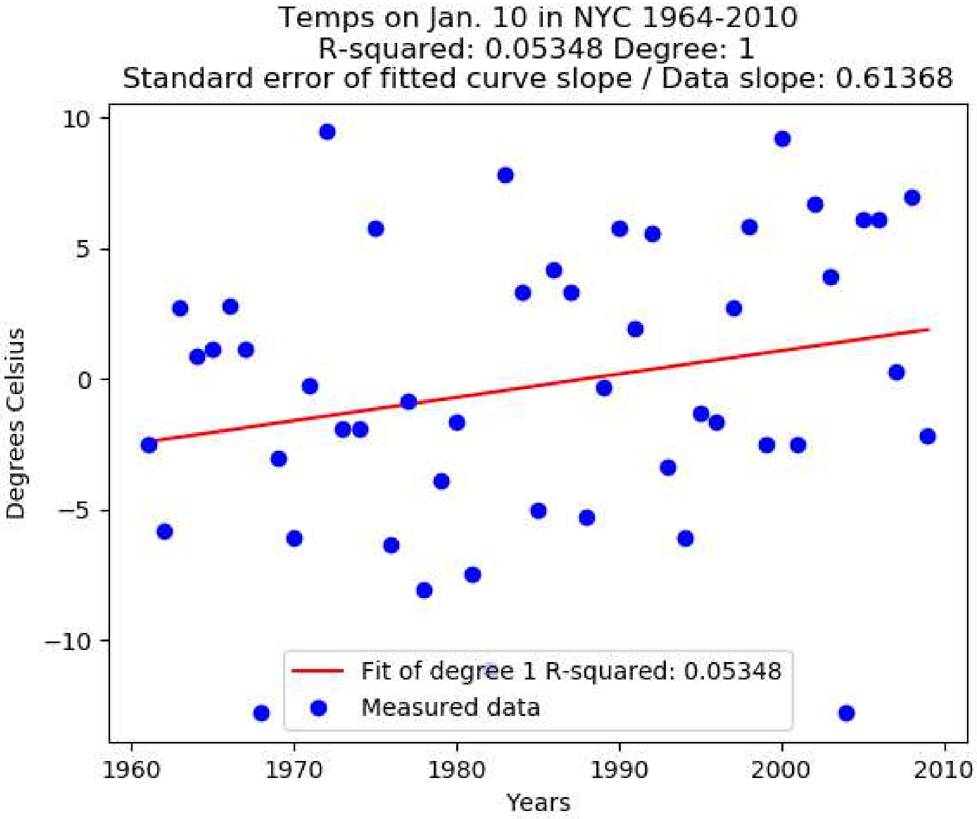
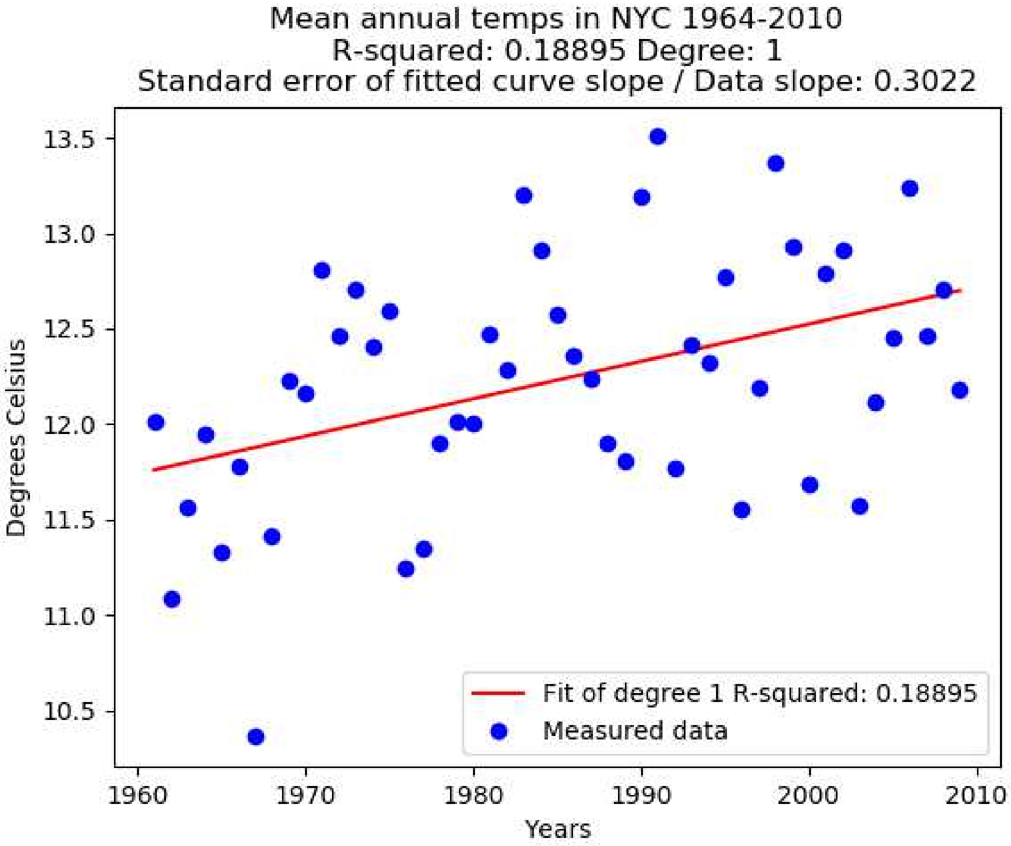
Ethan Fulbright

