Weather Station Records and Communicating Climate Change–Kentucky

Marc Los Huertos

June 29, 2022

Contents

1 Evaluating Terrestrial Meteorological Data

1.1 Selected History of Climate Science

Geologists have known the climate has been changing over the Earth's history. But what causes these changes has been a major research area for over 100 years. There are numerous drivers that contribute to changing climates – including the arrangement of the continents on the planet, the distance to the sun, energy generated by the sun, volanic activity, and the composition of the Earth's atmosphere.

It's the last one that we'll spend time because the Earth's temperature are changing pretty dramatically over the last 100 years and the cause is no mystery – the human activity that has released carbon dioxide (CO_2) into the atmosphere. The two main sources of CO_2 is from land use change, e.g. deforestration, and the burning of fossil fuels, e.g. coal, oil, and natural gas.

The first to propose the role of CO_2 on the Earth's atmosphere was a Swedish scientist Svante Arrhenius, who figured out that CO_2 absorbs infarred light. Moreover, he deduced that the Earth's temperature was actually warmer than it might otherwise be if CO_2 was not part of the Earth's atmoshere.

1.2 Why Look at Individual Stations?

I don't think there is a single, perfect way to analyze and communicate climate change. But the beauty of the network of stations in the USA and around the world is that these stations record weather as expecienced by local people. And while indiviudual stations may not represent the overall regional and global patterns well, this give us a mechanism to connect local experiences to regional or global processes.

Of course, some may fixate on the local pattern and remain unconvinced of the larger context and for those folks, there may be better ways to communicated climate data.

However, I would be remiss in failing to mention that some may fixate on local patterns and use these patterns to ignore or to dimiss the patterns in other regions.

Finally, the impacts of climate change are highly specific to the region in question. Thus, once someone understands the impacts on climate change in their region, they my not be able to appreciate how differnet the climate impacts might affect other peoples, who maybe more vulneratble, around the globe.

Thus, with these weaknessed in mind, I will pursue this project with an eye to address these other issues at later stages.

1.3 Approach

1.3.1 NOAA Data Records

The US National Oceanic and Atmospheric Adminstration (NOAA) maintains several sources of digital weather data from the USA and beyond. These data have been collected from stations around the country to support a wide range of human activities that include farming, aviation, shipping, and even armed conflict.

At various times, these records have been used to evaluate long-term climate change with varying success. Without a doubt, these data are not perfect, but they remain that foundation of an effective adn professionally maintained environmental monitoring program that engenders integrity, even when facing budget cuts.

I will use these data to select for a station with a long record for each state in the USA. Future projects might evaluate the record for stations around the world, but we will see about that.

1.3.2 rNOAA Package and R

R is an open source programming environment that has become one of the most popular tools for statiticians and data scientists. Capitalizing on the open source framework, a wide range of libraries or packages have been developed to faciliate data processing, analysis, and graphical displays. On such package is rNOAA developed to collect and display climate records stored on NOAA servers.

Using the package requires the use of a key. To maintain the integrity of the key, it's best to avoid posting the key in a public repository and to encryp the key to ensure it's not abused.

1.4 Selecting Weather Records by State

There are numerous ways to analyze temperature records, where stations can be analyzed individually or records could be sampled and analyzed in spatially in grids. Each of these are valid approaches depending on the question to be addressed.

In this case the question is "Based on the longest state meterological record, is there a temperature trend?"

1.4.1 Identify List of State IDs (FIPS)

Using the rNOAA library in R, we can queary NOAA's database to identify station codes (FIPS) by state. With the states and some territories, there are 55 FIPS for US weather stations.

rNOAA has a simple function to list for each of the states and the weather stations in each. I use ncdc_locs() functions to select each state and ncdc_station() to obtain the station ids with the longest records.

The function queries the NOAA website and retrieves state codes, "FIPS:XX". Each state has a number of weather stations,[1] some with a long record, some with a short record, and some with numerous interruptions. Our goal is to select a long record with few missing data.

1.4.2 Selection Stations

With the state ids, we can evaluate the metadata for all the weather stations, which will work to get the longest records, using `ncdc_stations()`.

First, we subset the data for stations that actively collecting data. Then we'll sort to the active stations to find the one with the longest records. We will use these stations for our analysis.

There were some records that didn't have robust TMAX/TMIN records, so there are some states that I had to manually select an alternative stations.

```
GSOM_Stations <- ncdc_stations(datasetid='GSOM',
               datatypeid = c("TMAX", "TMIN"),
               locationid=fips$id, limit=1000,
               sortfield = 'maxdate', sortorder='desc')

GSOM_Recent =
   GSOM_Stations$data[GSOM_Stations$data$maxdate>='2021-11-01',]

GSOM_Coverage =
```

