Is 2016 Really the Hottest Year on Record?

C. M. Wyss

January 24, 2017

Introduction

In this project, I aimed to reproduce four pieces of information gleaned from the NOAA site.

1. The annually averaged land temperature for 1901-2000 is 8.5°C, as in Global Analysis - Annual 2016 (https://www.ncdc.noaa.gov/sotc/global/201613).
2. 2016 was the hottest year on record, also in Global Analysis - Annual 2016.
3. The temperature increase for 2016 over the 20th century average was +1.43°C (also from Global Analysis - Annual 2016).
4. The following figure, from NOAA's Climate at a Glance page.

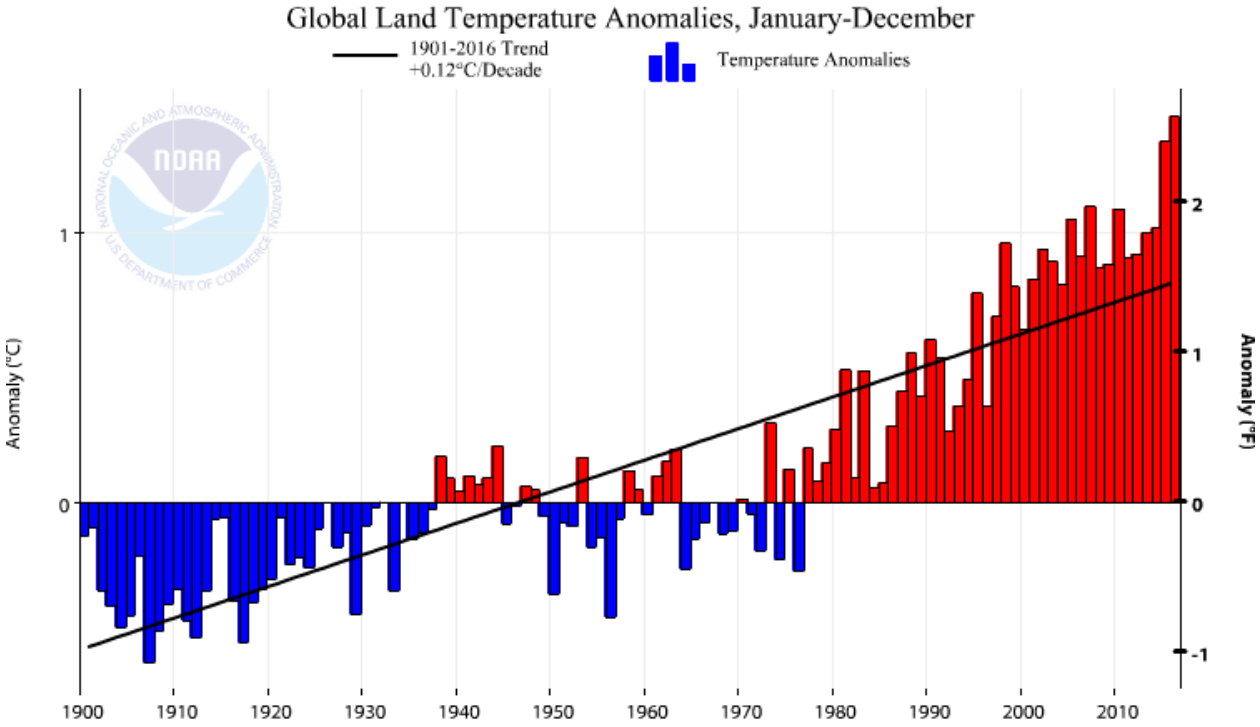


Figure 1: Global Land Temperature Anomalies, January-December

From NOAA's FTP site (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v3/) I was able to download the most recent Global Historical Climatology Network (GHCN) data, ghcnm.tavg.v3.3.0.20170122.qcu.dat (as of January 22, 2017). I chose the "quality control unadjusted" or "raw" dataset to work with, then performed the following cleaning steps.

- Where the QCF (quality control flag) indicated a value was erroneous, I set that value to NA.
- I removed rows where there were one or more missing values for the montly average temperature.