6.0002 – Problem set 5

A 4.1



A 4.2



A4.1 / A4.2 writeup

What difference does choosing a specific day to plot the data for versus calculating the yearly average have on our graphs (i.e., in terms of the R^2 values and the fit of the resulting curves)? Interpret the results.

Using the yearly average seems to smooth out some of the noise in the data, resulting in a higher R^2 and lower standard error to slope ratio. As a result, the resulting curves are a tighter fit (although, due to the high noise, the R^2 value is still quite low). This is because the yearly data consists of mean values, which results in statistical outliers influencing the overall data less.

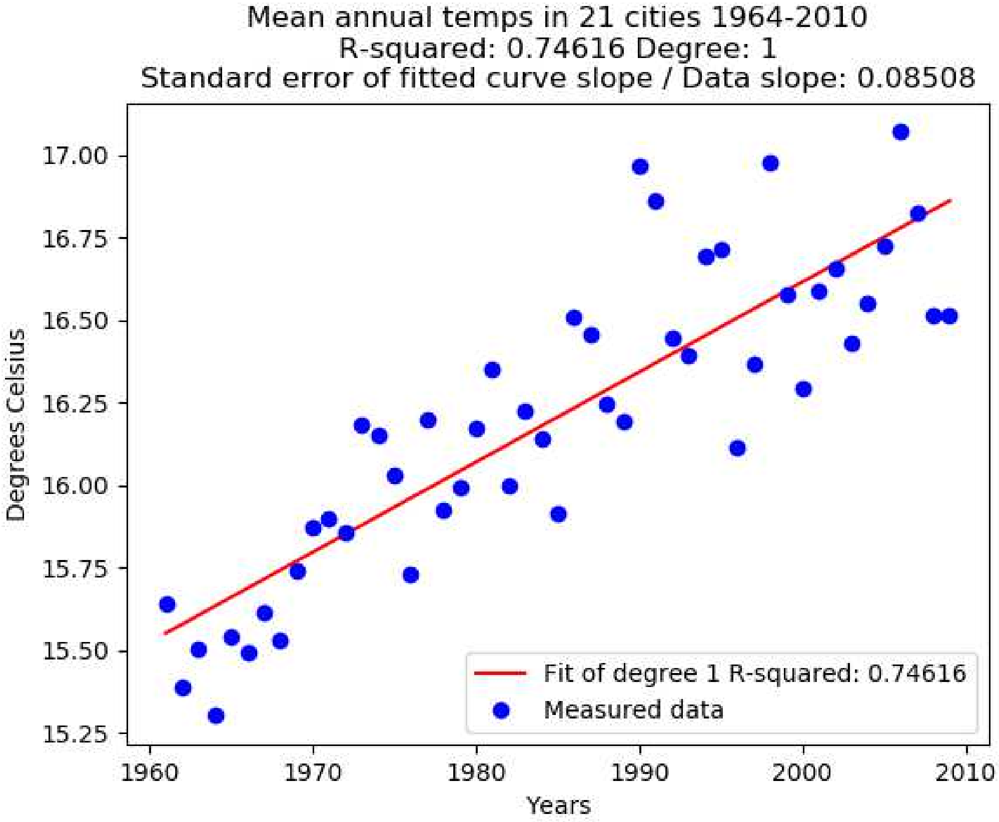
Why do you think these graphs are so noisy? Which one is more noisy?