[1] Project Idea: It would be nice to make a map of how concentrated the stations spatially.

```r
   GSOM_Recent[GSOM_Recent$datacoverage > 0.92,]
GSOM_Sorted =  GSOM_Coverage[order(GSOM_Coverage$mindate),]

GSOM_Longest = GSOM_Sorted[1,] #Pick longest
# Second and Third for Comparisons
# GSOM_Longest = GSOM_Sorted[2,]
# GSOM_Longest = GSOM_Sorted[3,]

# Exceptions Overrride by changing GSOM_Sorted[#,]
# Iowa (same as Hawaii)
if(fips$State=="Hawaii"){
   print("Manual Override for Hawaii")
   GSOM_Longest = GSOM_Sorted[1,]
}


# Iowa (same as Illonois)
if(fips$State=="Iowa" | fips$State=="Illinois"){
   print("Manual Override for Iowa/Illinois")
   GSOM_Longest = GSOM_Sorted[2,]
}

# Pennsylvania
if(fips$State=="Pennsylvania"){
   print("Manual Override for Penn")
   GSOM_Longest = GSOM_Sorted[2,]
}

# South Dakota

# Tennessee
if(fips$State=="Tennessee"){
   print("Manual Override for Tennessee")
   GSOM_Longest = GSOM_Sorted[2,]
}


#Puerto Rico
# Puerto Rico
PR = 0
if(PR == 1){
fips$State = "Puerto Rico"
fips$id = "FIPS:PR"

GSOM_Selected = ncdc_stations(datasetid="GSOM",
                              stationid = "GHCND:RQC00665097",
                              datatypeid = c("TMAX", "TMIN"))
"Doing Puerto Rico"
GSOM_Longest = GSOM_Selected$data

}

GQ = 0
if(GQ== 1){
fips$State = "Guam"
fips$id = "FIPS:GQ"

GSOM_Selected = ncdc_stations(datasetid="GSOM",
                              stationid = "GHCND:GQW00041415",
                              datatypeid = c("TMAX", "TMIN"))

GSOM_Longest = GSOM_Selected$data
}

str(GSOM_Longest)
# Change Case of Station Name -- Complicated!
#v2 <- gsub("[sty]", "", paste(letters, collapse=""))
#chartr(v2, toupper(v2), GSOM_Longest£name)
#sub('\\b([a-z])([0-9])', '\\L\\1\\2', GSOM_Longest£name, perl=TRUE)
#[1] "JAStADMMNIsyNDK" "LAUKsNDTUsAINS"

# Caution this might be a problem!
fips$State2 <- sub(" ", "_", fips$State)
```

The record selected has the following metadata associated with it, which will be used for nameing, labeling, and mapping.

elevation mindate maxdate latitude name 2 293.2 1872-11-01 2022-06-01 38.03391 LEXINGTON BLUEGRASS AIRPORT, KY US datacoverage id elevationUnit longitude 2 0.9243 GHCND:USW00093820 METERS -84.61138 [1] 1872 [1] 2022

2 Gathering Weather Record Datasets

2.1 Main Datesets of Interest

GSOM

CHCND

CHCNM

2.2 Functions to Collect and Clean GSOM

To collect the data, I used a short function, but the download time is painfully slow because only 1 year can be obtained at a time. Might want to get a work around for this at some point.

Functions to bin data into decades and scores.

The function relies on two inputs, the station id and the measured parameter – TMAX and TMIN in this case. After that, the data needs to be clean up quite a bit.

Furthermore, I have converted units to Farenheit, which is not my favorite, but important for US consumption.

2.3 Functions to Report Probabilities

2.4 GSOM: Retreive and Clean Data

2.5 CHCND: Retreive and Clean Data

CHCND have been bias corrected...

3 Data Analysis Processes

3.1 Map Weather Station Location

```
## Error in eval(expr, envir, enclos):  object 'AIzaSyDYUc3ExxqFTOHtyxyr6'
not found
```

3.2 Using a Linear Model Monthy Trends

I used a linear model (`lm()`) to evaluate the long term trend for each each month to determine which, if any, have long-term trends. At somepoint, I'll have to the stats correcting for the autocorrelation using a autoregressive model.

Evaluate both TMAX and TMIN in GSOM by Year using MonthEvalStats() function.

3.2.1 Trends in Tabular Formats

Admittedly, determining the months with the biggest changes isn't a very good approach for hypothesize testing – it's more like a fishing expedition, but as long as we understand the difference between an a priori hypothesis and an exploratory analysis, we should be okay if we make appropriate conclusions.

For this section, we'll look to see what months had the greatest changes for both TMIN and TMAX. By looking at significant slopes in whatever direction, we might learn if warming is really the dominiant pattern.

Table **??** summarizes the monthly trends for TMAX.

Table **??** summarizes the monthly trends for TMAX.

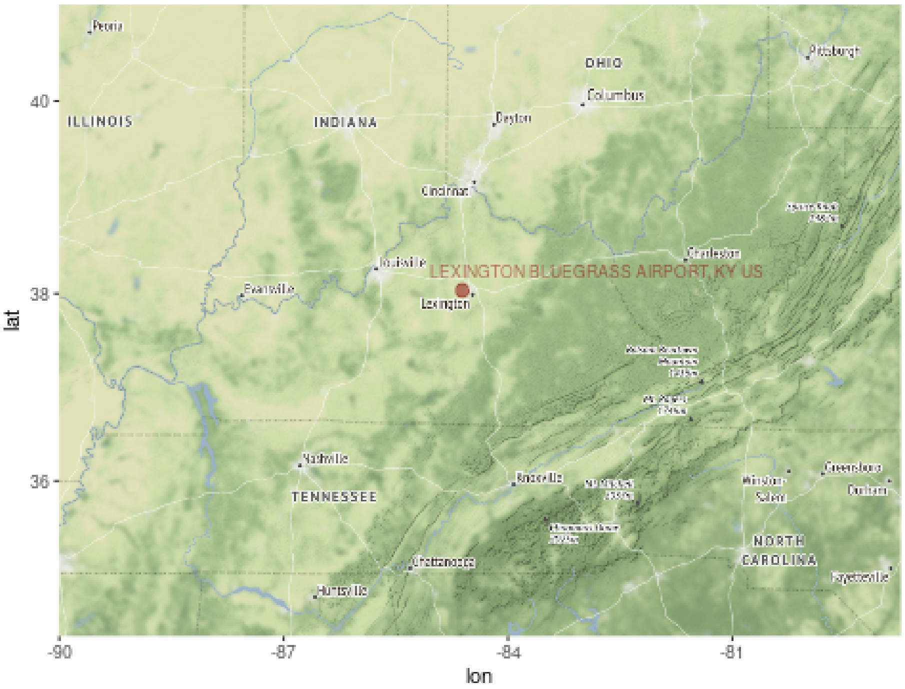


Figure 1: Weather Station Location (USW00093820).

