

Weather or Not You Believe It

Chuck Cartledge

April 23, 2020

Contents

1	Introduction	1
2	Discussion	1
3	Exploration	3
3.1	How does temperature vary over a wide area over time?	3
3.2	How does temperature vary at a single location over time?	6
4	Conclusion	7
5	References	10
6	Files	10
A	Data from selected cities and locations	11

List of Tables

1	An arbitrary collection of cities and places of interest.	7
---	---	---

List of Figures

1	Number of USAF and WBAN sites by year.	2
2	Times between measures (log-log scale).	4
3	Times between measures (semi-log scale).	5
4	Weather stations reporting time.	5
5	Sample surface temperature map.	6
6	Norfolk, VA, USA weather station coverage.	7
7	Norfolk, VA, USA monthly average temperature.	8
8	Norfolk, VA, USA yearly average temperature.	8
9	Norfolk, VA, USA weather station reporting (log scales).	9
10	Norfolk, VA, USA weather station reporting (semi-log scales).	9
11	Austin, TX, USA weather station coverage.	12
12	Austin, TX, USA monthly average temperature.	12
13	Austin, TX, USA yearly average temperature.	13

14	Austin, TX, USA weather station reporting (log scales).	13
15	Austin, TX, USA weather station reporting (semi-log scales).	14
16	Fairbanks, AK, USA weather station coverage.	15
17	Fairbanks, AK, USA monthly average temperature.	15
18	Fairbanks, AK, USA yearly average temperature.	16
19	Fairbanks, AK, USA weather station reporting (log scales).	16
20	Fairbanks, AK, USA weather station reporting (semi-log scales).	17
21	Gladys, VA, USA weather station coverage.	18
22	Gladys, VA, USA monthly average temperature.	18
23	Gladys, VA, USA yearly average temperature.	19
24	Gladys, VA, USA weather station reporting (log scales).	19
25	Gladys, VA, USA weather station reporting (semi-log scales).	20

1 Introduction

Weather is everywhere. Sounds trite to say that, but it is true and a lot of real-time data is available for free, just for the downloading. We will look at semi-realtime weather data available from the National Oceanic and Atmospheric Administration (NOAA) collected and consolidated from over 25,000 uniquely identified United States Air Force (USAF), and 3,000 Weather-Bureau-Army-Navy (WBAN) weather stations world wide. Some weather stations have both USAF and WBAN identifier, while others may have only one. During this exploration, we will be creating “heat maps” of temperature for the states of Virginia, and North and South Carolina. These states were chosen because of local interest. The attached R script can be modified to display the same data for any collection of the US states.

2 Discussion

Weather data is available via FTP download from the NOAA domain at:

`ftp://ftp.ncdc.noaa.gov/pub/data/noaa/`

There are number of files available at this URL, and we will give a short description of the ones of interest for this exploration.

Weather station reports are collected starting from 1901 are organized by year, United States Air Force (USAF) number, and Weather-Bureau-Army-Navy (WBAN) numbers. But the coverage is not complete for all all years, nor for all possible combinations (see Figure 1). USAF and WBAN numbers can not be used as an indicator that the organization reponsible for the installation and maintenance of the weather station actually belongs to the USAF or any US government organization. The numbers should be viewed as data base keys to ensure uniqueness.

- **udpates.txt** – the last date when data (typically yearly) was updated.
- **isd-history.csv** – weather station identification (by USAF and WBAN), location (country, state, latitude, longitude, and ICAO), and the time frame when the weather station was active (beginning and ending dates). The location data in this file can be sued to get the USAF and WBAN identifiers of weather stations in a geographic area.
- **isd-inventory.csv** – number weather station reports per month per year.
- **yearly directories** – all collected data for each weather station. One weather station per gzip file¹ In 2017, there are approximately 13,500 stations reporting.

¹gzip is a file format used for file compression and decompression. gzip files can be created and manipulated via the **gzip** program part of many *nix operating system installations, or 7-Zip an addon for Windows operating systems.

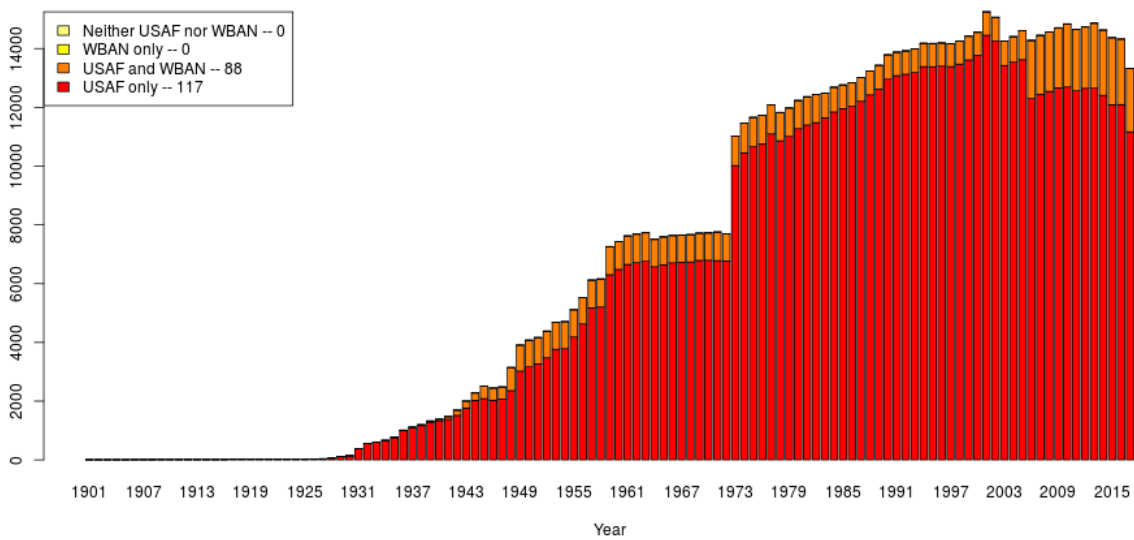


Figure 1: Number of USAF and WBAN sites by year. With two sources of data (USAF and WBAN), there are four possible combinations. From the data, it is apparent: (a) when air travel, and the organization that was to become the USAF started to become widespread (early 1930s), (b) when the USAF became separate from the Army (1947), (c) how the number of WBAN stations has remained relatively constant, while USAF stations has generally increased over time. The data show that USAF only designated weather stations cover 117 years. While there are data covering the range of years, not all years may have data. A weather station's identification may change over time. It could start as only a USAF station, and then later have a WBAN designation added, so for its entire life it would be classified as a USAF and WBAN station.

- **yearly directories/USAF-WBAN-YEAR.gz** – column formatted data, one record per measurement, from the beginning of the year until last update.

From an algorithmic perspective, the process is quite simple:

1. Use the `isd-history.csv` file to gather up the USAF and WBAN station ids of interest.
2. Decide which year to use.
3. Download the appropriate files of interest `yearly directories/USAF-WBAN-YEAR.gz`.
4. Parse each file IAW the ICD (see Section 6).
5. Remove invalid data from each file².
6. Evaluate and display the data.

About preserving data: The embedded R-script creates a series of directories based on the variable `downloadDirectory` whose value is controlled by the variable `persistData`. If `persistData` is set to `TRUE`, then data will be preserved in whatever directory is assigned to the variable `downloadDirectory`. If `persistData` is set to `FALSE`, then the program will run and when the current R session ends (as part of an RStudio session, or when an RScript command terminates), all data will be lost. Data preservation via `persistData` is all or nothing.

Weather station data records are variable length. Each record has these components[1]:

1. **Control data section** – information about the report including date, time, and station location information. Mandatory, 60 characters long.
2. **Mandatory data section** – meteorological information on the basic elements such as winds, visibility, and temperature. Mandatory, 45 characters long.
3. **Additional data section** – information of significance and/or which are received with varying degrees of frequency. Optional, 0 to 637 characters long.
4. **Remarks data** – plain language remarks are provided if they exist. Optional, 0 to 515 characters long.
5. **Element quality data section** – information on data that have been determined erroneous or suspect during quality control procedures. Optional, 0 to 1587 characters long.

The maximum data record size is 2,844 characters.

3 Exploration

3.1 How does temperature vary over a wide area over time?

Our interest and explorations led us to create an R script to interpret the weather data. The intent being to visualize how temperature changes over a reasonably sized area over time. The resulting R script (see Section 6) satisfied those initial needs, and opens the possibility of satisfying other needs.