The daily data is much noisier than the annual mean data. I believe the noise is present for a variety of reasons, including experimental and equipment variability as well as the fact that weather is extremely variable and subject to wide change based on many different factors. The mean annual data smooths out some of this noise because taking the mean for an entire year removes some of the outlier data points. Another reason for the observed noise is that the Y-axis values are relatively small – although the graph appears noisy, it's important to note that the actual variation in the annual data is less than 3° Celsius.

How do these graphs support or contradict the claim that global warming is leading to an increase in temperature? The slope and the standard error-to-slope ratio could be helpful in thinking about this.

They definitely support the idea that global warming is leading to an increase in temperature – the fitted curves in both cases show a significant upward slope, indicating that temperatures are increasing. Although the R^2 values are low, the standard error to slope ratio (i.e., the mean distance between the data points and the fitted curve) is quite low in both cases, meaning that the curve is a relatively good fit for the provided data points (since the average distance those data points lie from the fitted curve is less than 1° Celsius in both graphs).

B – Average 21 cities



B writeup

How does this graph compare to the graphs from part A (i.e., in terms of the R^2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.

This graph also supports the claim regarding global warming. In comparison to the graphs from Part A, however, the R^2 value of this graph is much higher, and the fit of the resulting curve is also better. The standard error to slope ratio is also lower (i.e., better) than the Part A graphs. In general, this graph presents much more compelling evidence to support the claim of increasing temperature.

Why do you think this is the case?

This graph, by taking the mean over 21 cities in different parts of the United States smooths out even more of the noise and statistical variance in the data, allowing us to more clearly spot the upward trend and fit a curve to it. In the same way that taking the mean temperature for a given year for each city smooths out the effect of random outliers in that city's data, taking the mean across 21 cities smooths out the outliers within the yearly data.

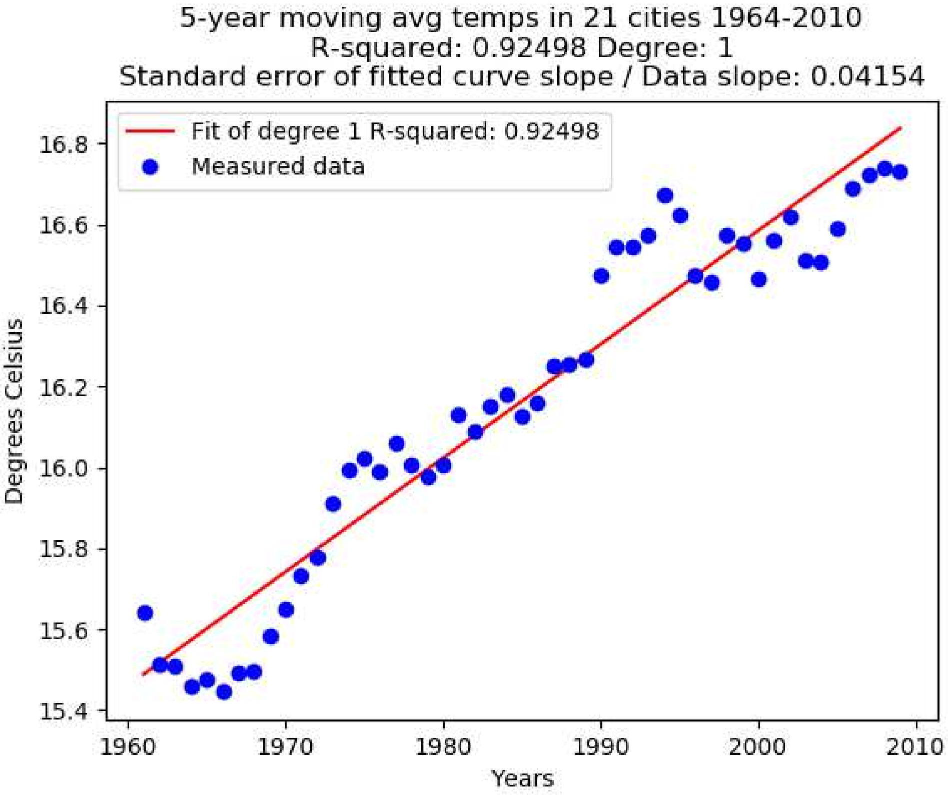
How would we expect the results to differ if we used 3 different cities? What about 100 different cities?

