiles



Here we plot the function

$$f(x) = \sum_i \left(|x - b_i| + (1 - 2p)(x - b_i) \right) .$$

to be minimized and its derivative for three different values of p. The raw data are the b_i's. There are 5 of them plotted along the x-axis. Changing p causes the function $f'(x)$ to slide up or down thereby changing where it crosses zero.

Confidence Intervals For Medians

Assume that $B_1, B_2, B_3, \ldots, B_n$ are independent identically distributed with median m.

Let

$$B_{(1)} < B_{(2)} < B_{(3)} < \cdots < B_{(n)}$$

denote the *order statistics*, i.e., the original variables rearranged into increasing order.

Note: $B_{(k)}$ is the (k/n)-th *sample percentile*.

Then,

$$\begin{aligned}
\mathbb{P}(B_{(k)} \leq m \leq B_{(k+1)}) &= \mathbb{P}(B_j \leq m \text{ for } k \text{ indices and} \\
&\qquad B_j \geq m \text{ for the remaining } n-k \text{ indices}) \\
&= \binom{n}{k} \left(\frac{1}{2}\right)^n .
\end{aligned}$$

Hence,

$$\mathbb{P}(B_{(k)} \leq m \leq B_{(n-k+1)}) = \sum_{j=k}^{n-k} \binom{n}{j} \left(\frac{1}{2}\right)^n .$$

For any given n, it is easy to choose k so that $\sum_{j=k}^{n-k} \binom{n}{j} \left(\frac{1}{2}\right)^n \approx 0.95.$

Confidence Intervals For LAD Regression

Suppose we have n pairs of measurements (a_i, b_i), $i = 1, 2, \ldots, n$.
We posit that there is an affine relationship between the pairs:

$$b_i = x_1 + x_2 a_i + \varepsilon_i.$$

The ε_i's are independent, identically distributed, and have median zero.
We don't know the coefficients x_1 and x_2. We wish to find an estimator and an associated confidence "interval" for these two parameters.
Following our median example, the analogous optimization problem for this regression model is:

$$\min_{x_1, x_2} \sum_i \left(|x_1 + x_2 a_i - b_i| + (1 - 2p)\,(x_1 + x_2 a_i - b_i) \right).$$

It is easy to convert this problem into a linear programming problem:

$$\begin{array}{lrcll} \text{minimize} & \displaystyle\sum_i \left(\delta_i + (1 - 2p)\,(x_1 + x_2 a_i - b_i) \right) & & & \\ \text{subject to} & x_1 + x_2 a_i - b_i & \le & \delta_i & i = 1, \ldots, n \\ & -\delta_i & \le & x_1 + x_2 a_i - b_i & i = 1, \ldots, n. \end{array}$$

Using the simplex method, it is straight-forward to find the pair (x_1^*, x_2^*) that achieves the minimum for any given p, say $p = 1/2$.