The script looks at data from 2017, for the weather stations in Virginia, North Carolina, and South Carolina, ending at a specific time (2017-02-01 08:00:00 EST). This data selection was chosen because the

²FEDERAL CLIMATE COMPLEX DATA DOCUMENTATION FOR INTEGRATED SURFACE DATA details how invalid data fields are identified

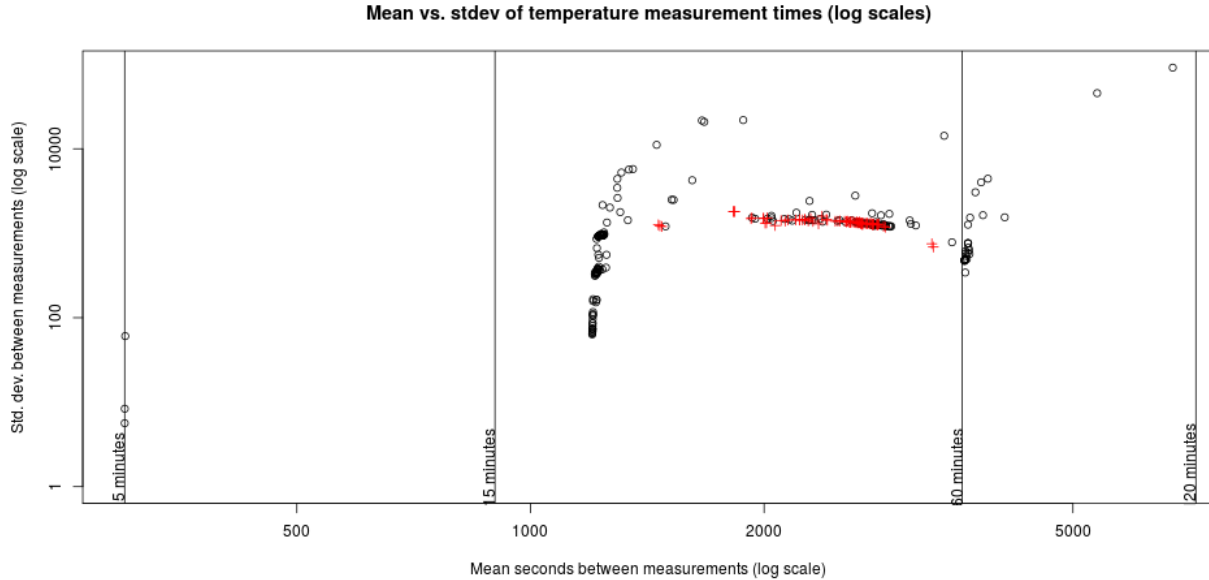


Figure 2: Times between measures (log-log scale). Raw data are plotted as black circles. Secondary processing of the data to remove outliers are shown as red crosses.

geographic area was of interest, and the time was far enough away from the beginning of the year so that simple boundary conditions were avoided. We focused our attention on the reported temperature, because of our interest. The approach is equally valid for other reported data as well.

The first area of interest was to attempt to “qualify” the reported data. The initial cut looked at determining how often data was reported, and how regular were the reports (see Figure 2). A log-log plot was chosen initially because the range of the data was unknown. The initial plot showed that there were a large number of delta times whose mean time and standard deviations were large. Secondary processing was performed on the raw time differences to remove those differences that were more than 25% or 75% away from the mean. The resulting data was plotted on the same log-log plot as red crosses. A significant number of red crosses were not plotable on the log-log plot because the standard deviation of the measurements was 0.

A similar approach was used to plot the data on a semi-log plot (see Figure 3). A significant number of red crosses with a standard deviation of 0 are now visible. From the figure, it is apparent that some stations report every 5 minutes, a reasonable number about every 25 minutes, a spread of stations in the 30 - 40 minute range, and a number around every hour. Based on the time between measurements, it makes the most sense to look at data on an hourly basis to get the greatest amount of current information. Requesting data any more frequently than that will generally result in stale data.

After achieving an understanding of how often data was updated, another question arose: when are the stations reporting their data (see Figure 4). If we attribute the “spikes” at 5 minute intervals as coming from those stations reporting every 5 minutes, it appears that most stations report on the hour, at a quarter past, 35 past (seems odd), and then at 55 past. The data almost looks as if there is an attempt at load leveling going on.

A surface temperature map is created at selected time intervals from the time of interest and going backwards in time. Each map is written to uniquely named file (see Figure 5) in an image directory. After all maps/images have been created, they are combined into a single GIF image showing how temperature

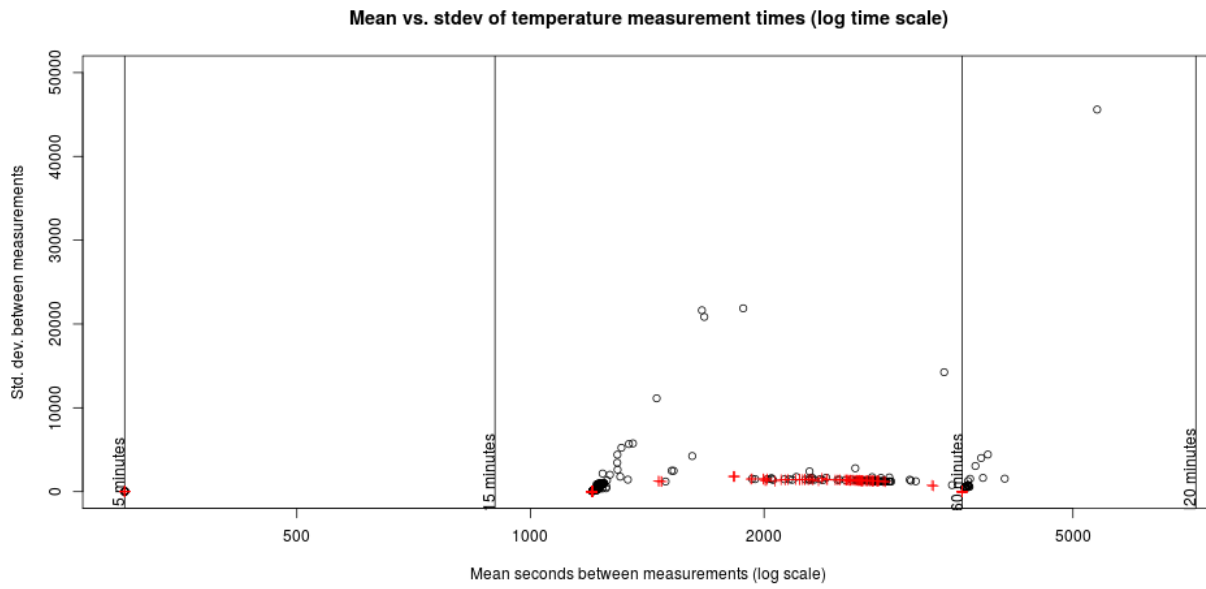


Figure 3: Times between measures (semi-log scale). Raw data are plotted as black circles. Secondary processing of the data to remove outliers are shown as red crosses.

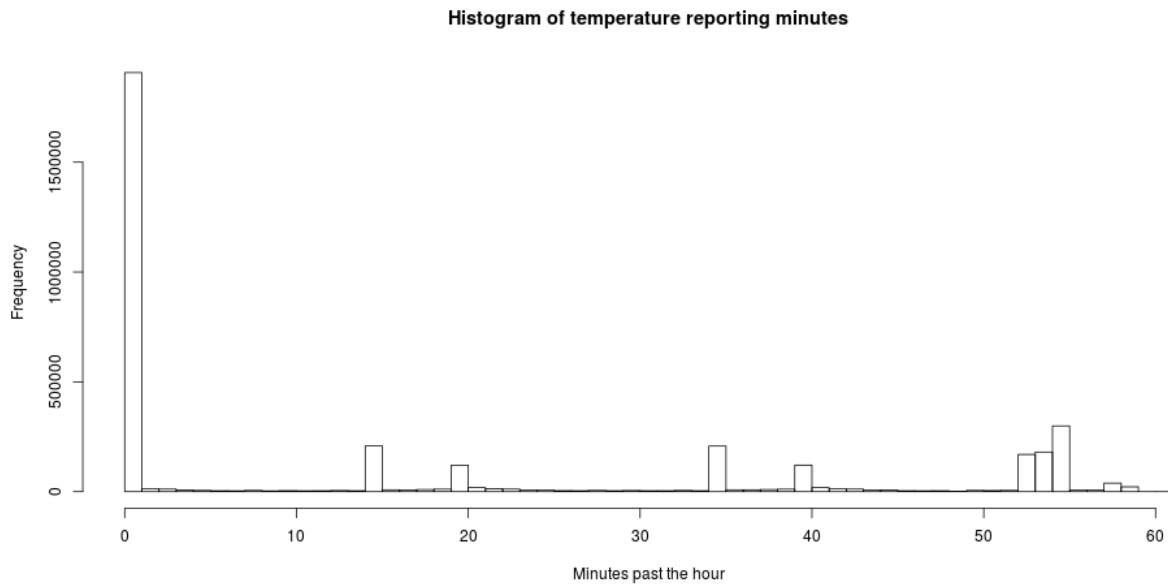


Figure 4: Weather stations reporting time.