I saved the data as a csv file so I was able to reload it easily in future sessions.

```
setwd("~/Documents/DataScience/Projects/Weather/qcu/ghcnm.v3.3.0.20170122")
data <- read.table("tavg_munged.csv", header = TRUE, sep = ",")
data <- data[,c('ID', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY',
                'JUN', 'JUL', 'AUG', 'SEP', 'OCT', 'NOV', 'DEC')]
```

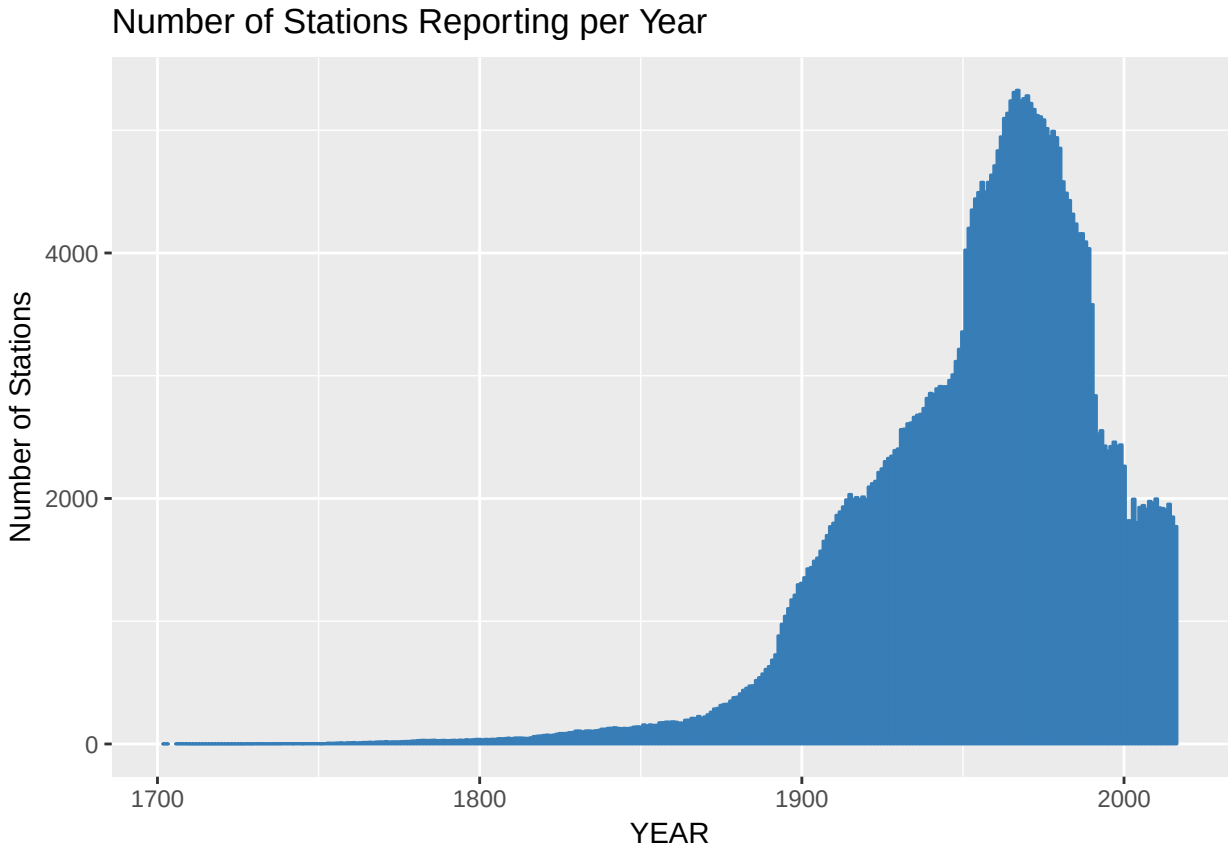
I ended up with 386,791 rows of data with a 14-column schema and no missing values. Columns are:

- ID - column 1, an 11 digit station identifier.
- YEAR - column 2, the 4 digit year of the observations.
- JAN through DEC - columns 3 through 14, each an observation of the respective monthly average temperature at that station, to four digits of significance, in degrees Celsius.

