November 20, 2017

WEATHER TRENDS – Global vs. Sacramento

By Alissa McBain

SQL Exploration and Export

First, I needed to pull the data from the closest city to myself. But, which city is nearest to me? And what could I filter by so that I don't have to scroll through 345 results? What does the table look like, and what information does it contain that I could filter on?

```
SELECT *
FROM city_list
LIMIT 5;
```

From this I saw that there are two columns, `city` and `country`. I filtered for only 'United States' to get a smaller list to scroll through.

```
SELECT city
FROM city_list
WHERE country = 'United States';
```

This gave me 52 results, which is much easier to scroll through to find my nearest city. Sacramento is on the list, so I continued with that as my nearest city.

Next, I needed to extract the Sacramento data. To see what type of data, and column headers are contained in the `city_data` sheet I ran this query.

```
SELECT *
FROM city_data
LIMIT 5;
```

There was `year` (number), `city` (string), `country` (string), and `avg_temp` (number).

To only pull Sacramento:

```
SELECT *
FROM city_data
WHERE city = 'Sacramento';
```

This gave 165 results, but the columns `city` ('Sacramento') and `country` ('United States') were not important and redundant, so I repulled the data without those columns.

```
SELECT year, avg_temp
FROM city_data
WHERE city = 'Sacramento';
```

This also gave 165 results, so I was confident I captured them all.

Next, I clicked "Download CSV."

Moving right along to downloading the global data. I was curious to see how this table is structured.

```
SELECT *
FROM global_data
LIMIT 5;
```

The only two columns in this table are `year` and `avg_temp`, so I downloaded it and exported to CSV.

```
SELECT *
FROM global_data;
```

I moved both of these files to my google drive (to keep organized) and then renamed them sacramento_weather.csv and global_weather.csv.

I used Excel from here on out.

Excel Exploration and Visualization

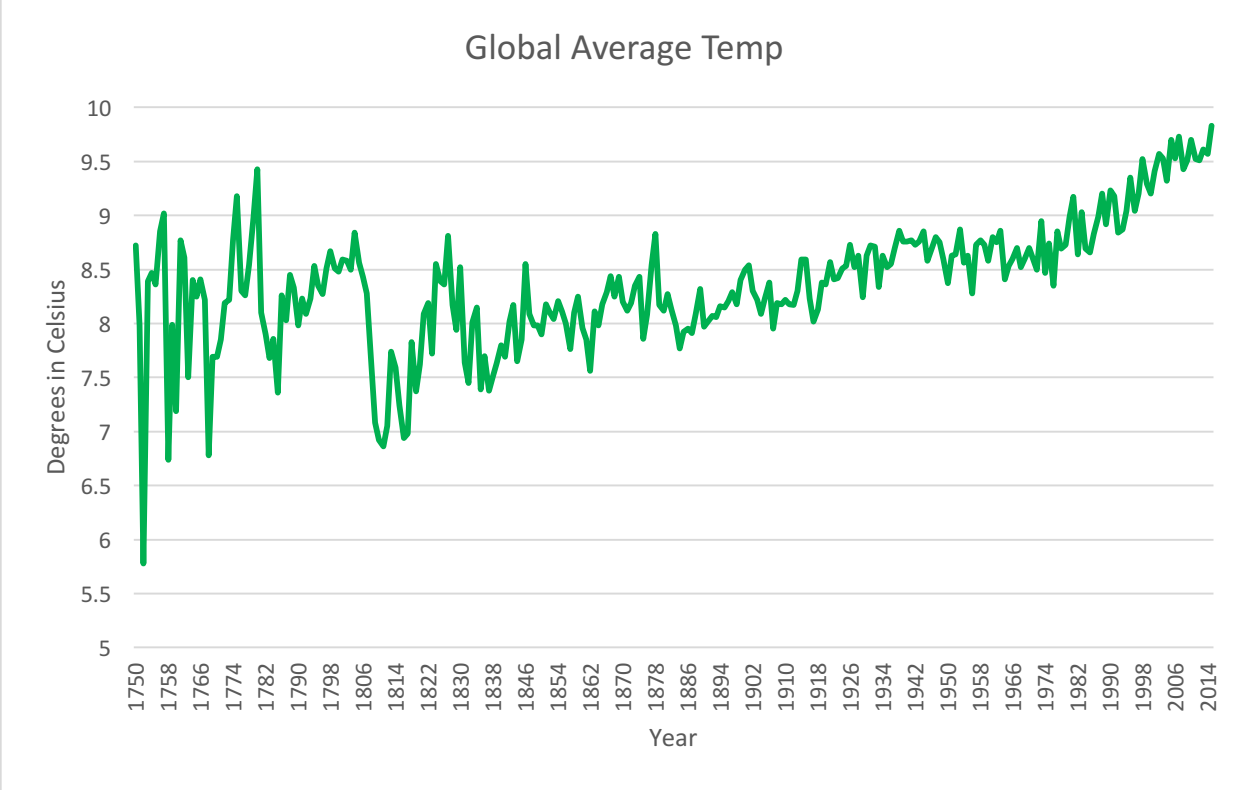
In Excel, I imported the data as text into two sheets, “global data” and “Sacramento data.”

First, I looked at the global data; I created the line graph and the line was quite jagged and didn’t fill the chart area well.

I did a couple quick formulas on the data to see what the minimum and maximum values are.