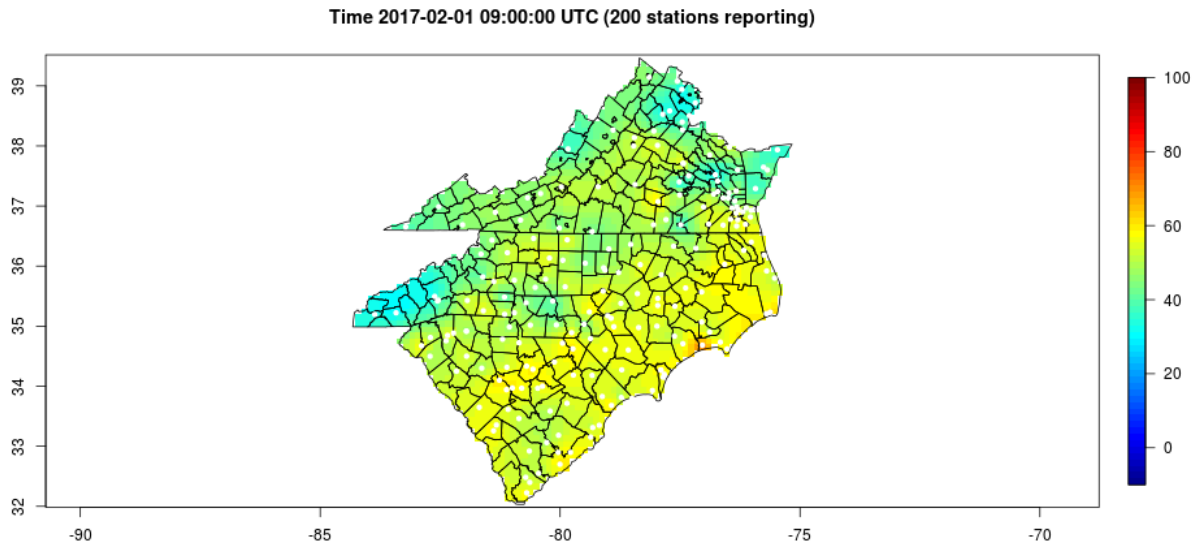


Figure 5: Sample surface temperature map.

changes over the region as a function of time (see Section 6).

3.2 How does temperature vary at a single location over time?

Our exploration into a how temperature changes over a large area as a function of time, led to the next logical question: How as temperature changed over time at a single location?

Algorithmically the question is easy to answer:

1. Choose a city/place of interest,
2. Convert the place of interest into a latitude and longitude. In the case of a city, estimate the center in latitude and longitude.
3. Identify the weather stations that are closest to the the computed latitude and longitude.
4. Select the weather station with the longest history of data.
5. Report the results.

A collection of cities and places is provided in the `citiesOfInterest.csv` file (see Table 1). The file is included in this report (see Section 6). `citiesOfInterest.csv` is a comma separated value file that can be modified (new places added, current ones removed) as desired. The `weather02.R` script maintains a local database of names and places based on the contents of the `citiesOfInterest.csv` file. If there is a new entry in the file, the Google GEOCODE API³ is queried to get latitude and longitude data, which is added to the location database.

Data from the weather station with the longest set of data is reported in a number of different ways. These include:

³<http://maps.googleapis.com/maps/api/geocode/>

Table 1: An arbitrary collection of cities and places of interest.

city	State	Country
Austin	TX	US
CAIRO		EG
Cairo International Airport		EG
Fairbanks	AK	USA
Gladys	VA	USA
Norfolk	VA	USA

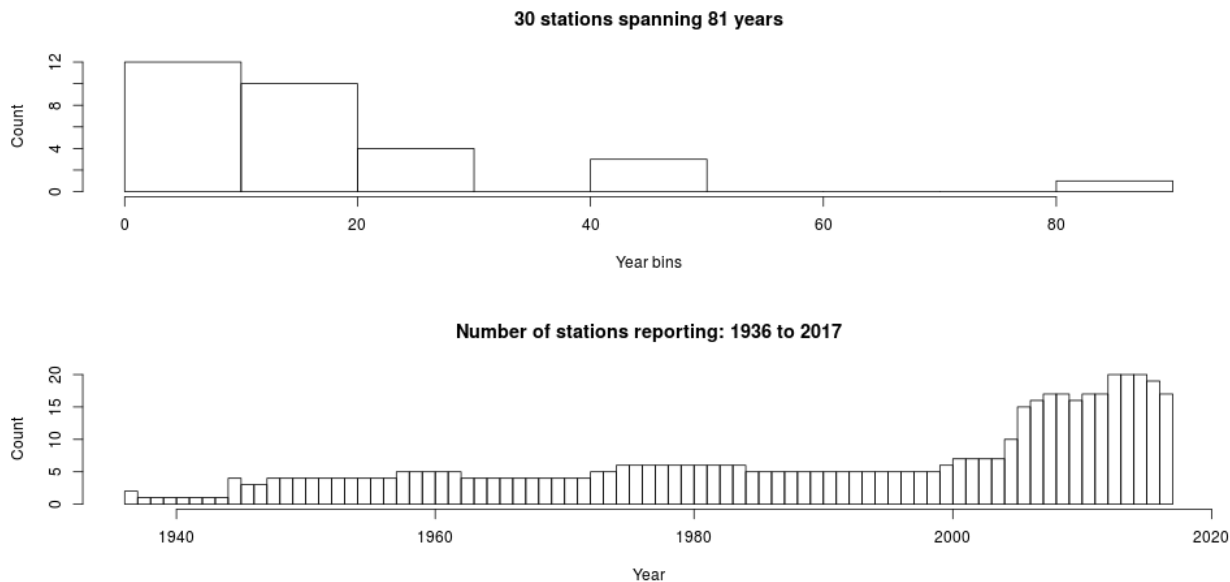


Figure 6: Norfolk, VA, USA weather station coverage.

- Histogram showing number of reporting stations based on years of reporting (see Figure 6),
- Bar plot showing number of reporting stations by year (see Figure 6),
- Monthly mean and standard deviation temperatures for the previous year (this assumes that the current year doesn't have 12 months of data) (see Figure 8),
- Yearly mean and standard deviation temperatures for all reported years (see Figure 8), and
- Mean and standard deviation of reporting times in log-log and semi-log formats (see Figures 9, and 10).

4 Conclusion

Weather data is freely available for many areas of the world via FTP from the NOAA site. Weather station data is fixed format character based, variable record length. Some weather stations report very frequently

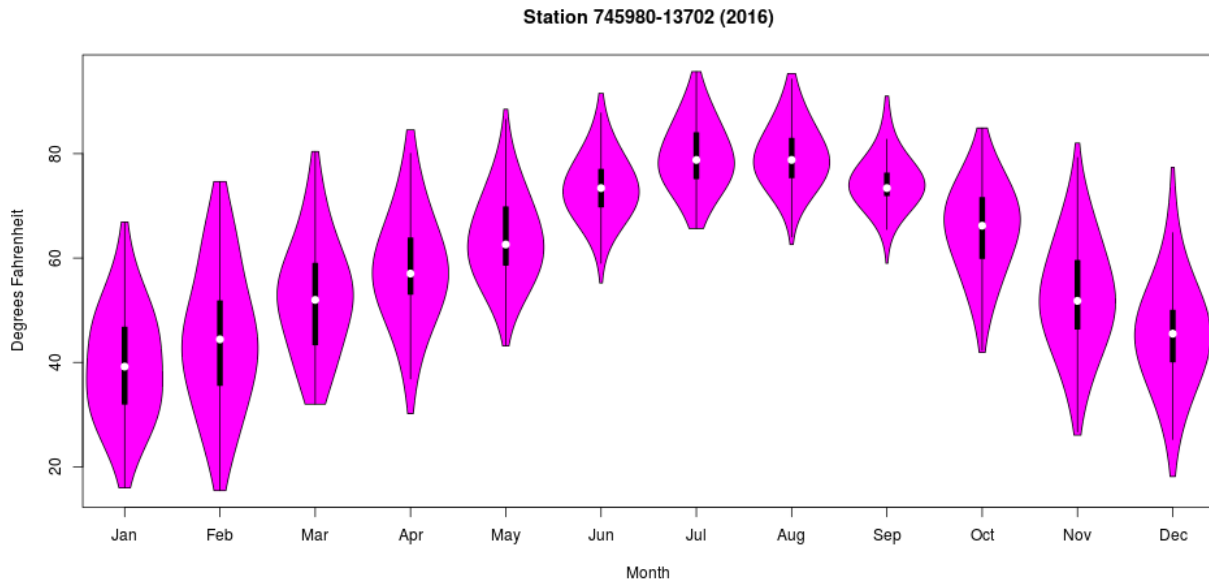


Figure 7: Norfolk, VA, USA monthly average temperature.