	Month	Slope100	r2	p_value	Symbol
1	1	-0.0028	0.00	0.6108	
4	2	0.0014	0.00	0.7867	
7	3	0.0091	0.04	0.0389	*
10	4	0.0050	0.03	0.0870	
13	5	0.0061	0.04	0.0560	
16	6	0.0026	0.02	0.2203	
19	7	0.0015	0.01	0.4592	
22	8	0.0024	0.01	0.2915	
25	9	0.0003	0.00	0.9294	
28	10	0.0006	0.00	0.8605	
31	11	0.0038	0.01	0.2595	
34	12	0.0077	0.03	0.1022	

Table 1: TMIN Trends

	Month	Slope100	r2	p_value	Symbol
2	1	-0.0044	0.01	0.4285	
5	2	0.0008	0.00	0.8876	
8	3	0.0105	0.04	0.0418	*
11	4	0.0063	0.03	0.0684	
14	5	-0.0008	0.00	0.8345	
17	6	-0.0010	0.00	0.7547	
20	7	-0.0045	0.03	0.1204	
23	8	-0.0008	0.00	0.7934	
26	9	-0.0036	0.01	0.3358	
29	10	-0.0038	0.01	0.2942	
32	11	0.0064	0.03	0.0908	
35	12	0.0076	0.02	0.1250	

Table 2: TMAX Trends

PPT changes are tricky to capture and I'll have to keep working on this (Table **??**).

3.2.2 Defining TMAXmonth and TMINmonth

The greatest changes for Station GHCND:USW00093820

4 Communicating Long-term Weather Records

4.1 Complete Records vs. Post 1975 Trends

Communicating climate change based on station records is tricky. The long-term record would on the surface to be the most robust, but several issues arise with a naive analytical approach – my favorite!

	Month	Slope100	r2	p_value	Symbol
3	1	-0.1422	0.05	0.0260	*
6	2	0.0618	0.02	0.2138	
9	3	-0.0258	0.00	0.6446	
12	4	0.0739	0.02	0.1611	
15	5	0.1599	0.08	0.0058	**
18	6	0.0325	0.00	0.5659	
21	7	0.0300	0.00	0.5844	
24	8	0.0530	0.01	0.3038	
27	9	0.0806	0.03	0.1167	
30	10	0.1561	0.12	0.0006	***
33	11	0.0156	0.00	0.6950	
36	12	0.1064	0.05	0.0256	*

The noise in the data may suggests that no trend is present (Figure 2). It's tricky because the seasonal variation dominates the source of varition. In other words the intra-annual variation exceeds the inter-annual variation, making signal detection very difficult. Is there a way to "filter" out the seasonal effect? Yes, let's see how that works next.

4.2 Filtering Seasonal Effect

There are several ways to filter out seasonal effects. The easiest way is subtract the mean value for each date, but that's tricky because every four years there is an extra day in Februrary – although there are ways to deal with this, a more straight forward way is to use mean monthly values to capture the seasonality for each month. With 12 months, this is a pretty good approach because there is pretty good resolution.

4.2.1 Method 1: Filtering by Monthly Mean

And to see what we created, see Figure 3.

4.2.2 Method 2: Polynomial Filter

Project to be followed up with.

```
# fit polynomial: x^2*b1 + x*b2 + ... + bn

# create time series object
#X = [i%365 for i in range(0, len(series))]
# y = series.values

# degree = 4
#coef = polyfit(X, y, degree)
```

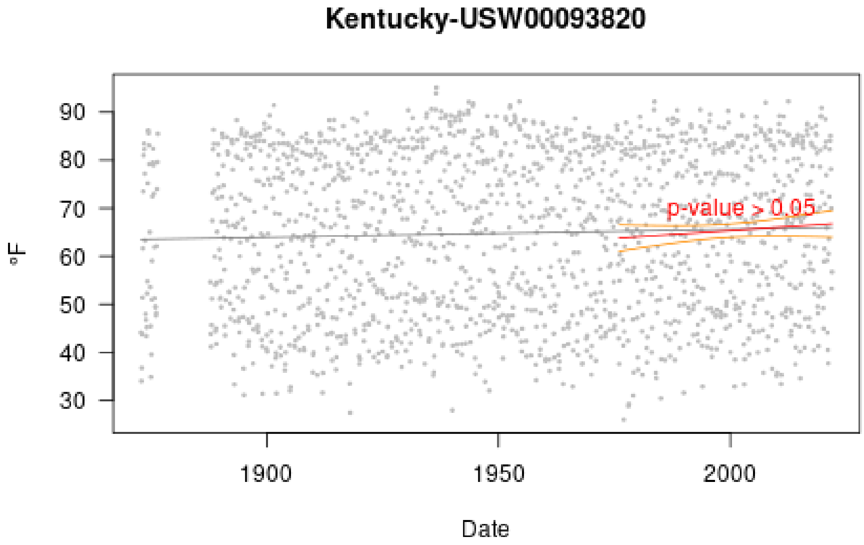


Figure 2: The climate trends from full record and post 1975 data. These data have a high level of variability due to seasonality effects that have not been filtered out.