If only 3 cities were used, more noise would show up in the graph (since we're taking the mean across fewer cities), and we could expect a lower R^2 value and a higher standard error to slope ratio. If we used 100 cities, we would expect to see even less noise, with an even better fit (i.e., a higher R^2 value and lower standard error to slope ratio).

How would the results have changed if all 21 cities were in the same region of the United States (for ex., New England)?

In the same way that using fewer cities would reintroduce noise to the graph, using cities that are all in the same region would also produce a noisier graph with a worse fit. This is because the data would be more subject to local variance in the weather that would have been smoothed out over a larger geographical sample.

C – 5-year moving average



C writeup

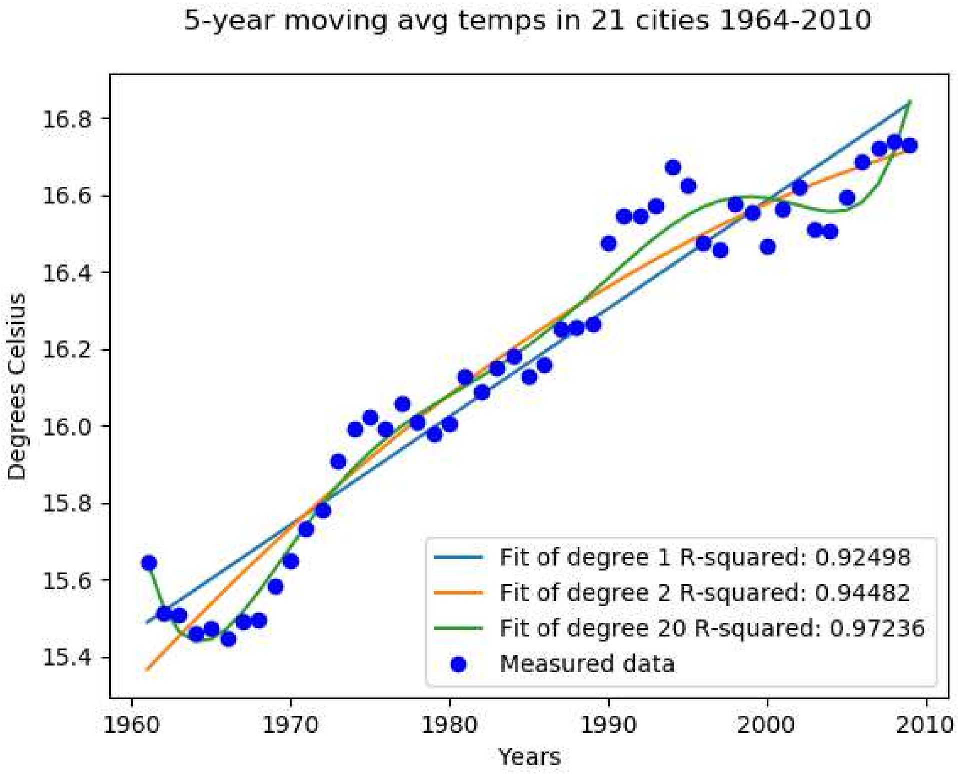
How does this graph compare to the graphs from part A and B (i.e., in terms of the R^2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.