Parametric Simplex Method

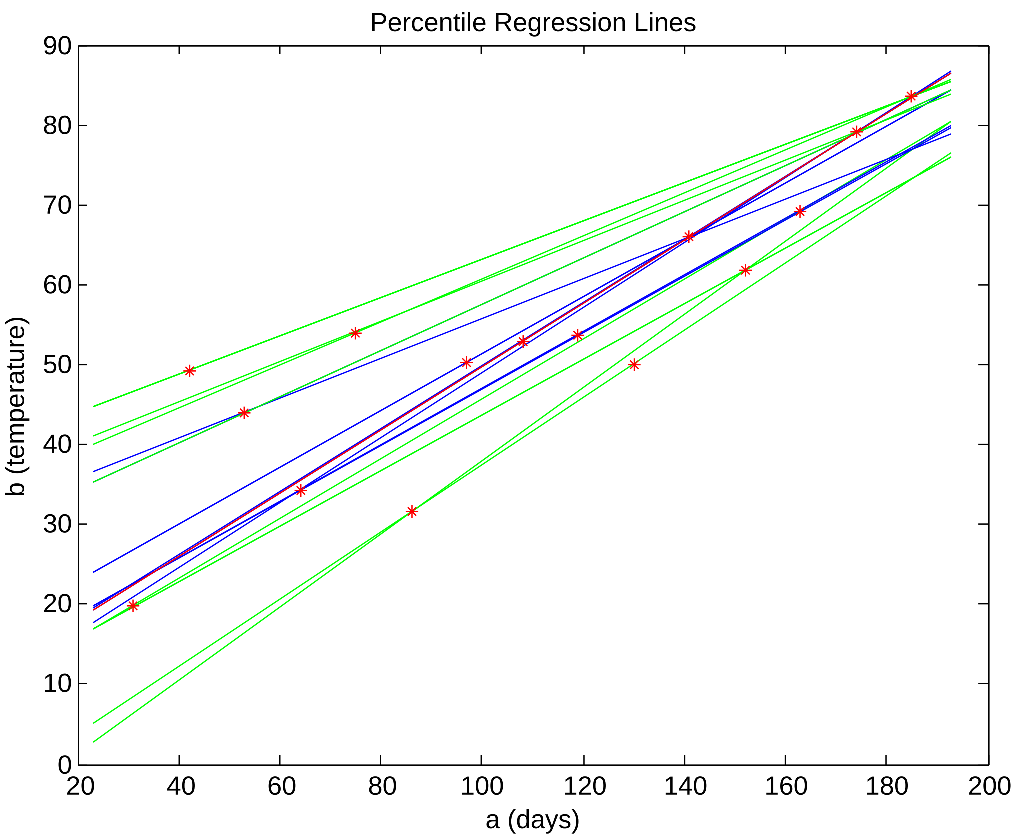
Better yet, using the parametric simplex method with p as the "parameter", one can solve this problem for every value of p in about the same time as the standard simplex method solves one instance of the problem.

Starting at $p = 1$ and sequentially pivoting toward $p = 0$, the parametric simplex method gives a set of thresholds $1 = p_0 \geq p_1 \geq p_2 \geq \cdots \geq p_K = 0$, at which the optimal solution changes.

In other words, over any interval, say $p \in [p_k, p_{k-1}]$, there is a certain fixed optimal solution, call it $(x_1^{(k)}, x_2^{(k)})$.

At the intersection of two intervals, say $[p_{k+1}, p_k]$ and $[p_k, p_{k-1}]$, both solutions $(x_1^{(k+1)}, x_2^{(k+1)})$ and $(x_1^{(k)}, x_2^{(k)})$ are optimal as are all convex combinations of these two solutions.

Quantile Regression Lines



Fifteen pairs of points, shown as red stars, and all of the regression lines associated with different intervals of p-values from $p = 1$ at the top to $p = 0$ at the bottom. The line associated with the interval that covers $p = 1/2$ is red and the lines within the confidence interval, computed using all p values between $p_{\min}$ and $p_{\max}$ are shown in blue.

Full 6D Regression Model

We can compute a confidence curve in $\mathbb{R}^6$ for the six regression coefficients in our local warming regression model.

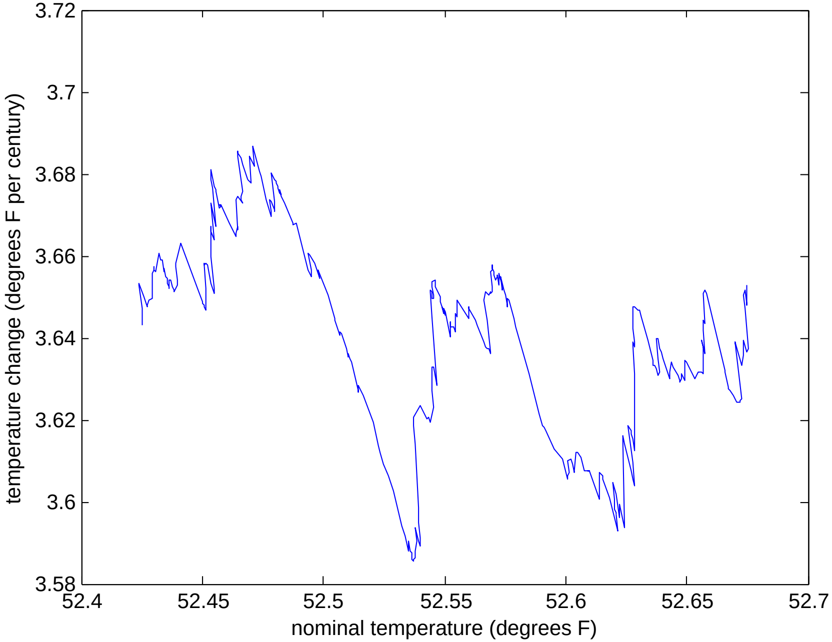
On the following pages we show a few 2-dimensional projections of this curve.