Sparse Data Before 1900

There were entries for the years from 1700 through 2016. The first graph I made was of number of observations (rows) for each year.

```
data %>%
  group_by(YEAR) %>%
  summarize(N = n()) %>%
  ggplot(aes(x = YEAR, y = N)) +
  geom_bar(stat = "identity", color = "#377eb8") +
  labs(title = "Number of Stations Reporting per Year", y = "Number of Stations")
```



There is a paucity of stations for the years prior to about 1900, so I removed observations for years earlier than 1901 since my main focus was the 20th century and beyond anyway.

```
data <- filter(data, YEAR > 1900)
```

I was left with 358,798 observations. (I used "$>$" instead of "$\geq$" because NOAA indicated that the years they took the average for are 1901-2000, not including 1900 itself.)

Anually Averaged Land Temperature for 1901-2000 is not 8.5°C

After ensuring a reasonably dense dataset, I computed the annually averaged land temperature for 1901-2000. First, I added a column MEAN to the table that contained the average of the monthly (averaged) temperatures for each observation. Then, the mean of this column for the years 1901 to 2000 (inclusive) is the value I sought. This value is the overall mean of (yearly) means of (monthly) means.

```
data <- mutate(data,
               MEAN = (JAN + FEB + MAR + APR + MAY + JUN +
                         JUL + AUG + SEP + OCT + NOV + DEC)/12)
data %>% filter((YEAR >= 1901) & (YEAR <= 2000)) %>%
  summarize(AVG = mean(MEAN), SD = sd(MEAN), N = n())
```

```
##        AVG      SD      N
## 1 12.51805 8.27343 328476
```

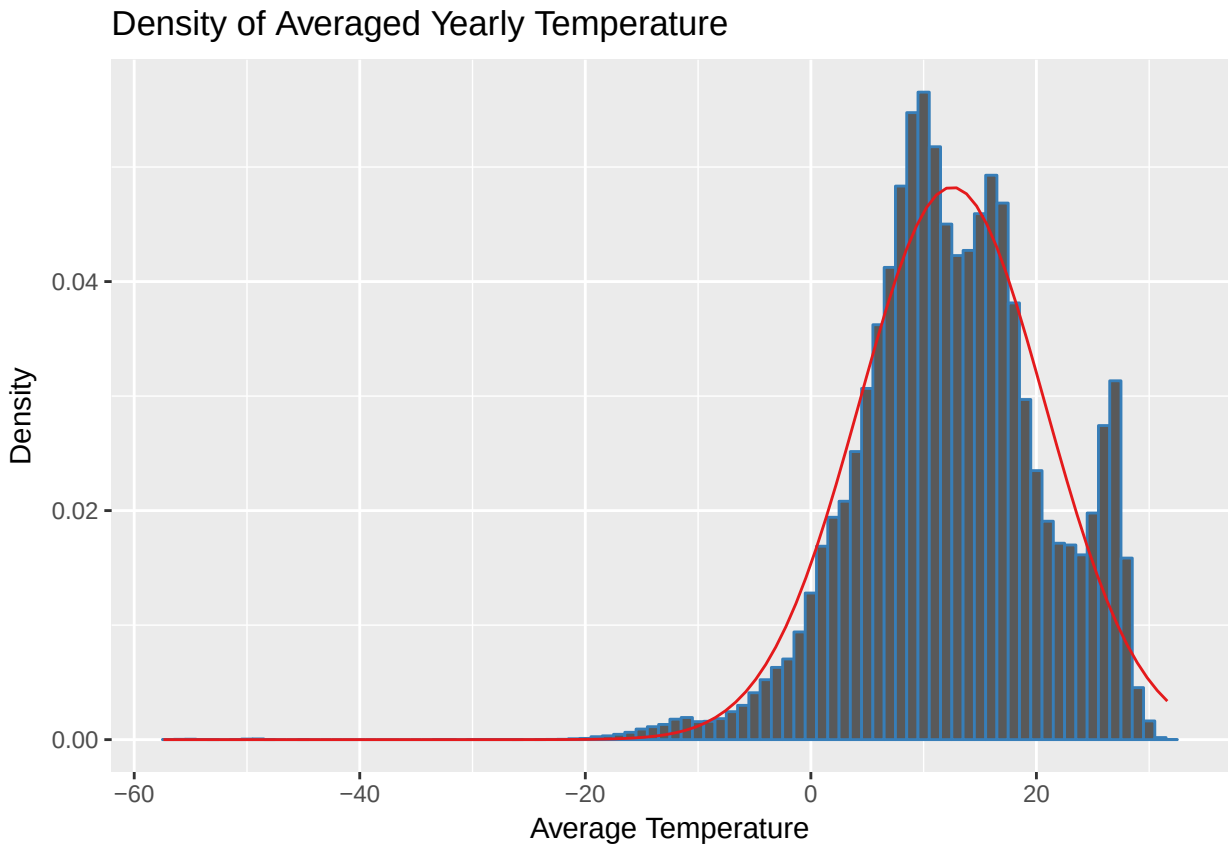
Next, I computed a 95% confidence interval for the annual averaged land temperature. I used the t-distribution since σ (the population standard deviation) is unknown.