=MIN(B2:B267) = 5.78
=MAX(B2:B267) = 9.83

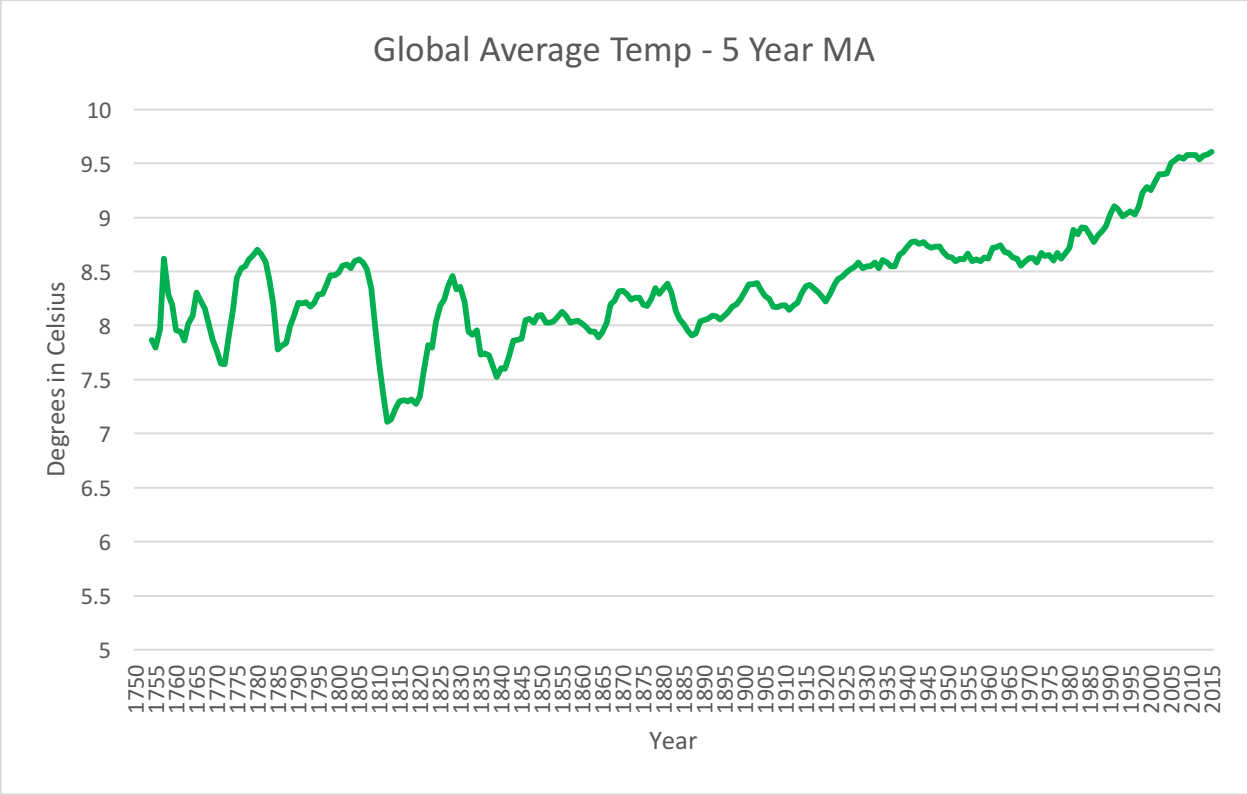
Based on these results, I changed the y-axis bounds from 0-12 to 5-10 so that the view was more narrow and condensed. The x-axis was Excel’s default.



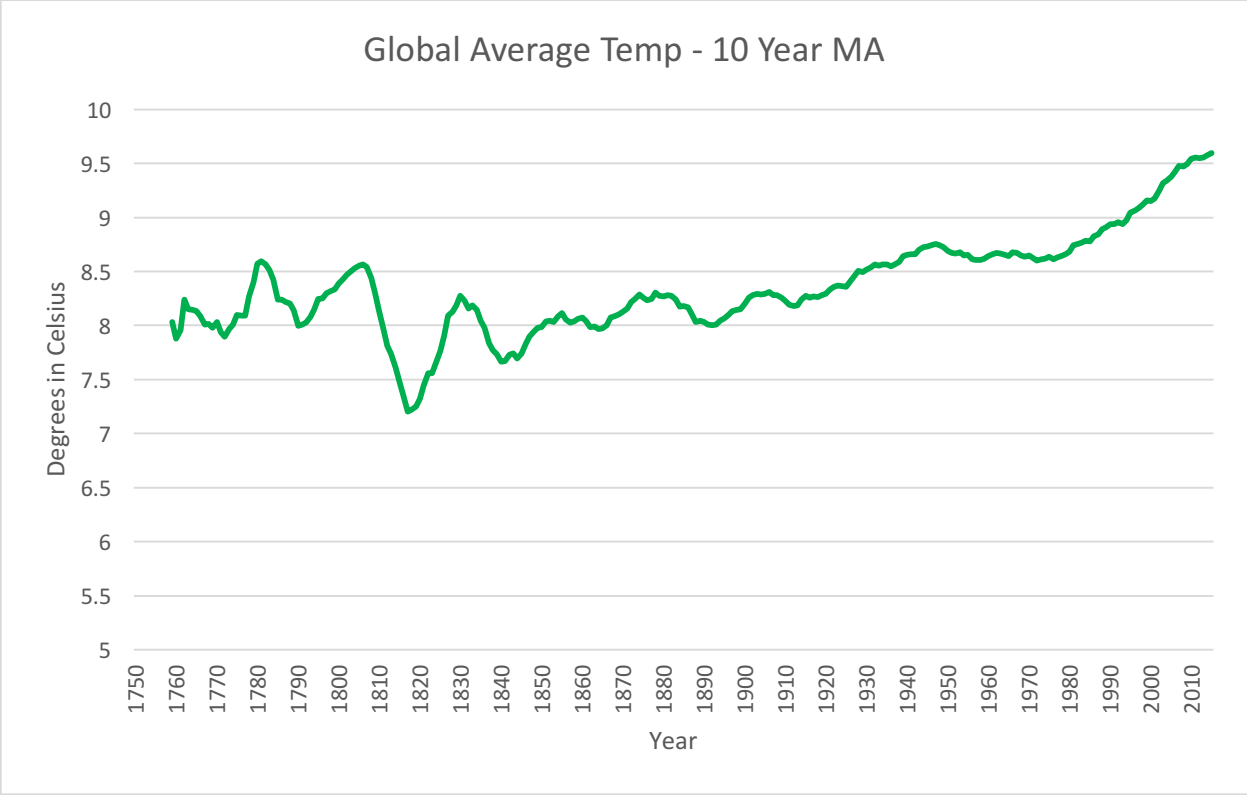
Next, I calculated the 5-year moving average in the column to the right of the avg_temp using this formula in cell C6:

=AVERAGE(B2:B6)

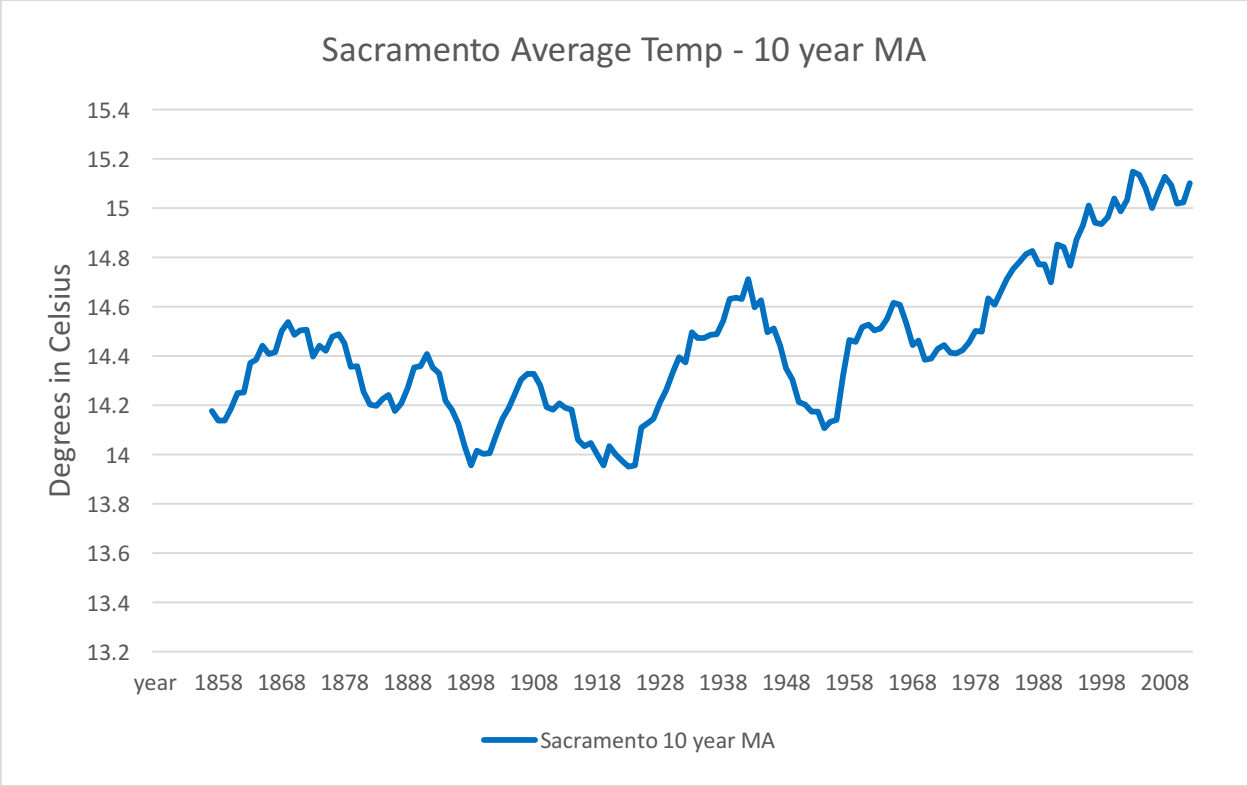
Then I filled that down the remainder of the data, and plotted it, although that as well was quite jagged. I also changed the x-axis labels so that they are labeled for every 5 years.



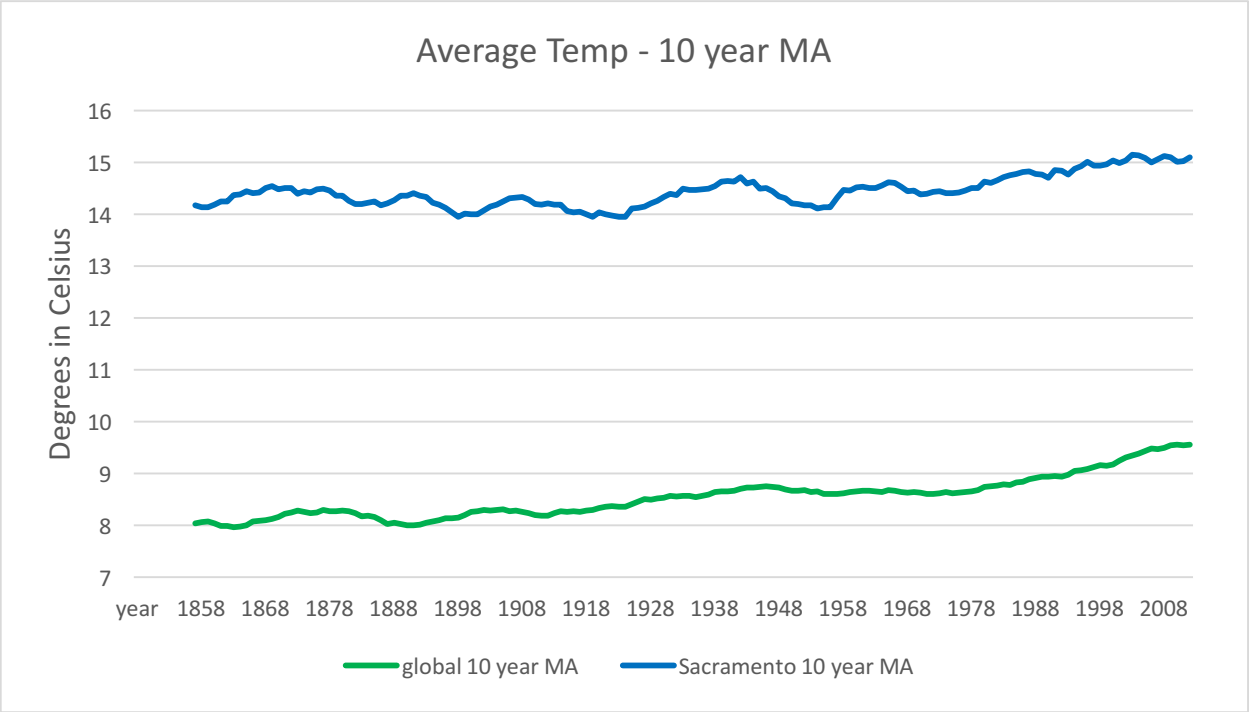
I did the same two steps (calculate and plot) the 10-year moving average and it looked much smoother, so I ran with that.