Any one-dimensional projection of the confidence curve defines a *confidence interval* for the associated quantity.

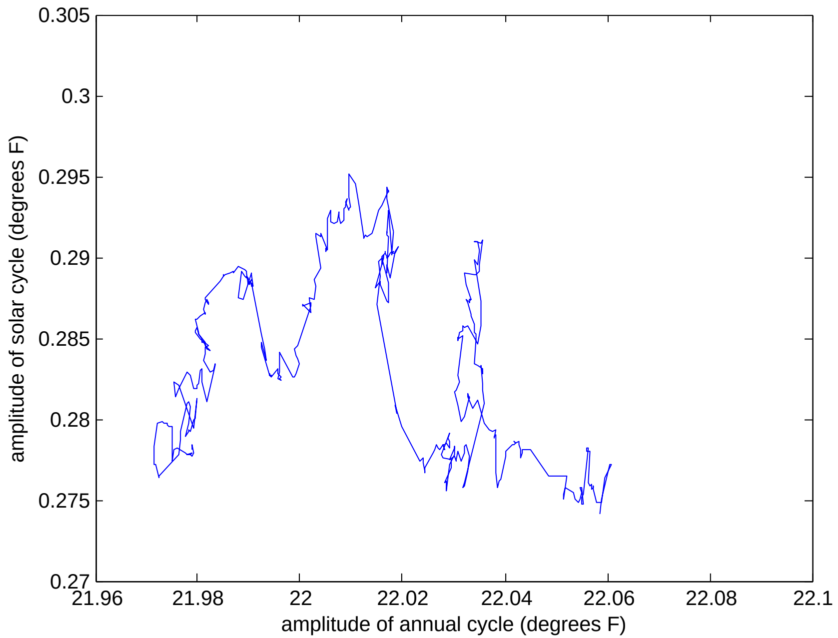
The 95% *confidence interval* for x_1 is $[3.588\ °\text{F},\ 3.687\ °\text{F}]/100$ yrs.

On the following page, the projection of the curve onto the vertical axis gives this interval. Note that the confidence interval is much wider than what one would deduce from looking just at the values associated with $p_{\min}$ and $p_{\max}$.

Confidence Curves



Plus/minus two-sigma confidence curve for the nominal temperature, x_0, and the rate of temperature change, x_1.



Plus/minus two-sigma confidence curve for the amplitude of the seasonal cycle, $\sqrt{x_2^2 + x_3^2}$, and the amplitude of the solar cycle, $\sqrt{x_4^2 + x_5^2}$.

Least Squares Solution (Mean instead of Median)

Suppose we change the objective to a sum of squares of deviations:

```
minimize sumdev: sum {d in DATES} dev[d]^2;
```

The resulting model is a *least squares model.*
The objective function is now convex and quadratic and the problem is still easy to solve.
The solution, however, is *sensitive* to outliers.
Here's the output:

$$
\begin{aligned}
x_0 &= 52.6 \text{ °F} \\
x_1 &= 1.2 \times 10^{-4} \text{ °F/day} \\
x_2 &= -20.3 \text{ °F} \\
x_3 &= -7.97 \text{ °F} \\
x_4 &= 0.275 \text{ °F} \\
x_5 &= 0.454 \text{ °F} \\
x_6 &= 0.730
\end{aligned}
$$

In this case, the rate of local warming is 4.37 °F per century.
However, the model produces the *wrong answer* for the period of the solar cycle.

Further Remarks

Close inspection of the output shows that:

- the January 22 is the coldest day in the winter,
- July 24 is nominally the hottest day of summer, and
- February 12, 2007, was the day of the last minimum in the 10.78 year solar cycle.

The coldest day in 2011 was January 23rd. It was -2 °F in the morning (very cold by NJ standards).

The ampl model and the shell scripts are available on my webpage.

Everyone is encouraged to grab data for any location they like.