$$\begin{aligned}\overline{x} \pm \left(t^*_{df} \times \frac{s}{\sqrt{n}}\right) &= 12.52 \pm \left(1.96 \times \frac{8.27}{\sqrt{328476}}\right) \\ &= 12.52 \pm (1.96 * 0.01) \\ &= (12.50, 12.54)\end{aligned}$$

This means I am 95% confident the annual averaged land temperature for 1901-2000 according to the GHCN dataset is between 12.50°C and 12.54°C. Unfortunately, this confidence interval does not contain the NOAA stated value of 8.5°C. I ran the corresponding hypothesis test with $H_0 : \mu = 8.5$ and $H_A : \mu \neq 8.5$ and obtained a p-value of 0.0000, meaning it is vanishingly unlikely that, with this data, the annual averaged land temperature is actually 8.5°C.

I did check the conditions for applying the theory (the Central Limit Theorem). First, independence. Any dependence between stations depends on whether they are close together geographically. I'll assume the stations are spread out since there are so few of them (at most 7280) located all over the world. Second, the distribution of means must be nearly normal. Is this the case? Here is a graph of the distribution with the corresponding normal distribution in red for comparison.

```
data %>% filter((YEAR >= 1901) & (YEAR <= 2000)) %>%
  ggplot(aes(x = MEAN)) +
  geom_histogram(binwidth = 1, color = "#377eb8", aes(y = ..density..)) +
  stat_function(fun = dnorm, args = list(mean = 12.52, sd = 8.27), color = "#e41a1c") +
  labs(title = "Density of Averaged Yearly Temperature",
       x = "Average Temperature",
       y = "Density")
```



The distribution is a bit skewed to the left and there is almost a second peak around what looks like 27°C but I believe the distribution is still nearly normal enough to perform the analysis.

2016 is not the Hottest Year on Record

Next, I computed the average temperature on a yearly basis, sorted the result from highest to lowest, and printed the first 10.

```
yrly_avgs <- data %>% group_by(YEAR) %>% summarize(AVG = mean(MEAN))
yrly_avgs <- yrly_avgs[order(yrly_avgs$AVG, decreasing = TRUE),]
head(yrly_avgs, 10)
```

```
## # A tibble: 10 × 2
##     YEAR      AVG
##    <int>    <dbl>
## 1   1998 14.24767
## 2   1991 13.91637
## 3   1990 13.87560
## 4   1999 13.82082
## 5   1995 13.74762
## 6   2016 13.68143
## 7   1994 13.66683
## 8   1953 13.47021
## 9   1997 13.38767
## 10  2012 13.34610
```

The hottest year on record in the GHCN dataset is 1998. 2016 is in sixth place, 0.57°C behind 1998.