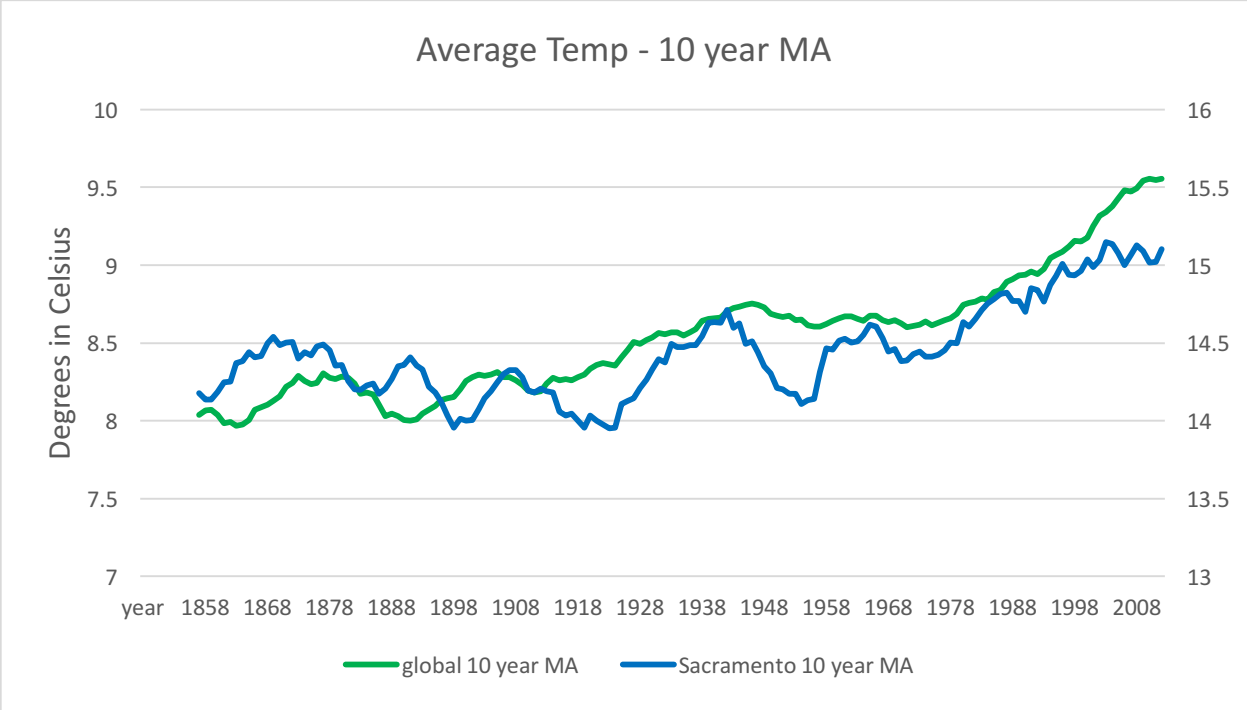
Next, I calculated the 10 year moving average for the Sacramento weather data, and plotted that on a line chart (changed the horizontal axis to have ticks every 10 years for consistency).



And then I combined the two charts so that I could see Sacramento *versus* the global temperature. Although the global data dates back to 1750, the Sacramento data begins at 1849 (1858 being the first record of the 10 year MA), so that is where I started the graph.



It's hard to get many meaningful insights from this chart, other than Sacramento is much warmer than the global average. So, I plotted the global data on the primary (left) axis, and the Sacramento data on the secondary (right) axis. This gave me a better idea of how each region changed over time in relation to each other.



From here on out, I truncated the global data and only worked with it from 1849 onward, so that my comparisons were fair.