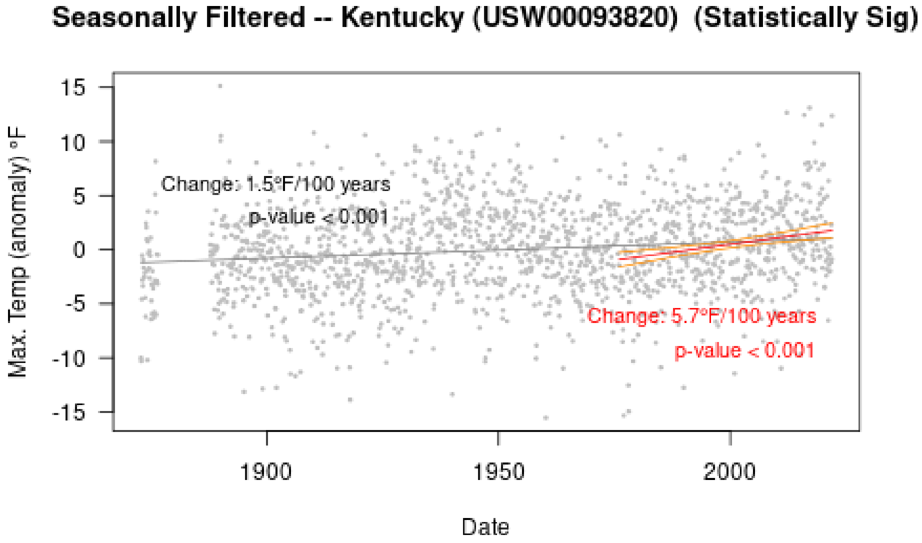


Figure 3: The changing in monthly temperature data.

```
# print('Coefficients: %s' % coef)
# create curve
```

4.3 Extreme Events–Using Daily Records

4.3.1 Complicated Nature of Rainfall Patterns

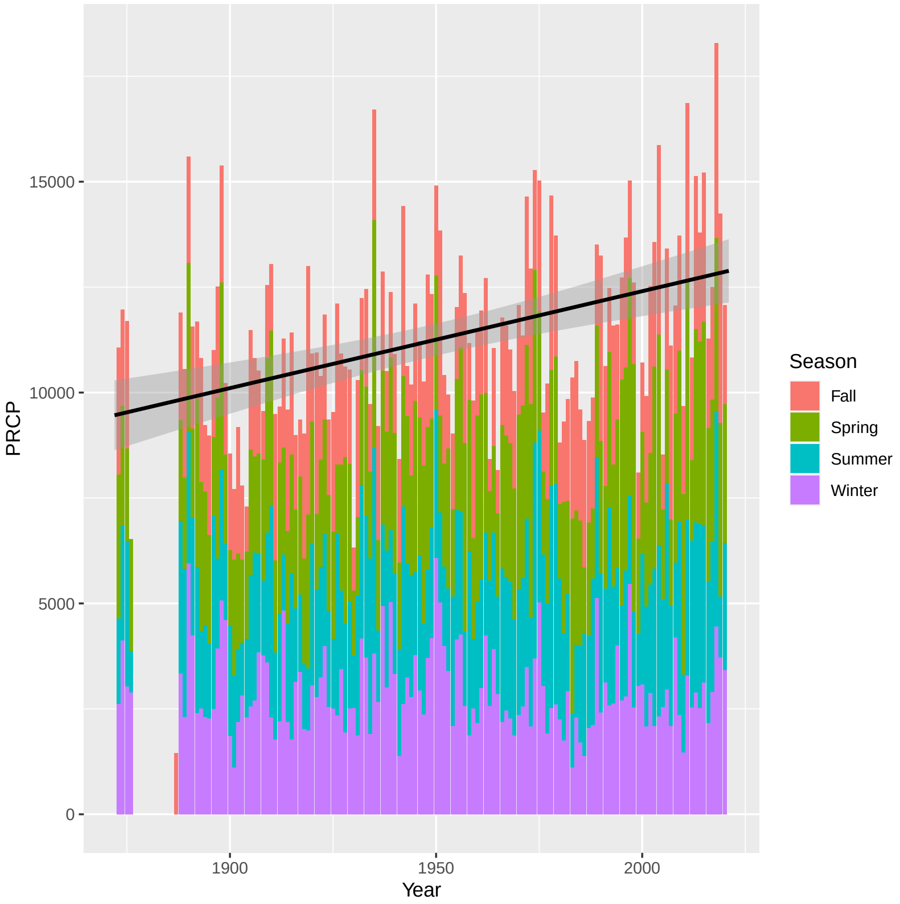
Rainfall trends are tough. Exteme events can occur in 24 hours or over long periods that might result in floods or droughts. Each region might have different patterns, so developing a consistent approach is tough.

We can look for trends in monthly averages, number of days without rain (important in tropics), and/or extreme events based on daily or hourly data.

I don't know of a robust way to look at this for the entire globe.

Rainfall totals by season might be a useful way to think about changes, because the rainfall is often seasonal, I wonder if we can see pattners by season.