The NOAA Temperature Increase for 2016 over the 20th Century Average is Correct and Statistically Significant

```
filter(data, YEAR == 2016) %>% summarize(AVG = mean(MEAN), SD = sd(MEAN), N = n())
```

```
##        AVG       SD    N
## 1 13.68143 8.745908 1773
```

Thus, the average temperature for 2016 was 13.68°C. How does this compare to my computed annual average of 12.52°C? To find out, I used the method of estimating the difference of means (with a 95% confidence interval). Let $\overline{x}_{2016} = 13.68$ and $\overline{x}_{20c} = 12.52$. Then a 95% confidence interval for the difference is:

$$\text{point estimate} \pm \text{margin of error}$$

$$(\overline{x}_{2016} - \overline{x}_{20c}) \pm \left(t^*_{df} \times \sqrt{\frac{s^2_{2016}}{n_{2016}} + \frac{s^2_{20c}}{n_{20c}}}\right)$$

where $n_{20c} = 328476$ and $n_{2016} = 1773$. The degrees of freedom for the t-distribution is $1773 - 1 = 1772$. It is easy to find t^*_{df} using R and it is basically the value for the normal distribution since the degrees of freedom is so high.

```
abs(qt(0.025, df = 1772))
```

```
## [1] 1.961304
```

Plugging in the numbers gives:

$$\begin{aligned}(\overline{x}_{2016} - \overline{x}_{20c}) \pm \left(t^*_{df} \times \sqrt{\frac{s^2_{2016}}{n_{2016}} + \frac{s^2_{20c}}{n_{20c}}}\right) &= (13.68 - 12.52) \pm \left(1.96 \times \sqrt{\frac{8.75^2}{1773} + \frac{8.27^2}{328476}}\right) \\ &= 1.16 \pm (1.96 \times 0.21) \\ &= 1.16 \pm 0.41 \\ &= (0.75, 1.57)\end{aligned}$$

Since the confidence interval does not contain 0, the difference in temperature is statistically significant. I am 95% confident that the difference between the average 2016 temperature and the annually averaged 20th century temperature is between 0.75°C and 1.57°C. Note that this CI contains the NOAA figure of +1.43°C for the increase.

NOAA's Climate at a Glance Figure Could not be Reproduced

Here is my attempt to reproduce the figure from NOAA's Climate at a Glance widget.

```
yrly_avgs %>% mutate(DIFF = AVG - 12.52) %>%
  filter(YEAR >= 1901) %>%
  ggplot(aes(x = YEAR, y = DIFF)) +
  geom_bar(stat = "identity", color = "#377eb8") +
  geom_smooth(method = "lm", se = TRUE, color = "#e41a1c") +
  labs(title = "Global Land Temperature Anomalies, Yearly",
       x = "Year",
       y = "Anomaly (Difference from 20th Century Avg, °C)")
```