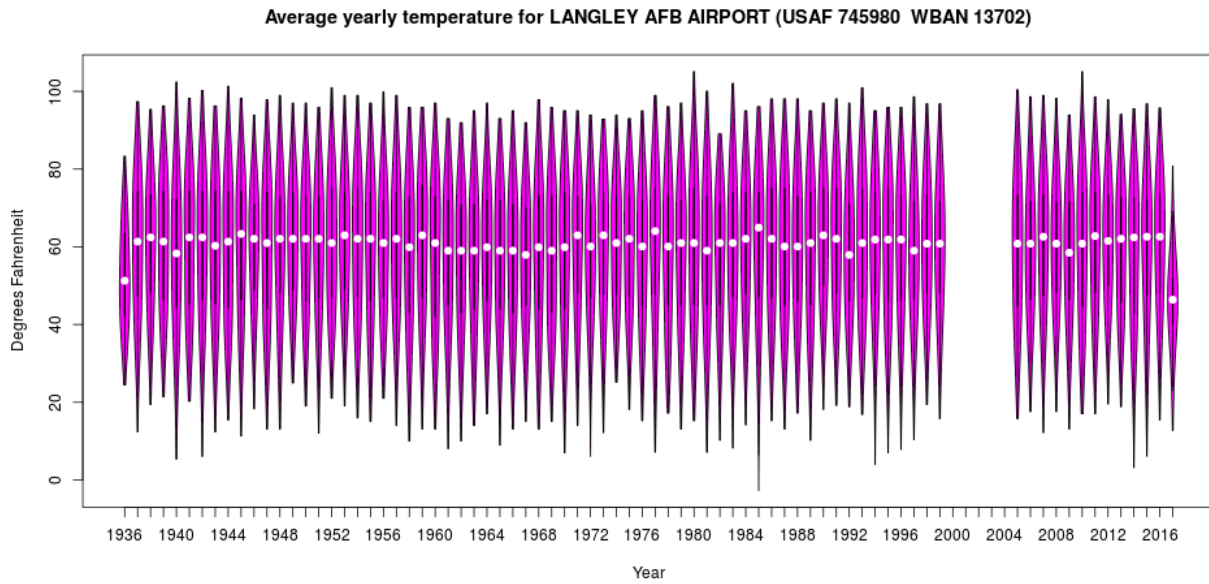


Figure 8: Norfolk, VA, USA yearly average temperature.

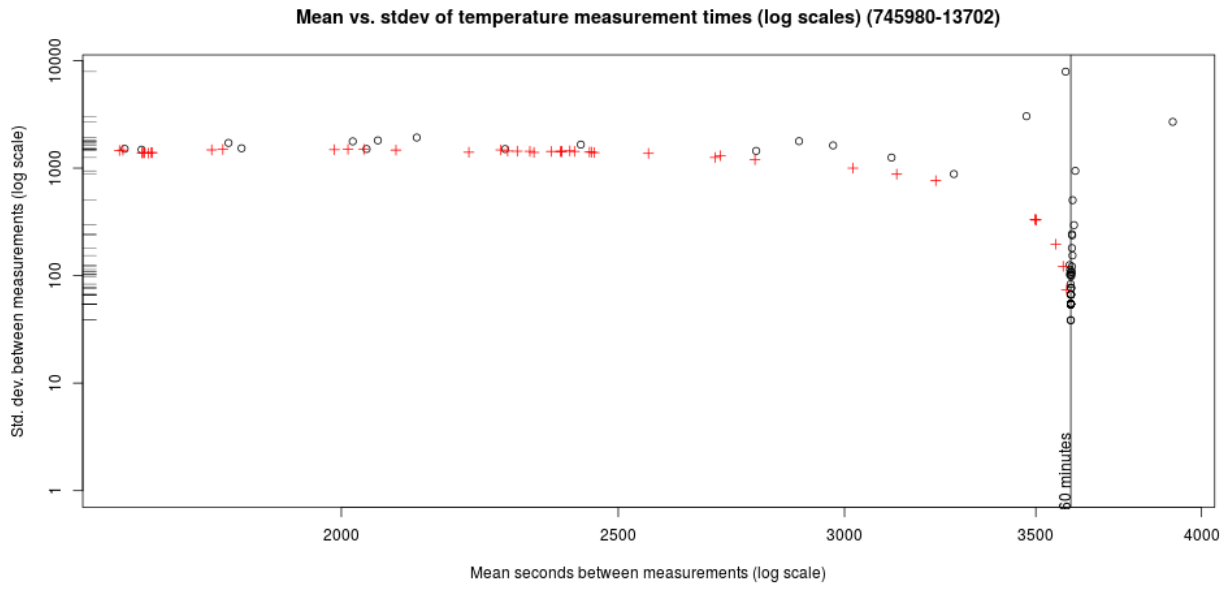


Figure 9: Norfolk, VA, USA weather station reporting (log scales).

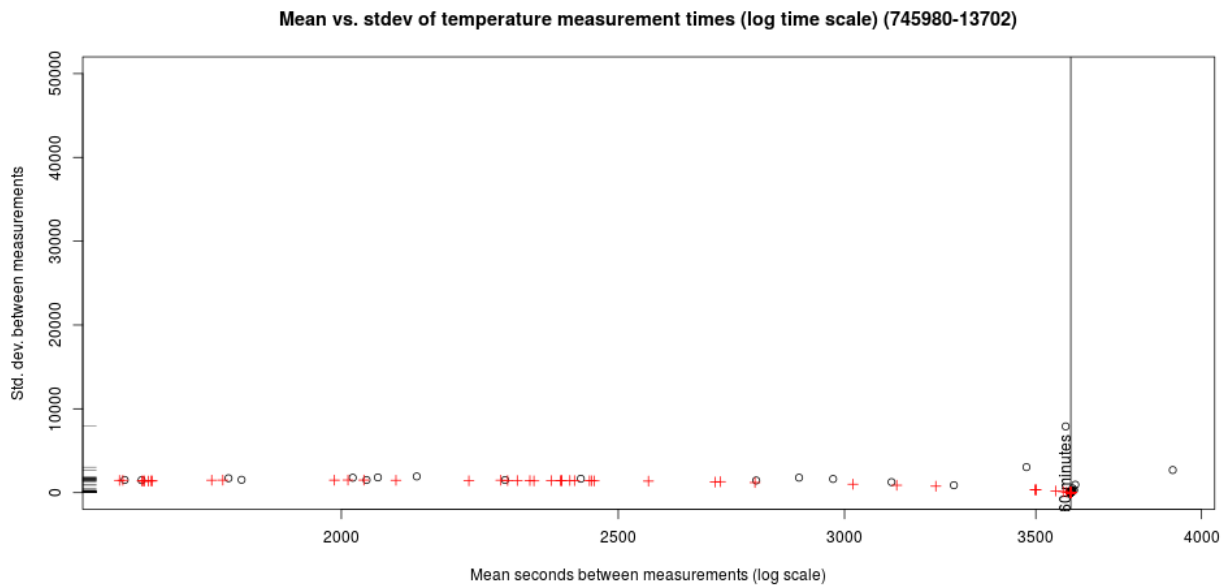


Figure 10: Norfolk, VA, USA weather station reporting (semi-log scales).

(every 5 minutes), while others report less frequently (once an hour). Data can be used to create animated maps and overlays showing data as a function of time.

A century of free weather data has the potential to answer lots of questions, including things like: has the average temperature changed over time?





5 References

References

[1] NCDC, *Federal Climate Complex Data Documentation for Integrated Surface*, (2016).

6 Files

A collection of miscellaneous files mentioned in the report.

- citiesOfInterest.csv – A CSV file collection of cities and places that are of personal interest. 
- weather02.R – The R script used to download data, purge invalid data, evaluate and plot remaining data. This is the program used to create the images in this report. 
- FEDERAL CLIMATE COMPLEX DATA DOCUMENTATION FOR INTEGRATED SURFACE DATA – The specification for parsing and interpreting weather station data. 
- weather.gif – temperature as a function of time across VA, NC, and SC. 