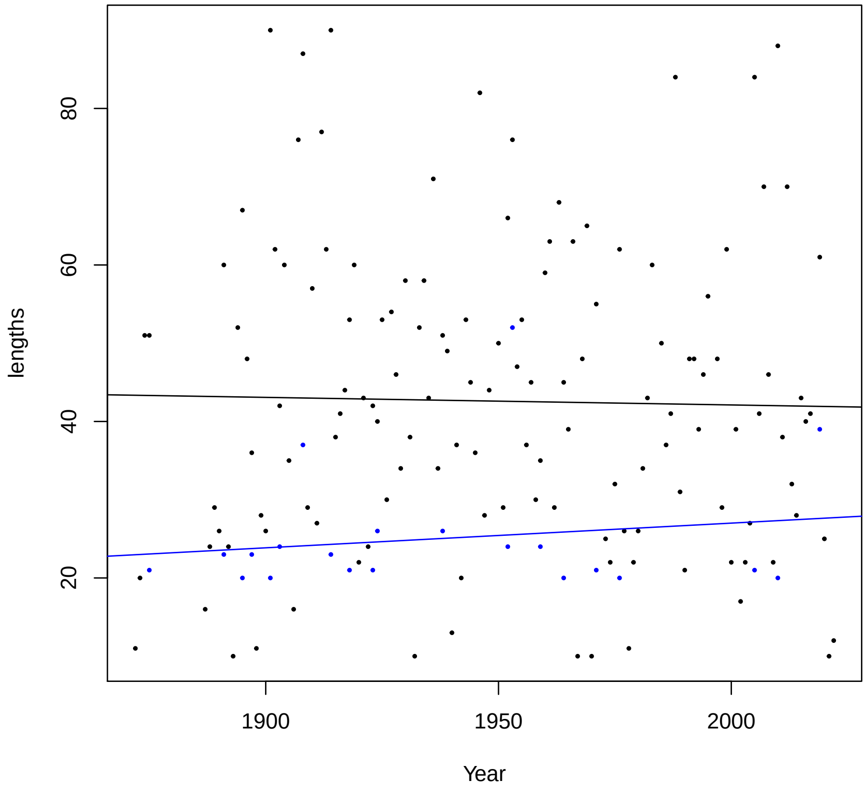
```
## `geom_smooth()` using formula 'y ~ x'
```



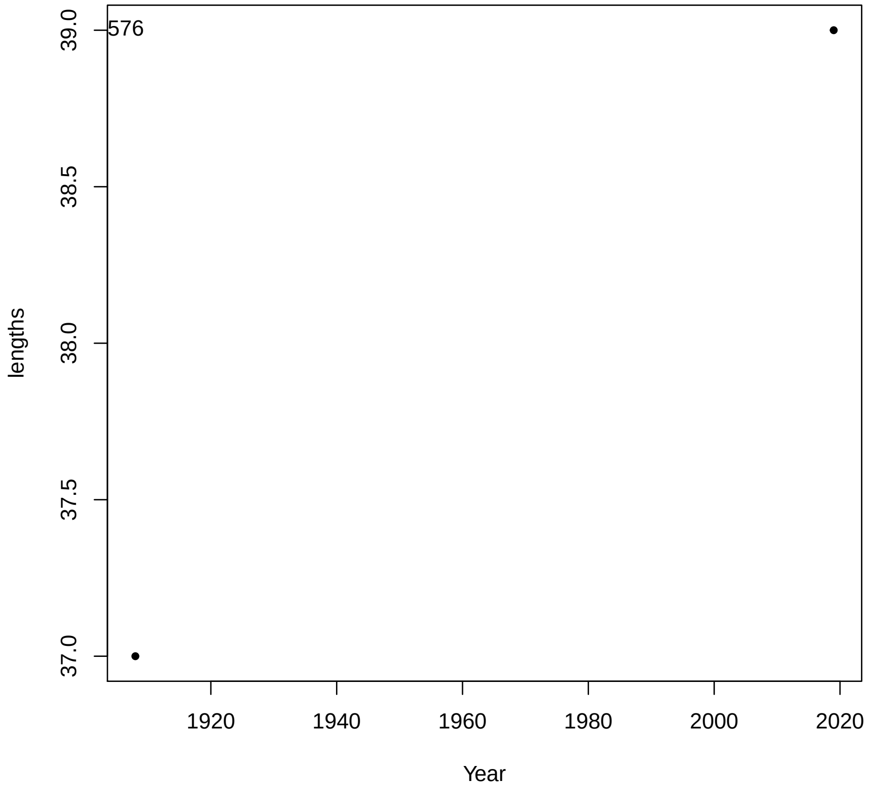
4.3.2 Drought

Days without rain...within a calendar year... bleed over between years isn't captured.. This is screwed up, Drought.run needs work.

```
## Error in aggregate.data.frame(mf[1L], mf[-1L], FUN = FUN, ...):
no rows to aggregate
## Warning in if (Drought.run$lengths > 100) {: the condition has
length > 1 and only the first element will be used
## Error in if (Drought.run$lengths > 100) {: missing value where
TRUE/FALSE needed
## Error in eval(m$data, eframe): object 'Drought.run.40' not found
## Error in eval(m$data, eframe): object 'Drought.run.100' not found
```



```
## Error in is.data.frame(data):  object 'Drought.run.40' not found
## Error in is.data.frame(data):  object 'Drought.run.100' not found
## Error in is.data.frame(data):  object 'Drought.run.100' not found
```



Rainfall Probability Distributions by decade... to be developed.

4.4 Record Setting Temperature Records

In many cases, people seem to ”feel” how temperature has been changing over time, and new records seem to capture the attention in the media. So, we’ll create a updated record of maximum temperatures and display them.

This is a common way to communicate temperatures changes. I suspect we have a better sense of change when we notice ”extreme” events...

I tried to use a for loop and in then statements and it was painfully slow, so I converted the data to a matrix that can be used by barplots with much more effeciency!