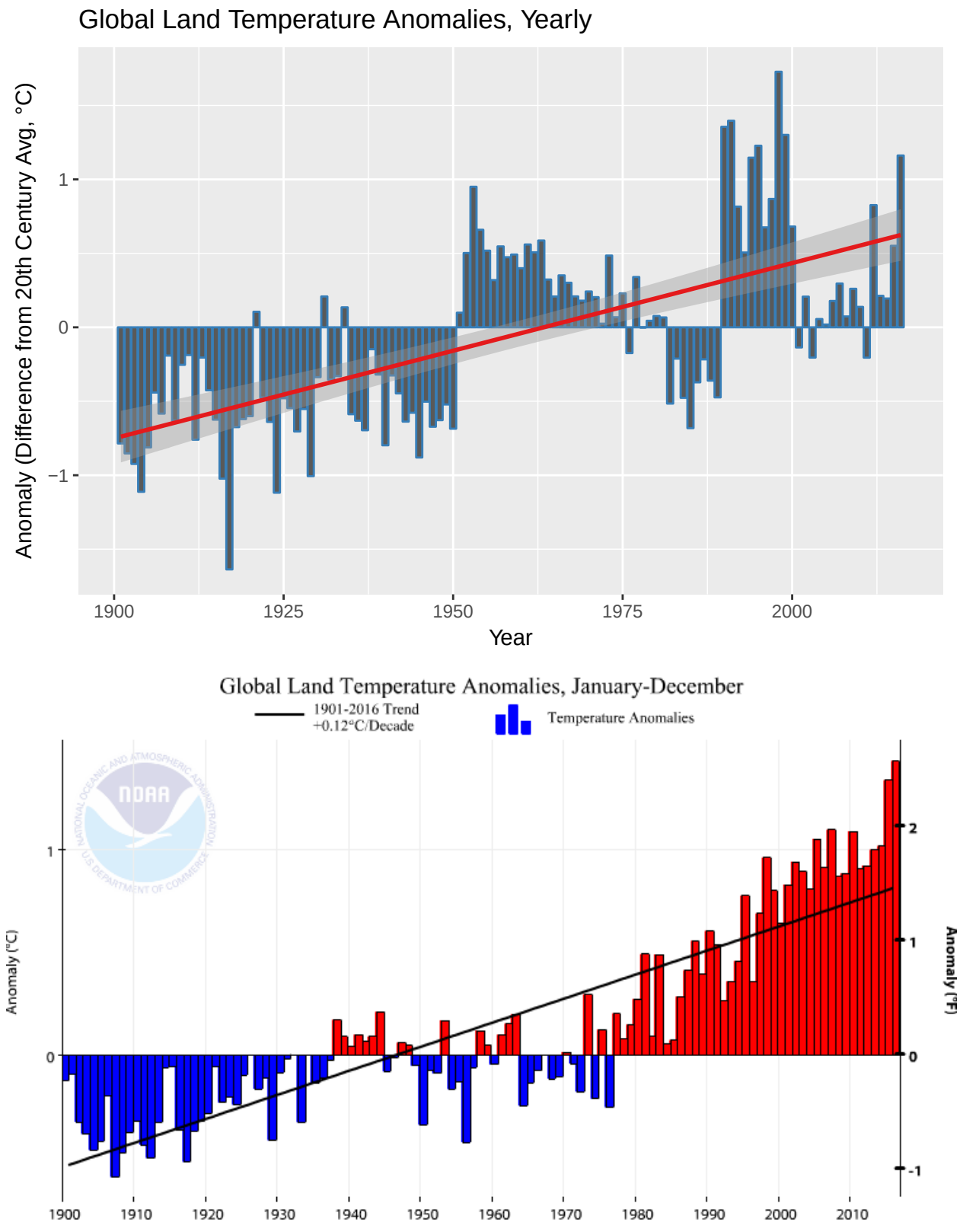


Figure 2: Global Land Temperature Anomalies, January-December

There are similarities in the two graphs.

- Both trend lines are increasing monotonically with time.
- Both graphs show it was colder during the first half of the 20th century than the last half.

However, there are also many differences.

- The NOAA graph shows cooler averages during the 50s, 60s, and early 70s. Mine does not show these, but instead shows a significant period of cooler averages in the late 70s/early 80s.
- The "pause" in warming during the early 2000s is clear in my graph but completely absent from the NOAA graph.
- In my graph, recent magnitudes are smaller than those during the 90s; in the NOAA data, temperature differences monotonically increase from about 1986 onward, throughout the early 2000s until the present.

Summary

I set out to reproduce the facts presented in the NOAA webpages and ended up mostly disappointed. Here is a summary of my findings.

1. I could not reproduce the claim that the annually averaged land temperature for 1901-2000 was 8.5°C, instead finding this temperature to be 12.52°C.
2. According to the data I downloaded, 2016 is *not* the hottest year on record, that honor goes to 1998.
3. I was able to reproduce the temperature increase for 2016 as +1.43°C since that value was within the 95% confidence interval I calculated.
4. I was not able to reproduce their graph of global land temperature anomalies (differences from the 20th century average), yet I was able to reproduce the overall warming trend.

Three out of four items were irreproducible using the available GHCN data. The actual dataset on which NOAA based their findings has been restructured from the raw GHCN data together with a sea-surface temperature (SST) dataset known as ICOADS. During the restructuring, filtering and interpolation (averaging) have been applied. The process is described in Smith-Reynolds 2005, available from the NOAA repository at https://www.ncdc.noaa.gov/monitoring-references/docs/smith-reynolds-2005.pdf. I looked at the 2005 paper (as well as its 2004 predecessor) and didn't see anything that should result in the large differences displayed in my figure versus NOAA's. The discrepancies remain a mystery.