A Data from selected cities and locations

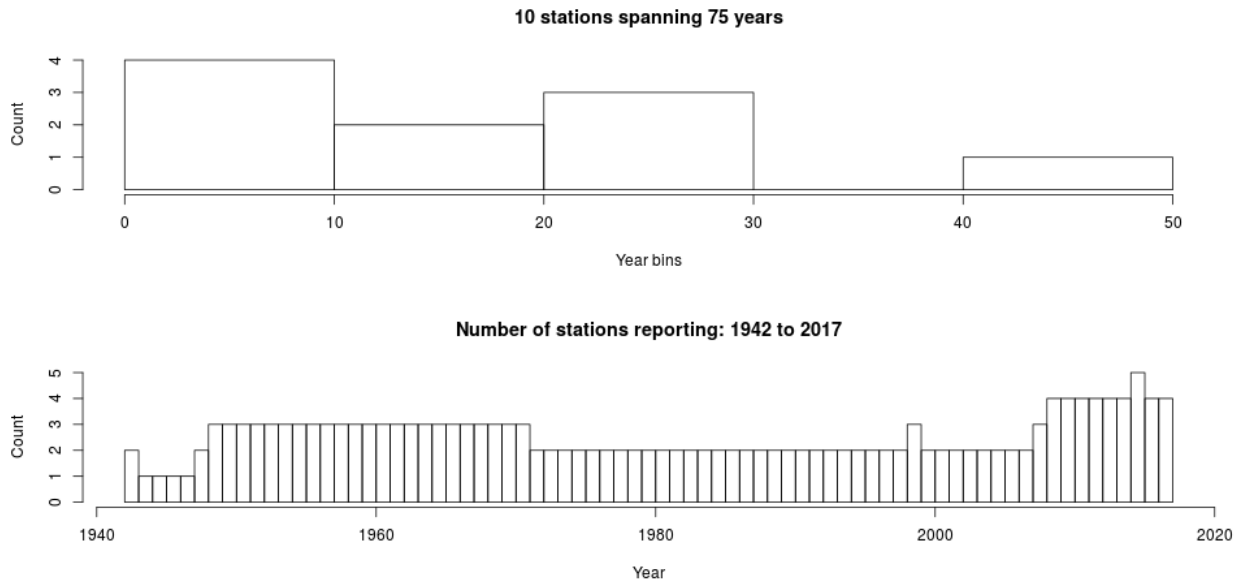


Figure 11: Austin, TX, USA weather station coverage.

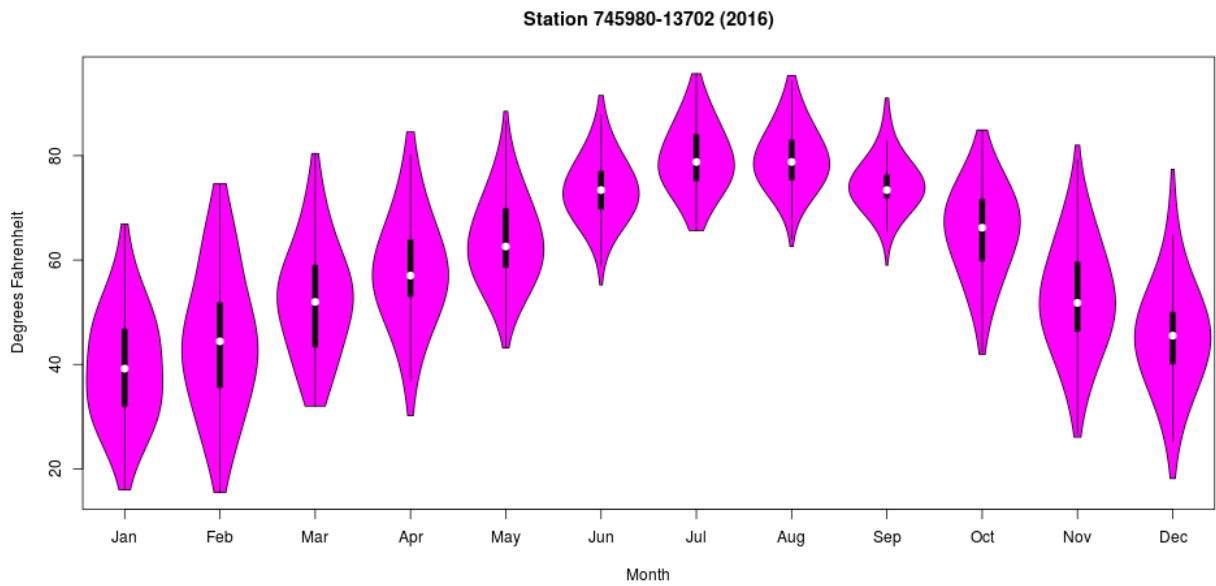


Figure 12: Austin, TX, USA monthly average temperature.

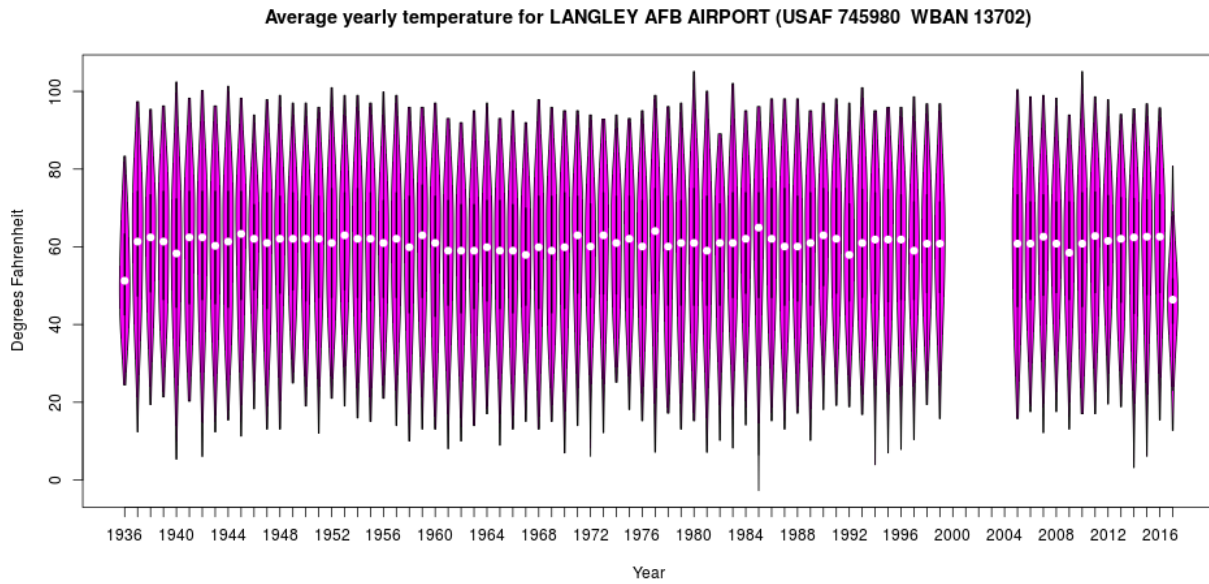


Figure 13: Austin, TX, USA yearly average temperature.

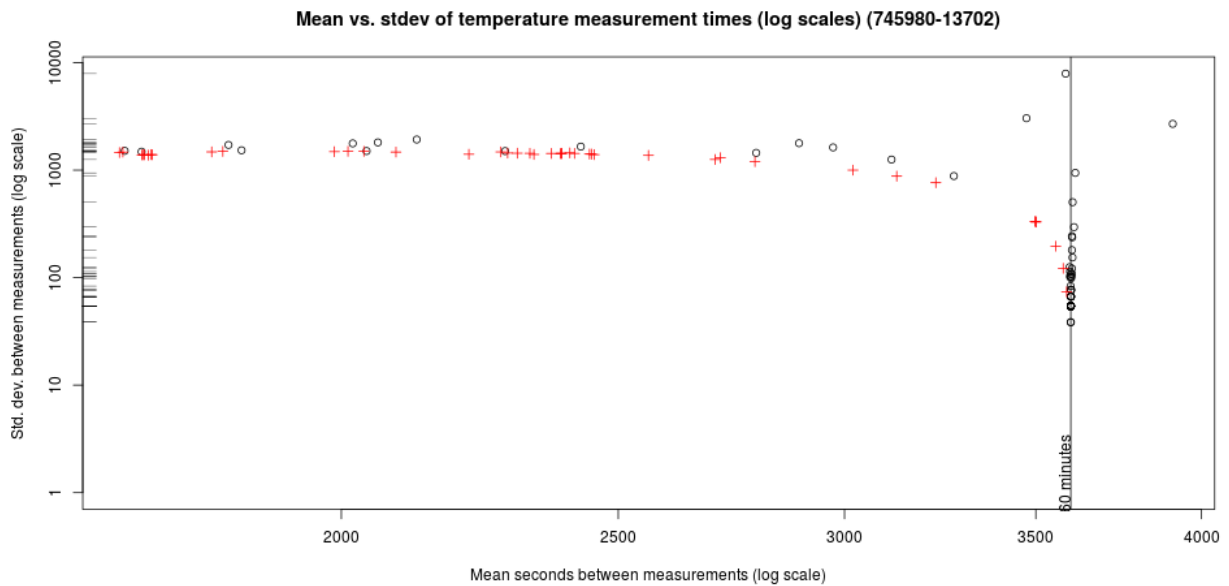


Figure 14: Austin, TX, USA weather station reporting (log scales).