Create the matrix

```
## Error in TMAX.mat.noleap[j, year.seq$Col[year.seq$Year == i]] <-
CHCND.noleap$TMAX[CHCND.noleap$Year == :  replacement has length zero
```

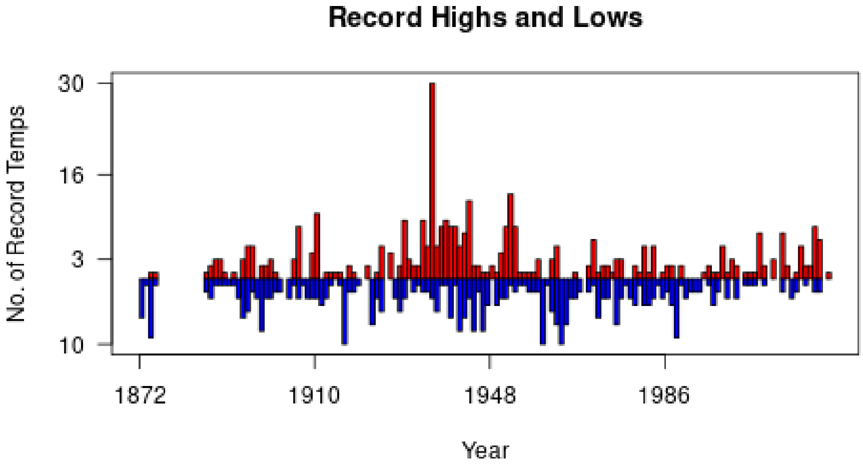


Figure 4: Daily temperatures that have been the highest on record (in red) and lowest on record (in blue). In some cases, climate change has created more records in the recent decades, while other stations seem don't show that trend.

The patterns of record temperatures often shows increasing number of new high temperature records and fewer record low temperatures more recently, but as usual, it depends on the location (Figure 4).

4.5 Iterate TMAX vs. Month Boxplots

Evaluating the changes in TMAX and Monthly temperatures might be useful, but for now, I think it's hard to see the patterns.

4.6 Four Plots Compelling Figures

To test the code, I have created graphics that can then be used in the animation process, i.e. try to create code that doesn't get too complicated and then fail!

4.7 KISS

Keeping it simple is critical in communicating scientific information. In this section, I try to come up with a consistent message for every state and a simple graphic.

4.7.1 Change Point Analysis

First, TMIN and TMAX and change point analysis...

https://cran.r-project.org/web/packages/mcp/readme/README.html

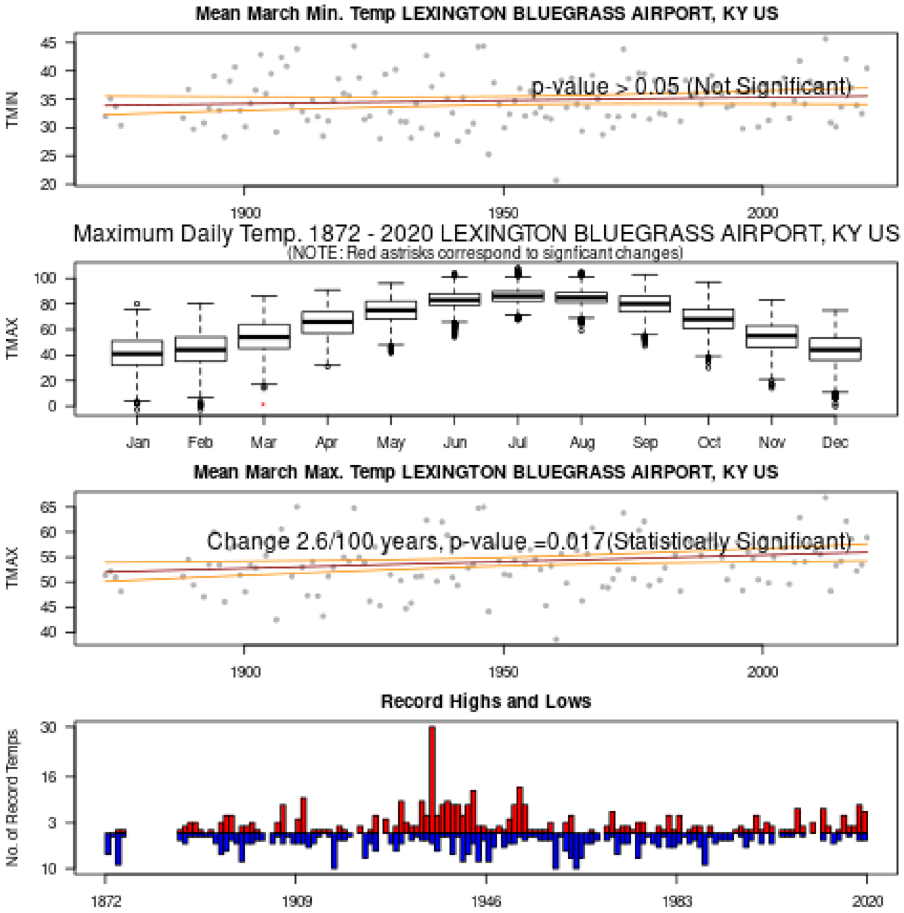


Figure 5: Climate can be analyzed using several types of lenses. In this case, we have analyzed show the months with the greatest changes. The first figure is monthly average of TMINs (daily low temperatures) with a best fit line. The second figure shows the monthly TMAX range and asterisks indicate singificant changes over the station record and the third figure is the trend for these TMAXs over time and includes the best fit line. The final figure shows the daily temperatures that have been the highest on record (in red) and the lowest minimum temperatures (in blue). In some cases, climate change has created more records in the recent decades, while other stations seem don't show that trend.