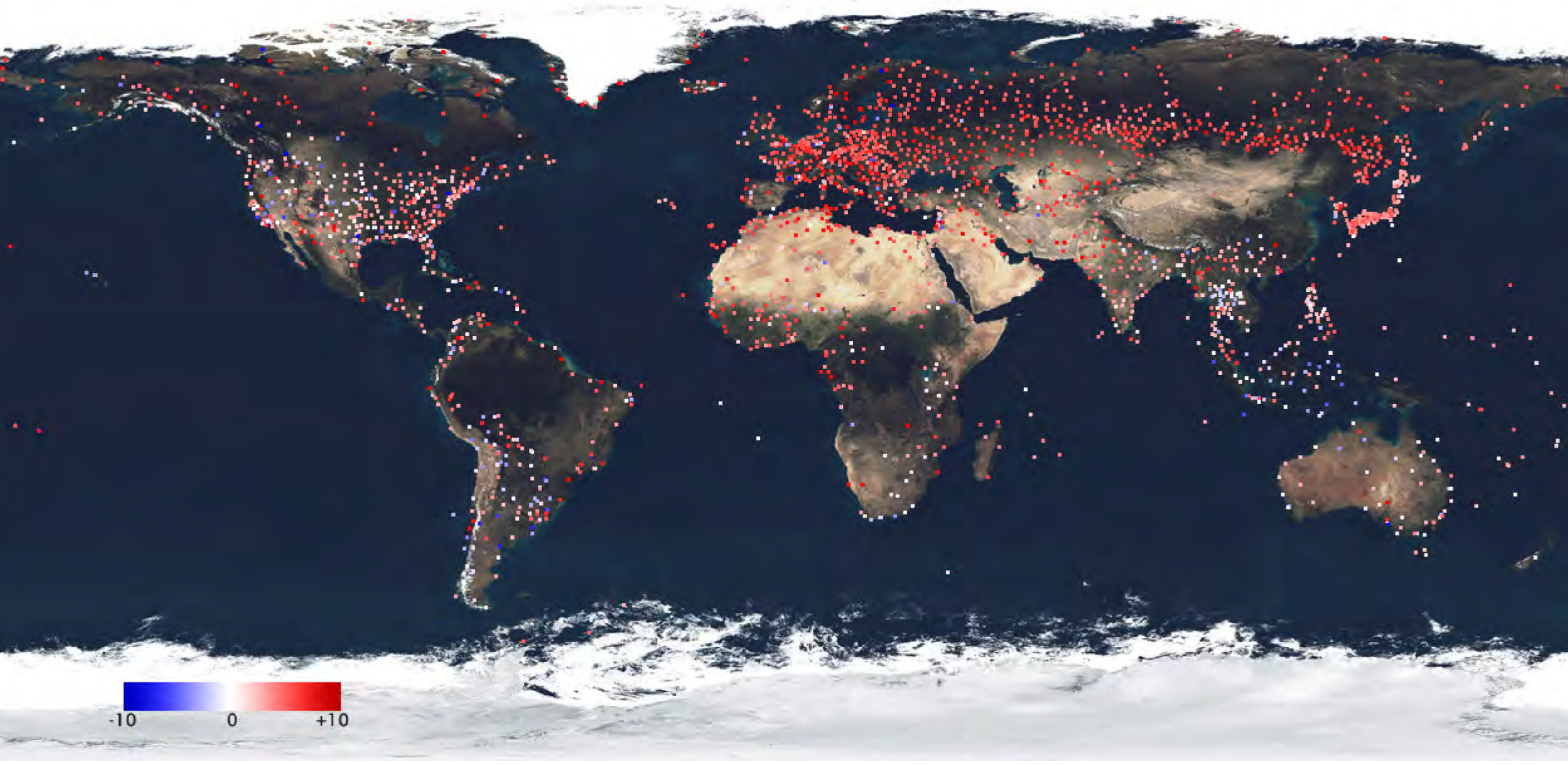
Send me the results and I'll compute a *global average*.