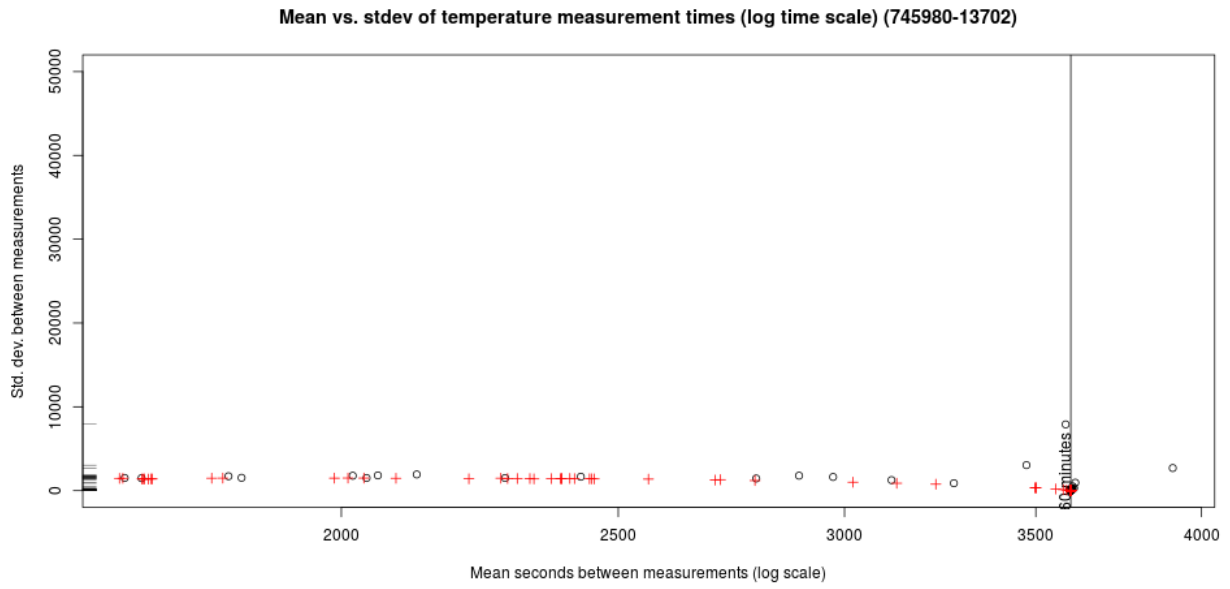


Figure 15: Austin, TX, USA weather station reporting (semi-log scales).

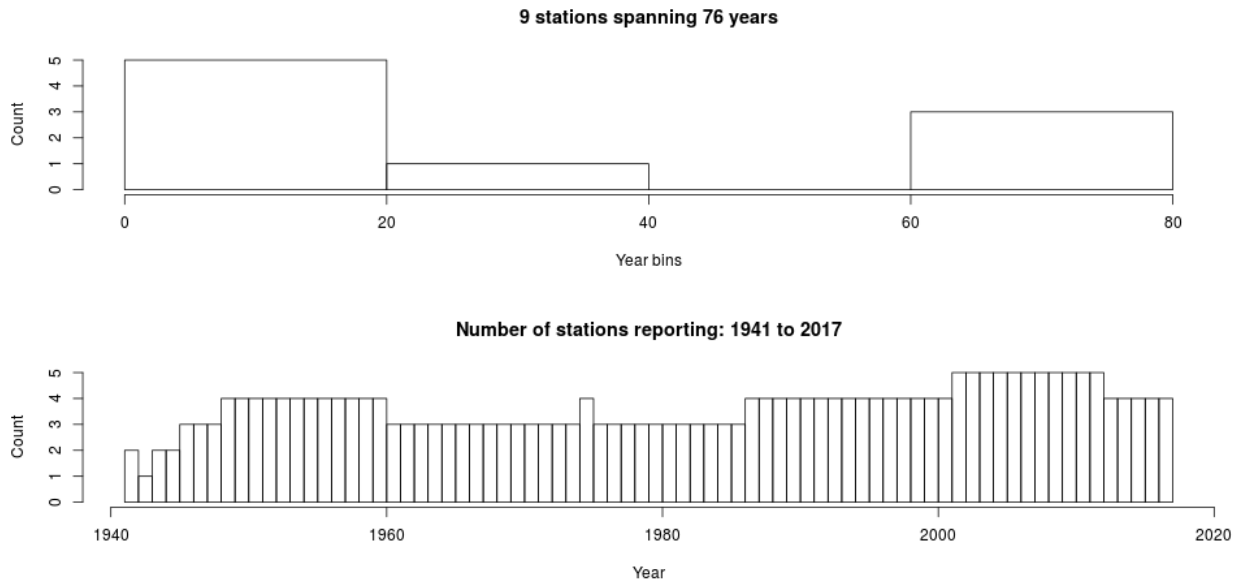


Figure 16: Fairbanks, AK, USA weather station coverage.

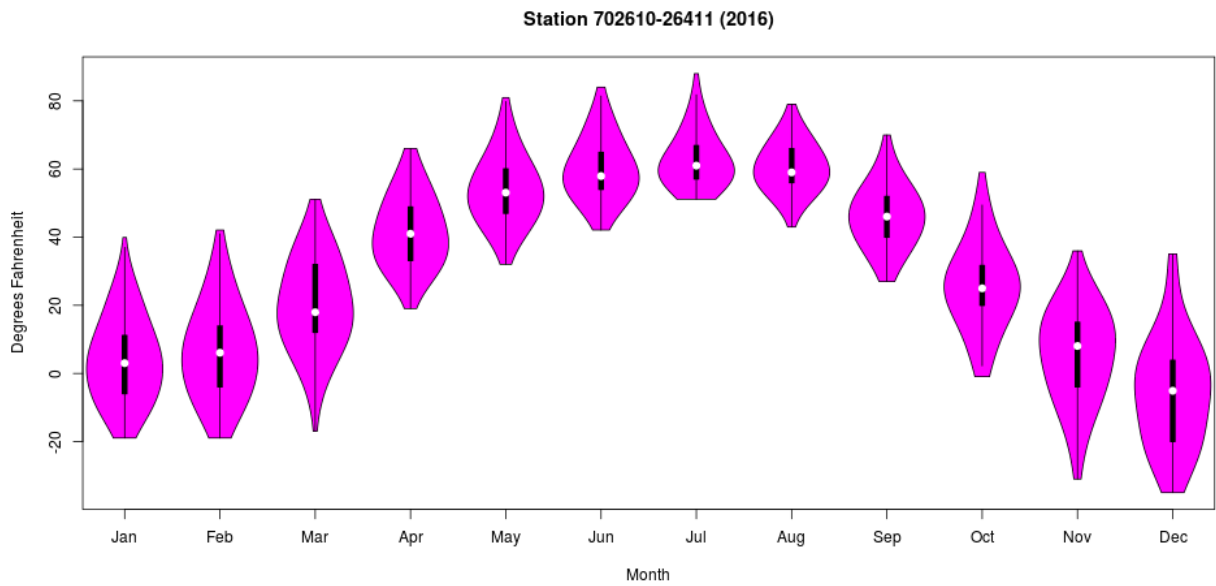


Figure 17: Fairbanks, AK, USA monthly average temperature.

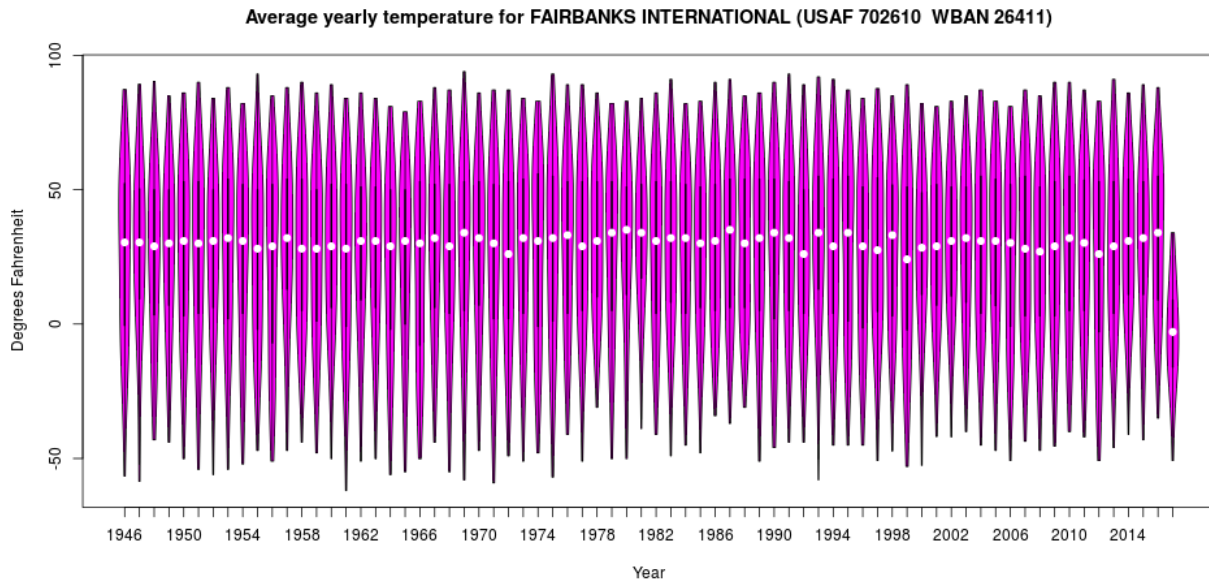


Figure 18: Fairbanks, AK, USA yearly average temperature.