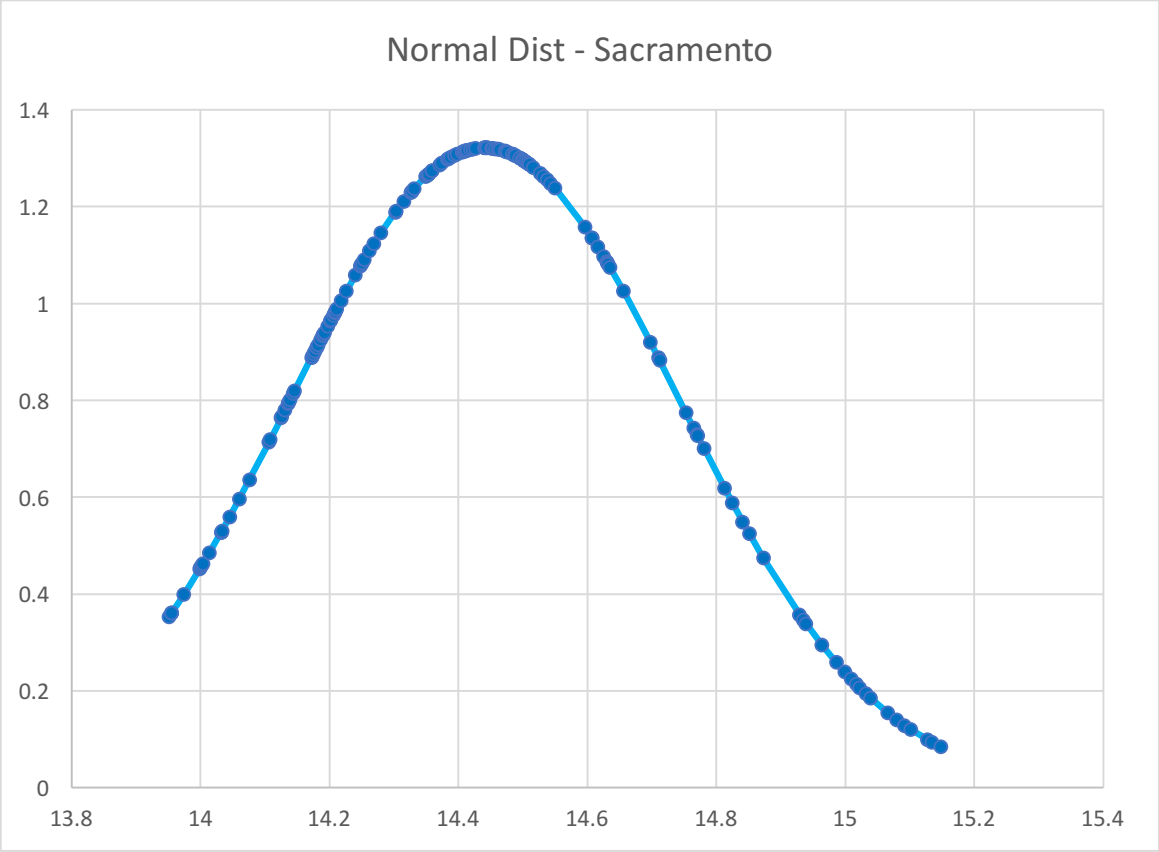
Here are a few things I noticed off hand about the data, that I explored further.

- While both temperatures went up over time, it seems the global temperatures went up more than the Sacramento temperatures.
- The Sacramento data has more volatility than the global data (as to be expected from one city versus the world).
- They seem to be correlated, slightly running in parallel to each other.

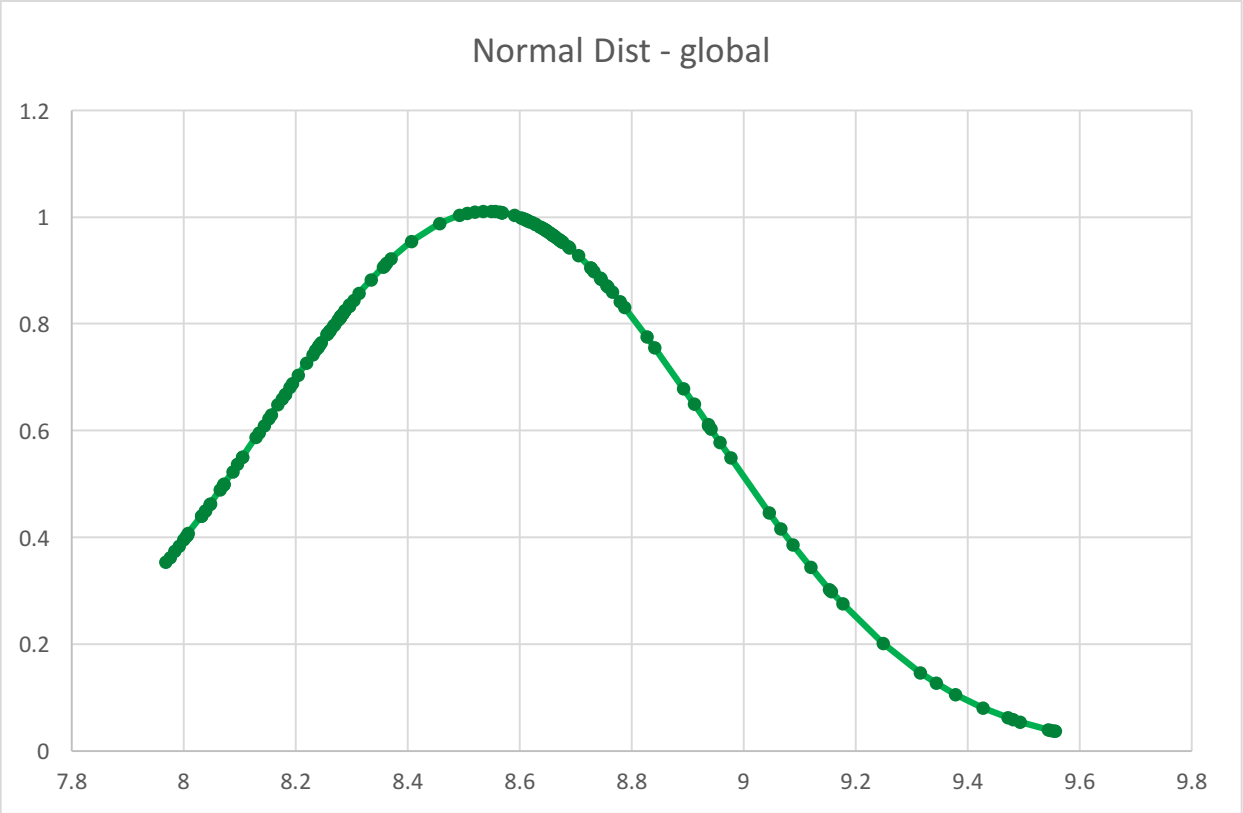
I began with a quick calculation to see how large the range was for each 10-year MA data set for the 156 data points I had. The minimum for Sacramento was in the year 1924 with a 10-year MA temperature of 13.951, while the maximum was 2004 with a temperature of 15.148. This gives us a range of 1.197. Likewise, for the global data 10-year MA, the minimum was in 1864 with a temperature of 7.968, and the maximum was in 2013 with a temperature of 9.556. The global range, therefore, is 1.588. The global range is higher than the range of the Sacramento temperatures, which I speculated on by looking at the charts.