Repeat the Analysis Everywhere

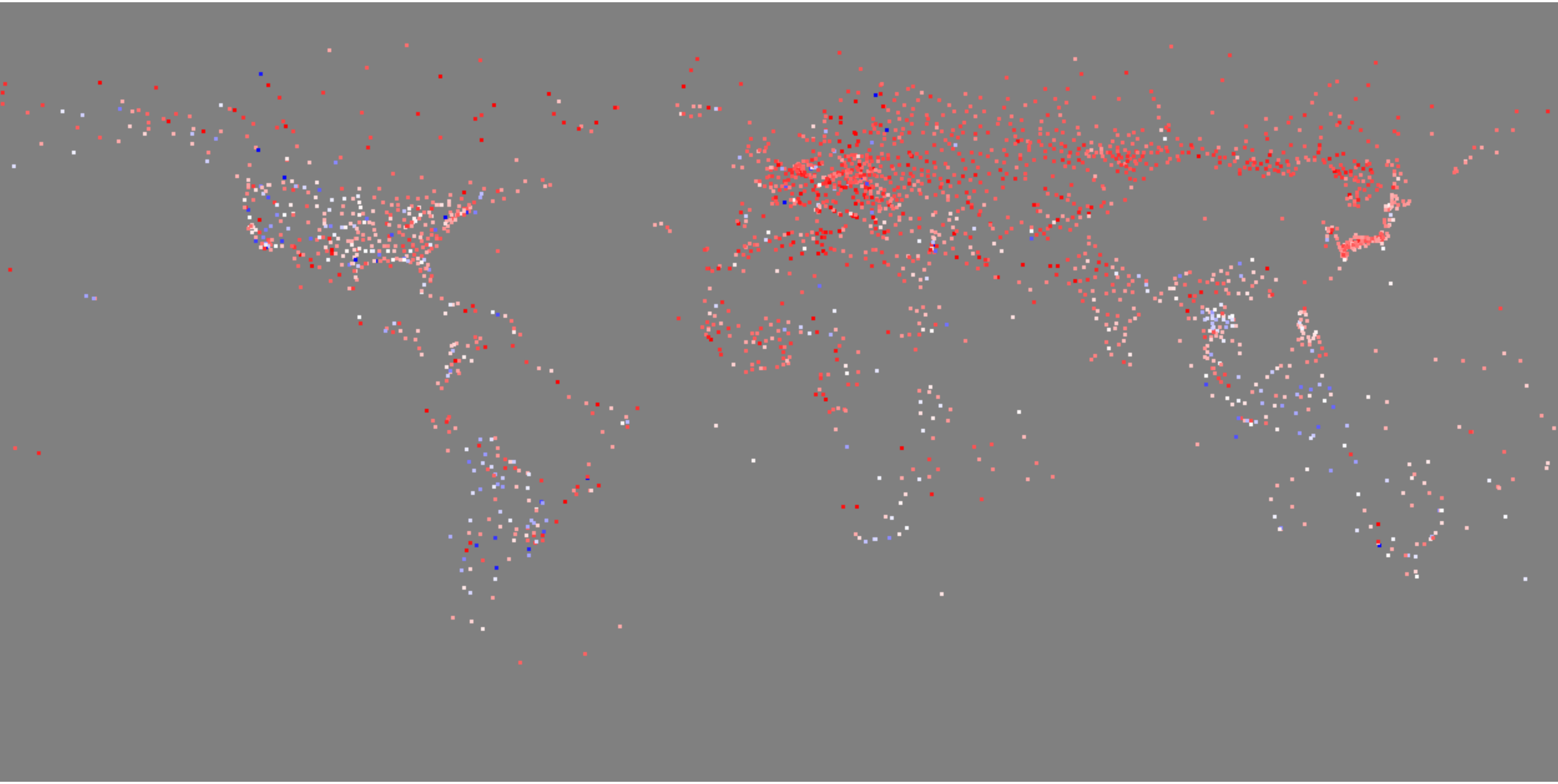
Criteria: Data collection commenced prior to Jan 1, 1955 and is currently in operation. There may be, and usually are, gaps in the data—the sight must have collected 3650 days of data (i.e., 10 years worth).

Caveats

- No attempt was made to filter out "bad data".
- Seasonal variations are not sinusoidal in the tropics.
- A site need not have been in continuous operation.
- No attempt has been made to purge anomolous data.



Mean value $= 4.18$ °F per century.
Median value $= 4.53$ °F per century.
Std Dev $= 2.94$ °F per century.



Mean value $= 4.18$ °F per century.
Median value $= 4.53$ °F per century.
Std Dev $= 2.94$ °F per century.