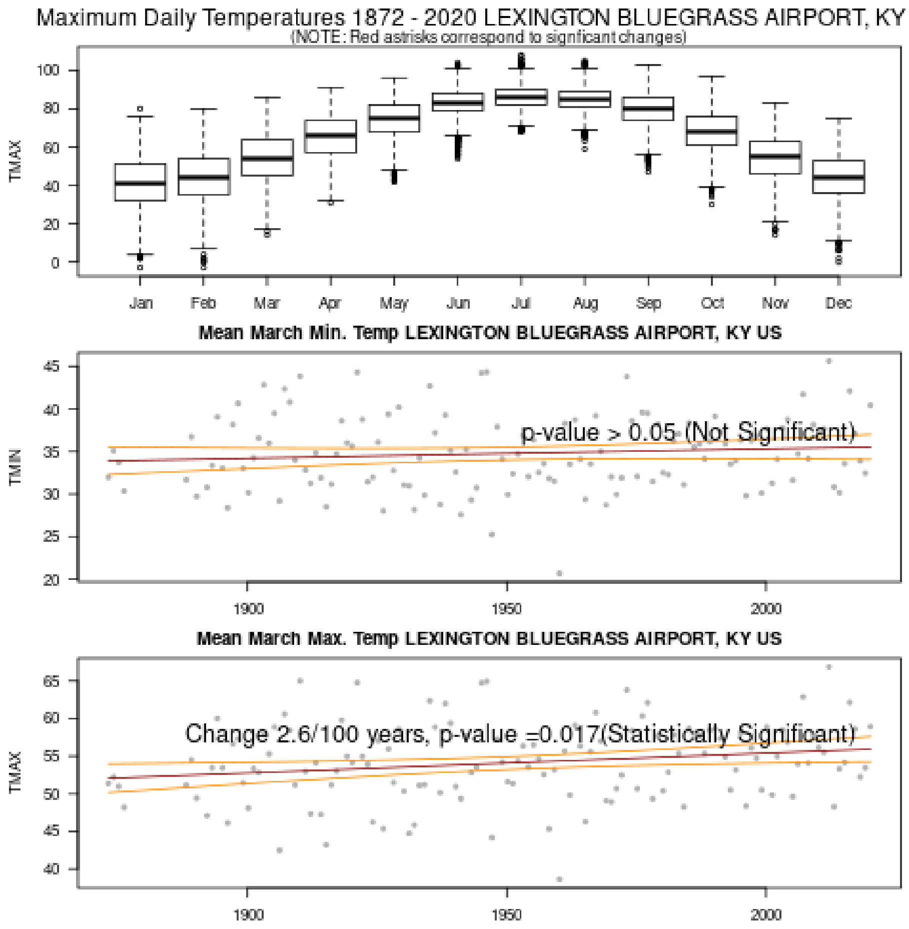


Figure 6: Keep it simple stupid!

Let's create a figure that simplifies the narrative, if we can!

4.8 Temp & Precipitation Probability

To highlight the patterns of change, it might be useful to analyze how the probability ditributuion might change – we can use a normal probability distribion as a theoretical distribution (and we can check if this distribuion is appropriate with a Chi-Square test), or we can use the data to create a emperical distribution, which is my favored approach.

I started with decade bins, but used 20 years bins (scores) to simplify the graphics while keeping a pretty good temporal resolution.

This figure is pretty effective, but still needs work.

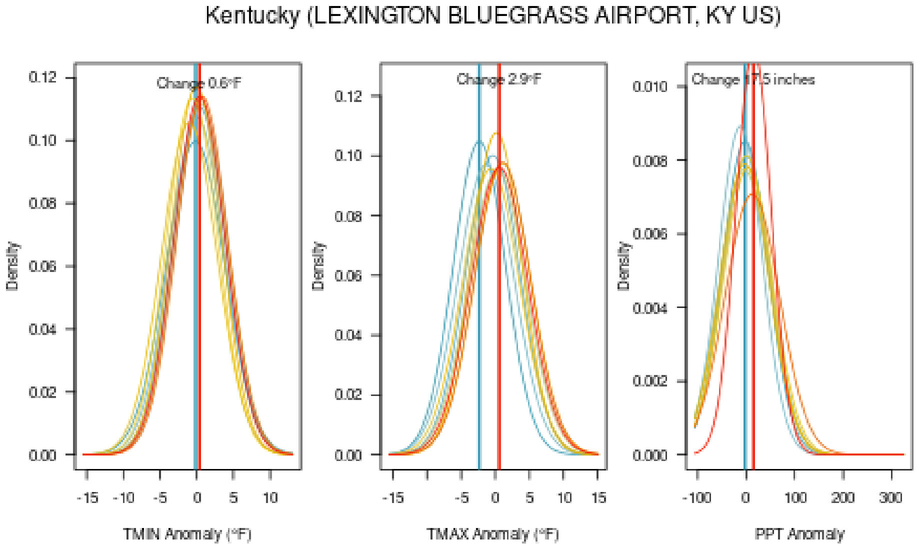
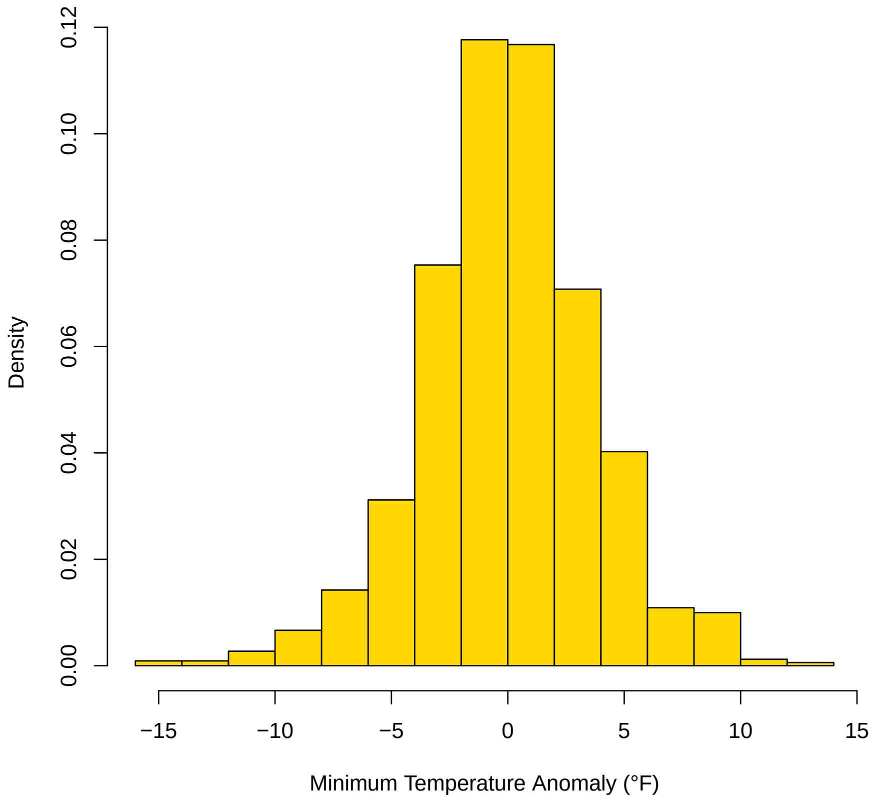