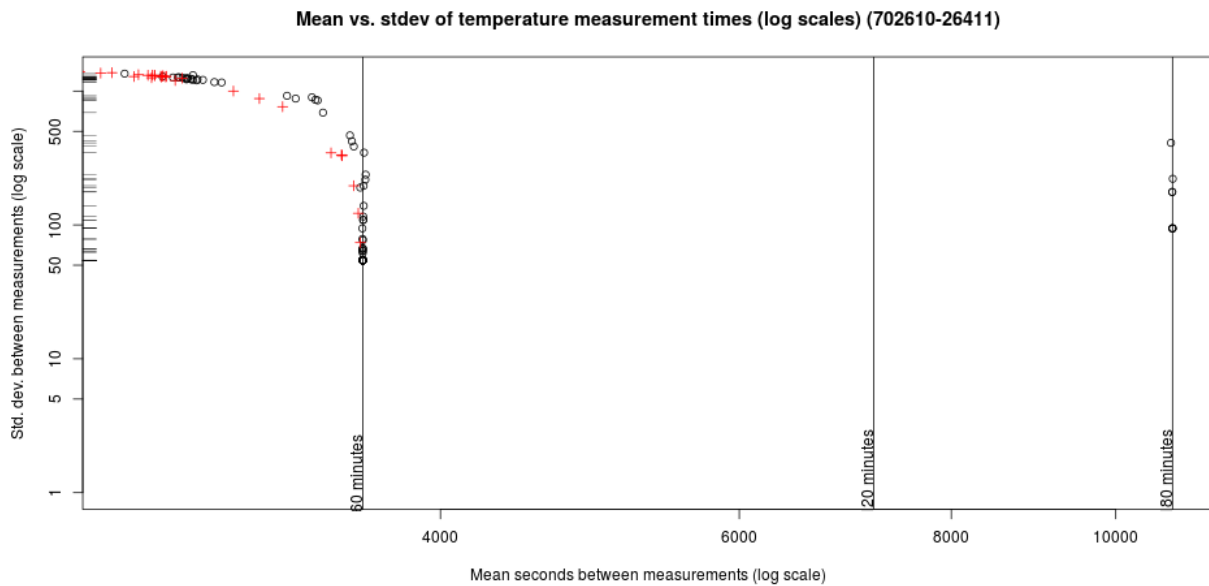


Figure 19: Fairbanks, AK, USA weather station reporting (log scales).

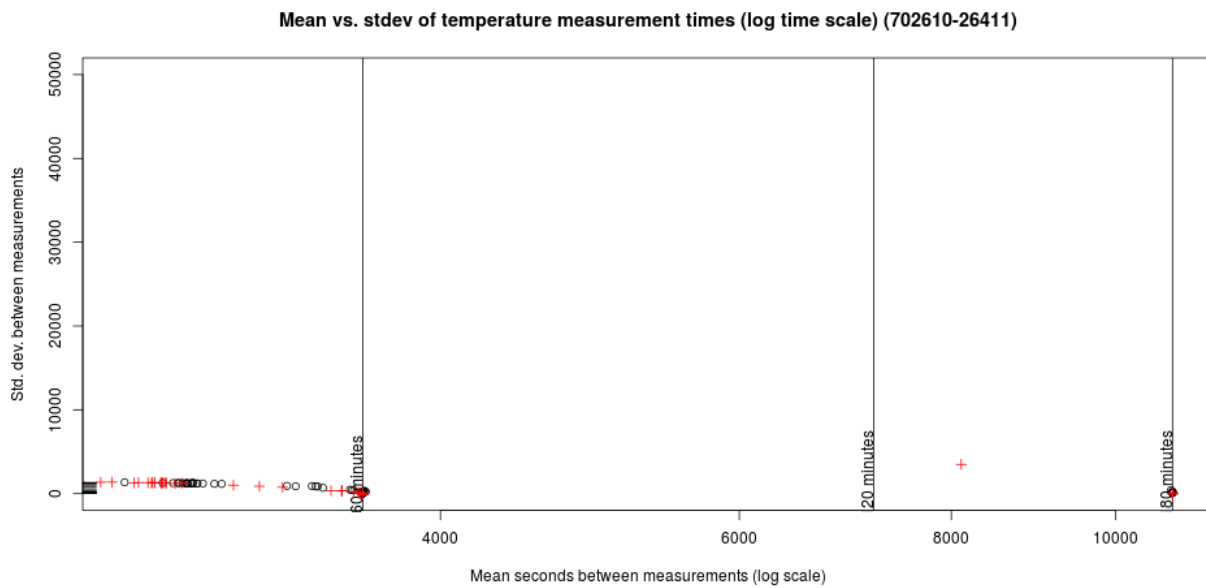


Figure 20: Fairbanks, AK, USA weather station reporting (semi-log scales).

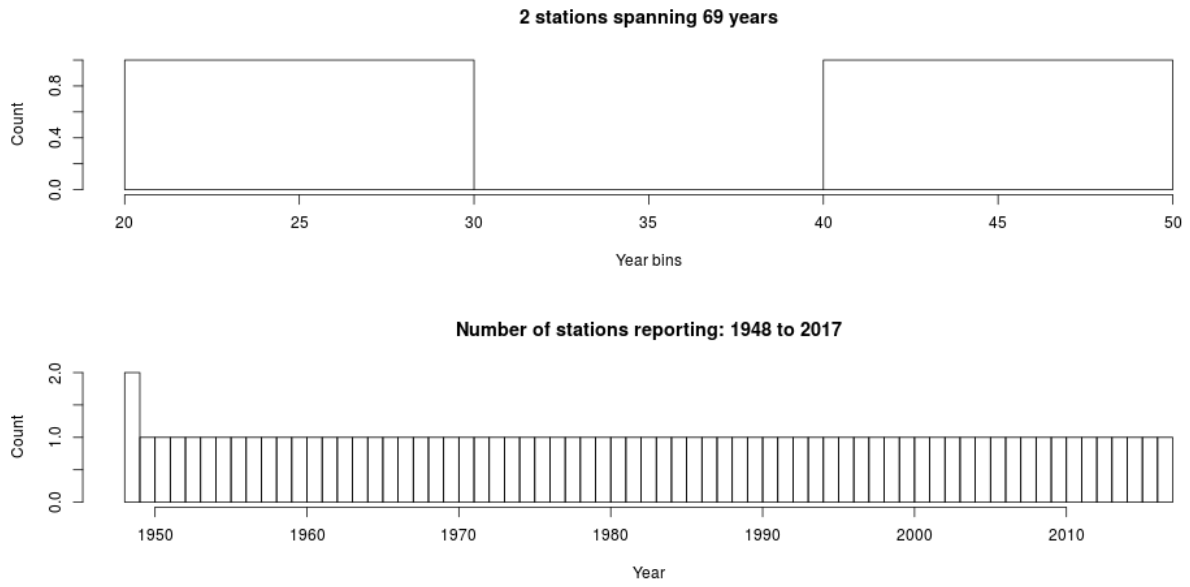


Figure 21: Gladys, VA, USA weather station coverage.