Also, I was interested in how much the temperatures have changed since the beginning of the data set. The 1858 10-year MA for Sacramento was 14.177 and 8.038 globally. The 2013 values were 15.102 for Sacramento and 9.556 globally. This is a 6.52% change for Sacramento, and a 18.89% change globally. Part of this discrepancy can be attributed to the fact that the global values (temperatures) are smaller overall, so changes will look more drastic.

To explore these observations even further, I normalized each data set and set them each on a bell curve.



Sacramento mean = 14.44
Sacramento standard deviation = 0.302

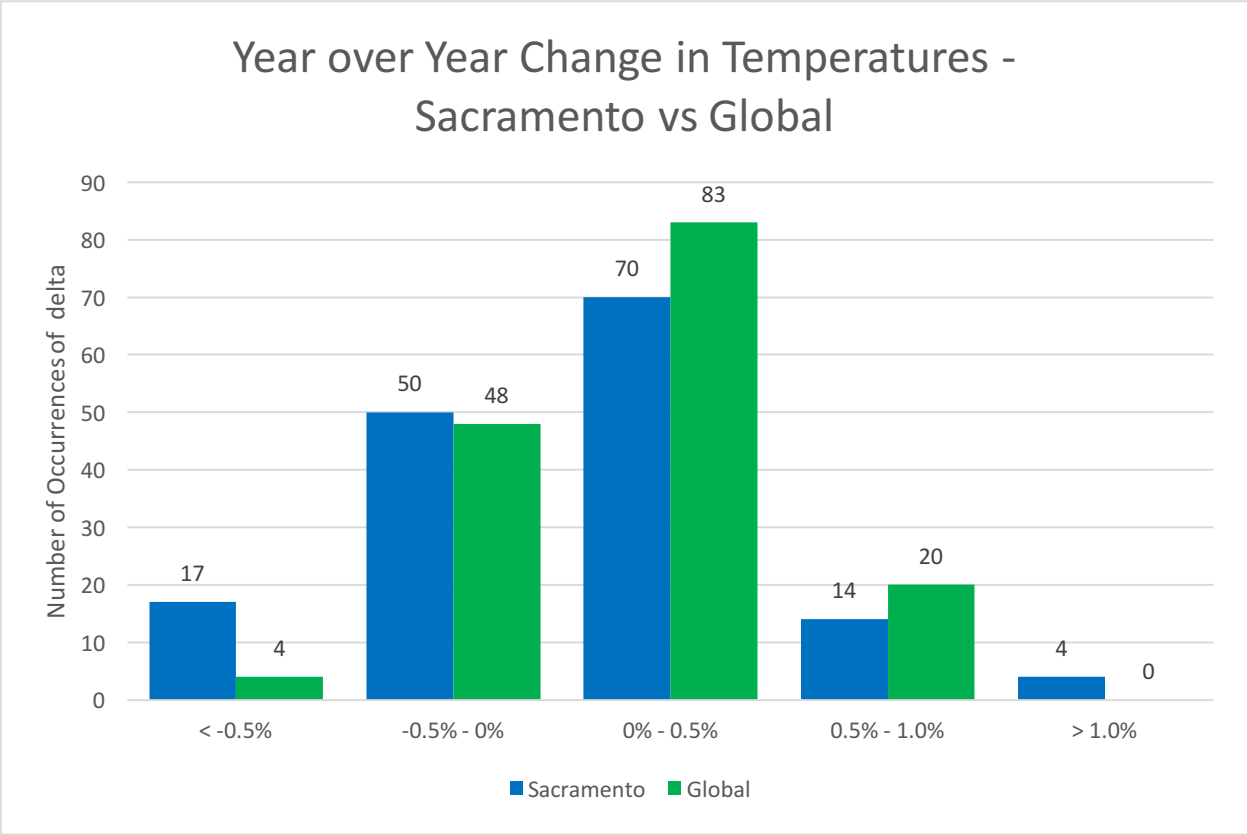


Global mean = 8.54
Global standard deviation = 0.395

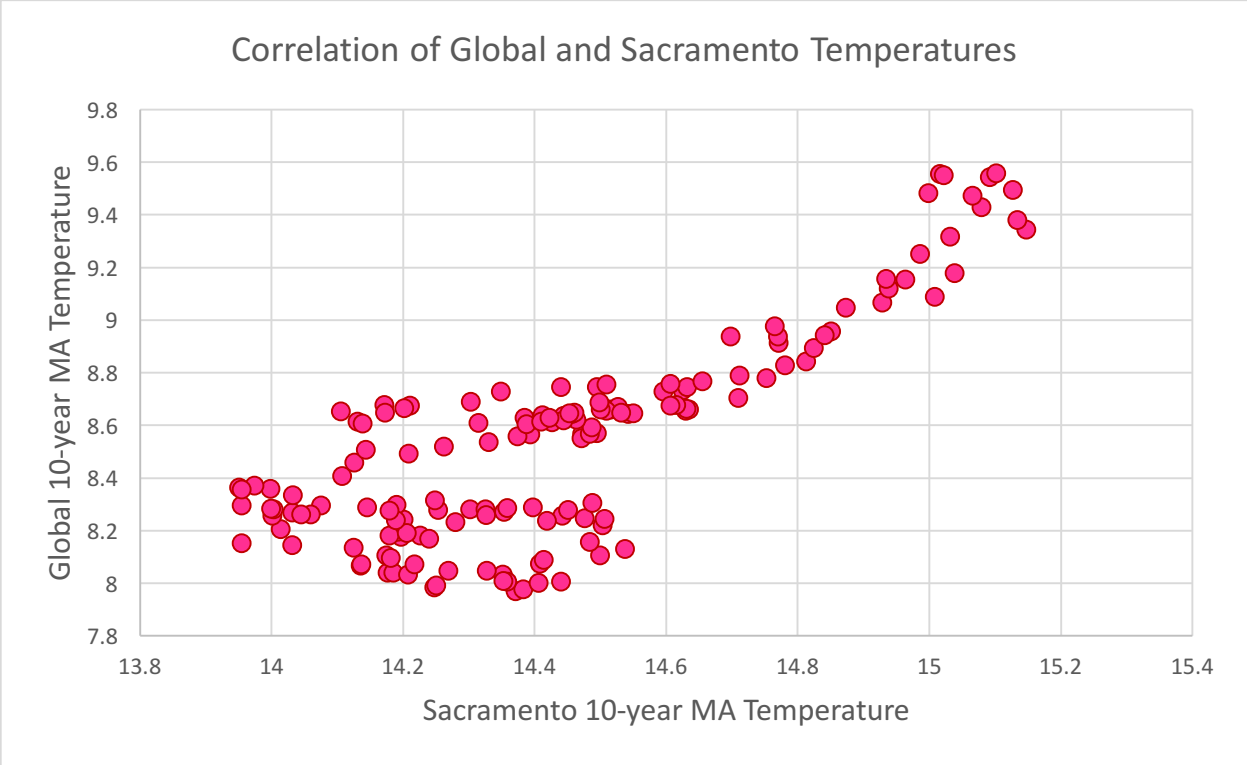
The standard deviation of the global data was larger, which indicated to me that the range of the global data is more spread out, and the range of the Sacramento data is more concentrated.