Figure 7: The changing in monthly temperature data, assuming a normal probability distribution.

4.9 Using library densEstBayes

Now, I used a screen split to look at the distribution of the temperate anomolies. First, we look at a simple histogram of the entire dataset.



The data center around zero, as expected, but are these normally distributed?

These values suggest that there is good reason to avoid the normal probability distribution.

Next we use a function to estimate the probability distribution using a markof chain the creates an estimated probability distribution. This doesn't always work when the distribution is not even and their only 10 years of data per slot. I suspect, I should make this by every 20 years. Plus that will go way faster and I think the data visualization will be more robust.

The process to create these figures is very time consuming, so in general, I need to come up with an if then statement to avoid creating these everytime!

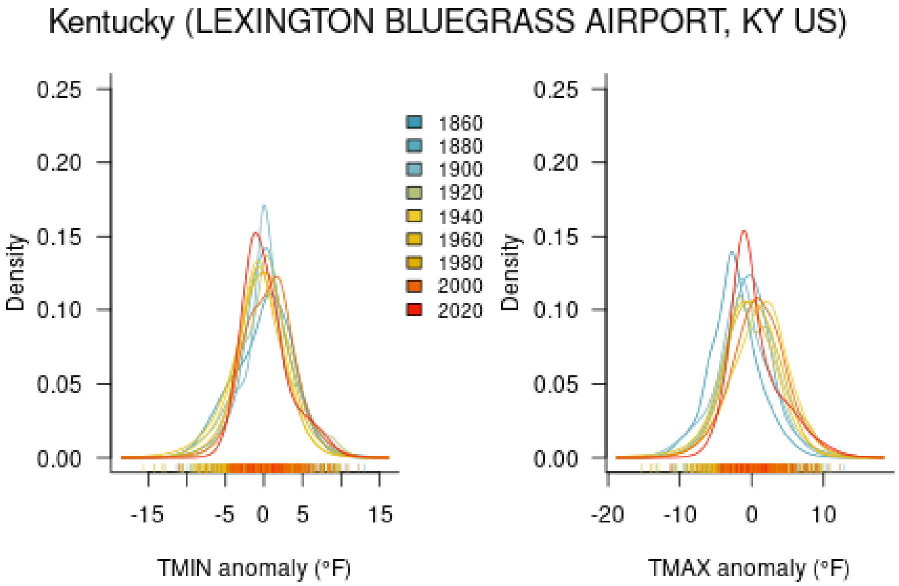


Figure 8: The changing in monthly temperature data.

5 Animated GIFs

So far, this creates a gif file, but I haven't been able get the gif in the pdf directly yet. I will need an additional package or create separate png that are combined. For now, we'll create a gif file to be used in separate documents.

5.1 Probability Distributions

The file is saved in the main directory.

5.2 4 Weather Trend Plots

The file is saved in the main directory.

5.3 Evaluating Records

TBD

5.4 Export Options

TBD

6 Sea Surface Temperature Data – SURP PROJECT WAITING TO HAPPEN

In contrast to terrestrial data, sea surface temperature (SST) is quite difficult to obtain and process. There are numerous tools to access the data, but they often require knowledge of complex software tools that are not easy to set up or programming experience with python or others.

`https://climexp.knmi.nl/select.cgi?id=someone@somewhere&field=ersstv5`

There are, however, a few tools build for R users that seem to accomplish all that we need.

`https://rda.ucar.edu/index.html?hash=data_user&action=register`

`https://rda.ucar.edu/datasets/ds277.9/`

Alternatively, we can download flat ascII tables of gridded data:

`https://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v5/ascii/`

7 Satellite Data

TBD

8 Ice-Core Data

TBD

9 Conclusions

Developing a robust method to analyze weather stations is both time consuming and difficult to justify the outcome. In part because the data suggest that each station (region) requires different types of analysis, based on the expected patterns of temperature and rainfall. As climate scientists have known for decades, the terminology of global warming is not very useful. Not because scientists are trying to hide something or promote some biased agenda, but that even as warming of the global average is well documented, the impacts of climate change on each region is highly specific, requiring specificity in the analysis.

Hopefully, this little analysis has created some mechanism for others to appreciate this compexity.

The document took 6.1 minutes to process and compile. My next goal will be to optimize the process and streamline the time to analyze.