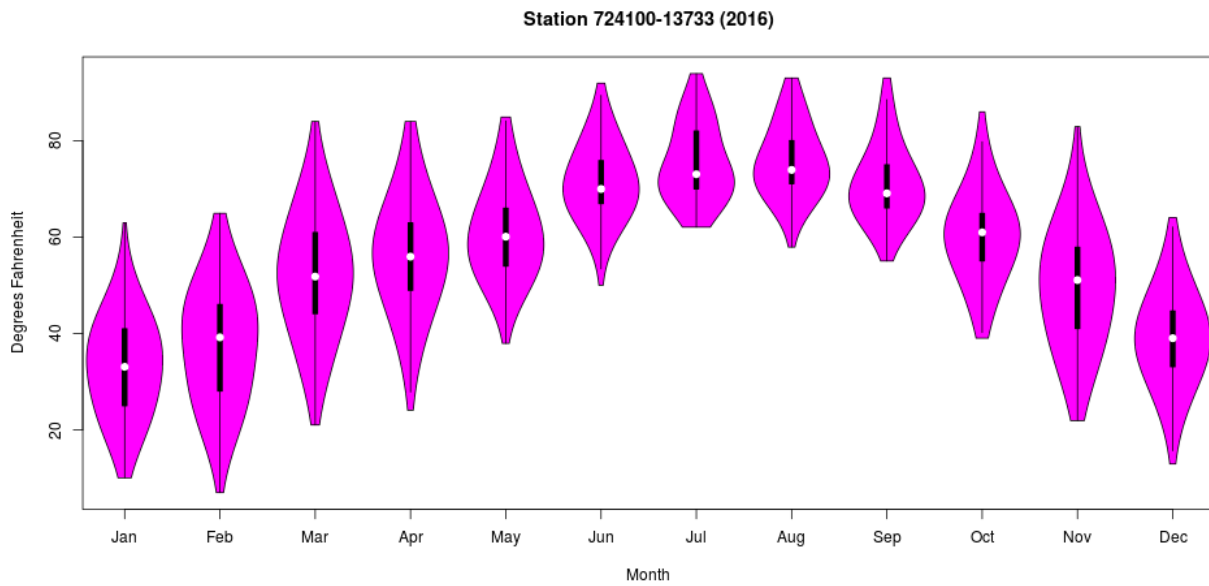


Figure 22: Gladys, VA, USA monthly average temperature.

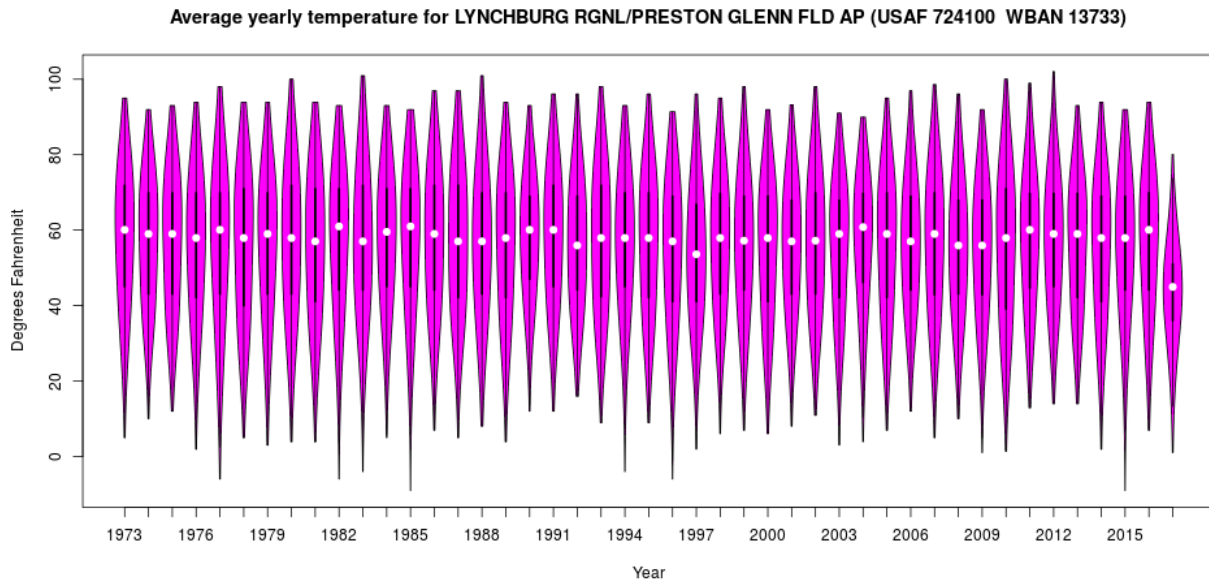


Figure 23: Gladys, VA, USA yearly average temperature.

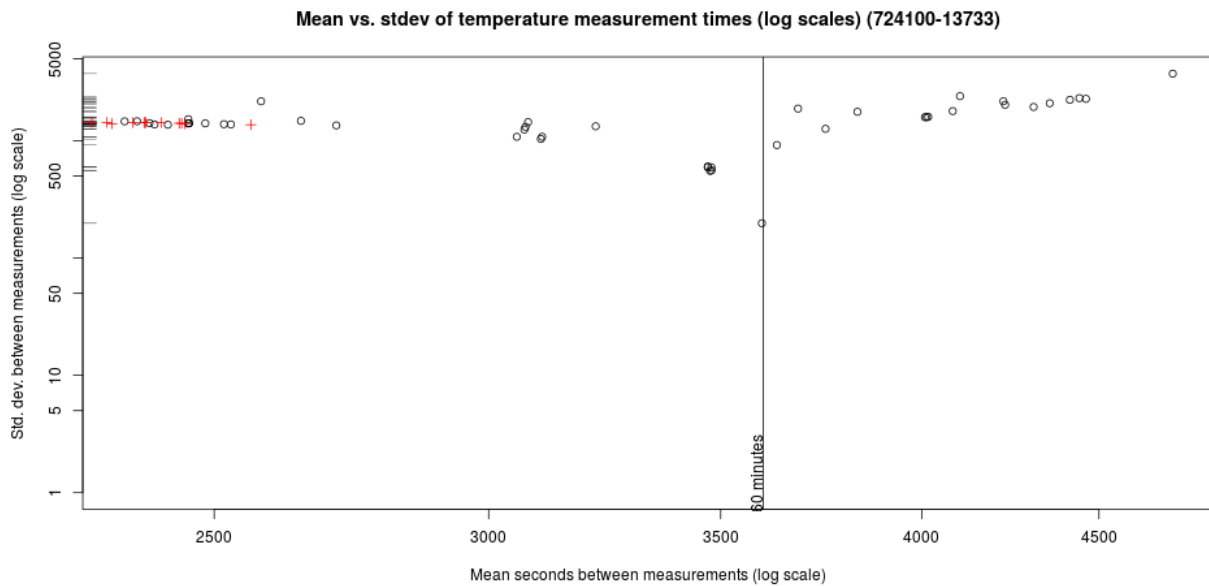


Figure 24: Gladys, VA, USA weather station reporting (log scales).

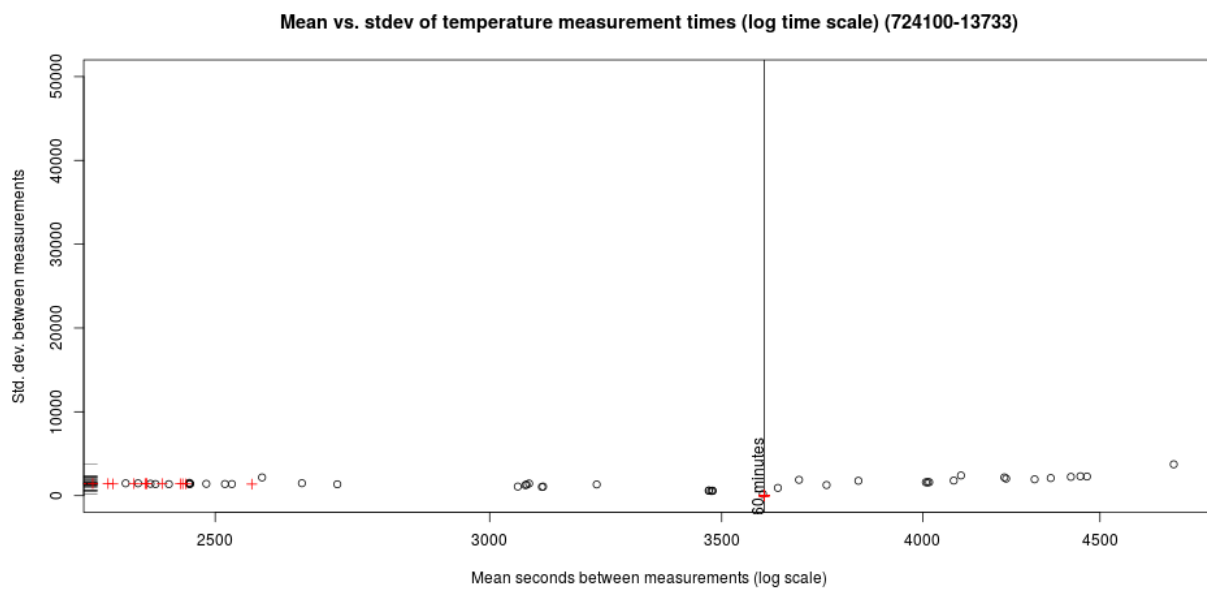


Figure 25: Gladys, VA, USA weather station reporting (semi-log scales).