Next, I looked at the "year over year" change in the 10-year moving averages for both Sacramento and globally. I'll call these changes "deltas" for short. The deltas in the Sacramento data went from -0.89% to 1.24%, which gave a range of 2.13%. The mean was 0.04%. The deltas of the global data on the other hand ranged from -0.91% to 0.85%, giving a range of 1.76%. The mean of the global deltas was 0.11%. The range of deltas of the Sacramento data were higher than that of the global data, which confirms the volatility of the Sacramento data.

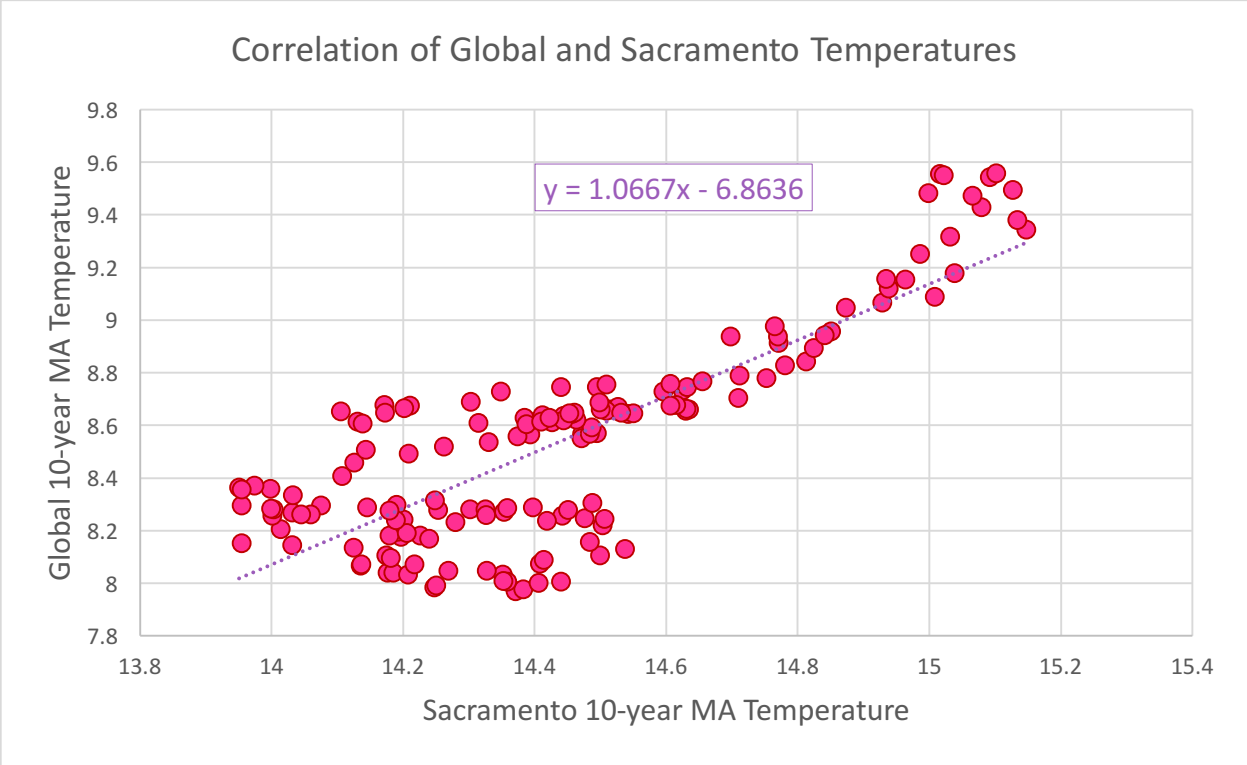
I charted both sets of these deltas with a histogram of five bins. The deltas of the Sacramento data were slightly more spread out than the frequency of the global data, which also corroborates the volatility of the Sacramento data. Said another way, the year-over-year changes in the global data were more likely to be minimal, than the year-over-year changes in the Sacramento data.



Lastly, I wanted to see how much, if at all, the global weather data was correlated to the Sacramento weather data. I started by calculating the correlation coefficient for both sets, which came to 0.816. This is a strong correlation, although not a perfect one. Here is the scatter plot of Sacramento versus global.



Given a global temperature, could I predict the Sacramento temperature, and vice versa? To do this, I needed to do a regression analysis to derive an equation relating the two data sets. This was done very simply and automatically by Excel; the below graph shows the best fit line and the corresponding equation.



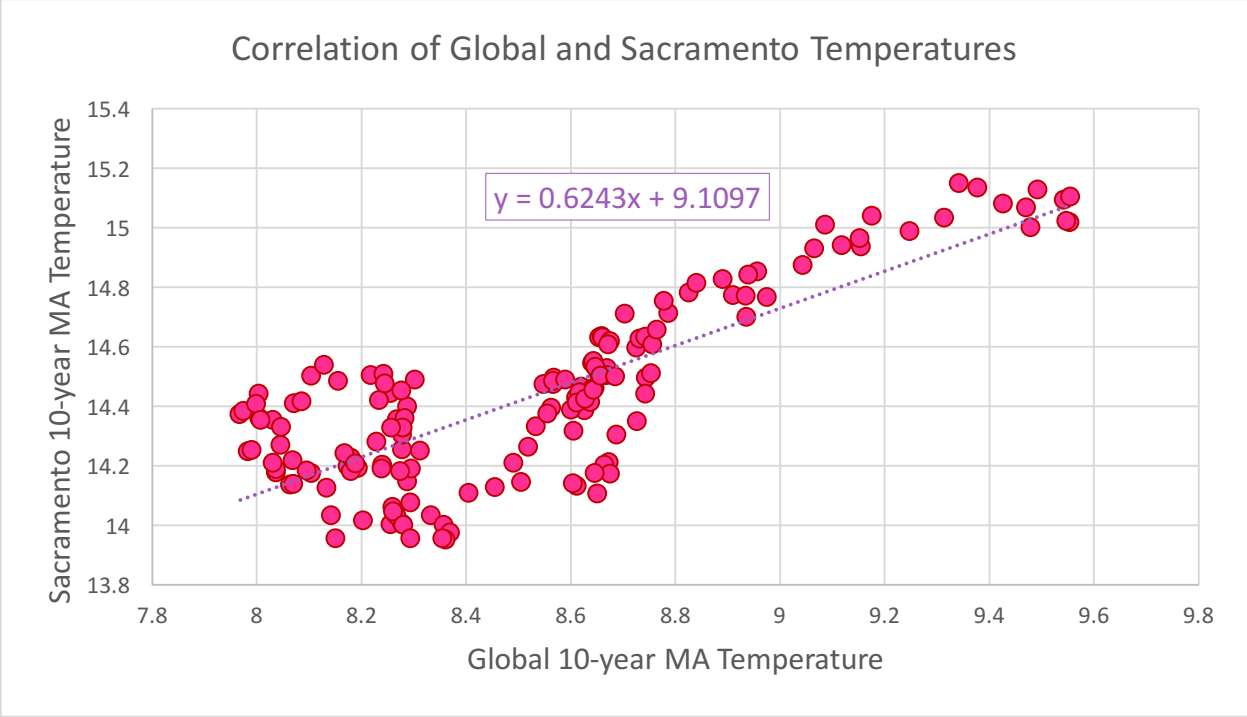
The equation is y = 1.0667x – 6.8636, where y equals the global temperature, and x equals the Sacramento temperature.

To test this, I took a selection of five random years, and their corresponding Sacramento temperatures. After I calculated the estimated global temperature, I compared that with the actual global temperature.

Year	Sacramento Temperature	Estimated Global Temperature	Actual Global Temperature	Difference
1961	14.516	8.621	8.659	0.44%
1966	14.616	8.727	8.676	-0.59%
1989	14.771	8.893	8.911	0.21%
2005	15.134	9.280	9.378	1.05%
2008	15.066	9.207	9.471	2.78%

Those results are about what is to be expected with a 0.816 correlation coefficient.

I did this whole calculation over again, this time with the global temperatures as the x value, and Sacramento temperatures as the y values. The scatter plot looks very similar, but the best fit equation has changed.



Our best fit equation is now y = 0.6243x + 9.1097, where y is the Sacramento temperature and x is the global temperature.

Year	Global Temperature	Estimated Sacramento Temperature	Actual Sacramento Temperature	Difference
1871	8.156	14.201	14.484	1.95%
1874	8.288	14.284	14.398	0.79%
1901	8.256	14.264	14.002	-1.87%
1999	9.156	14.826	14.935	0.73%
2000	9.153	14.824	14.964	0.94%

Again, this is about what we would expect, and not a bad estimation.

In conclusion, I confirmed that yes, since 1849, Sacramento temperatures didn't increase as much as global temperatures did. Also, the change in the 10-year average temperature for Sacramento was more extreme than the same globally. Given the temperature of Sacramento, one could estimate the global temperature, and vice versa, within a reasonable error of margin.