This graph presents even more compelling evidence for the claim of increasing temperature. Its R^2 value is VERY high, and the standard error to slope ratio is even better (that is, it's lower). The resulting curve is a much better fit than the curves from Parts A or B, and shows an unmistakable upward trend.

Why do you think this is the case?

Again, this graph smooths out even more noise from the data by using a 5-year moving average instead of annual averages. Just as using multiple cities in different parts of the United States and using mean data for each city removes the impact of experimental error and random noise, using a 5-year moving average removes the impact of statistically outlying years as well.

D – 2.1 – Training



D – 2.1 writeup

How do these models compare to each other?

Although they all communicate the upward trend, their R^2 values are different, increasing as the degree of the model increases.

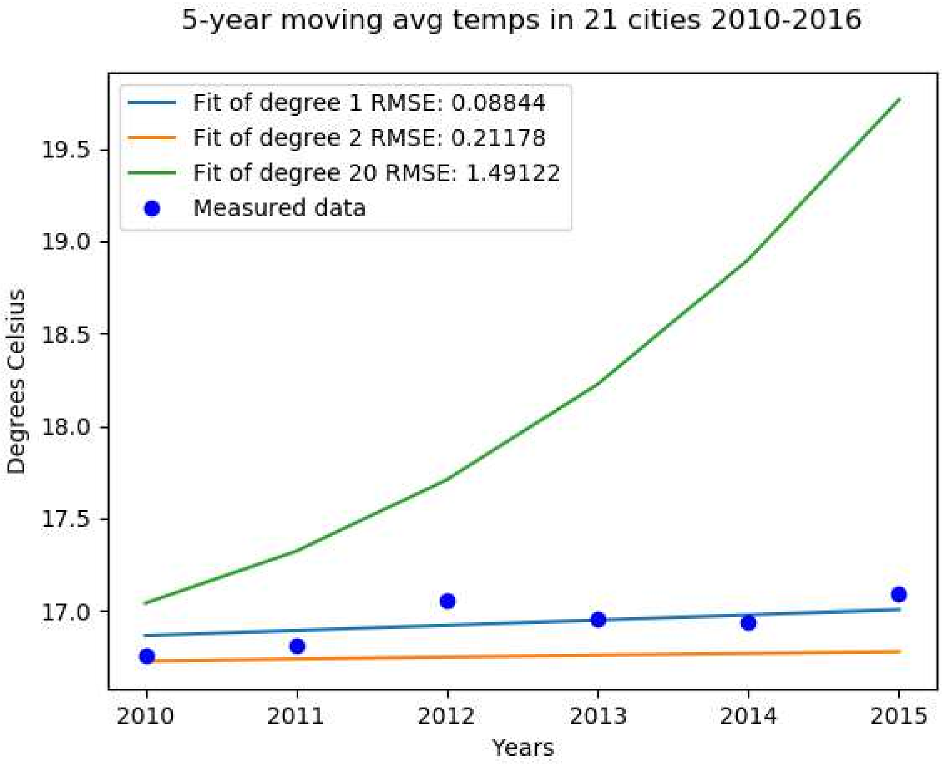
Which one has the best R2 ? Why?

The degree 20 model has the best R^2 value, because it adheres very closely to the data points. In fact, it is over-fitted. It's plotting the noise in the data as well as the trends.

Which model best fits the data? Why?

The degree 1 model best fits the data – although it has the lowest R^2 value ("lowest" being relative, as its R^2 is still 0.92), it avoids overfitting the data points and clearly shows the upward trend. It should also have the best predictive value.

D – 2.2 – Testing



D – 2.2 writeup

How did the different models perform? How did their RMSEs compare?

The degree 1 and degree 2 models predicted results which were close to the data. The degree 20 model failed miserably to predict accurate values. The degree 1 model had the lowest RMSE, followed by the degree 2, then the degree 20.

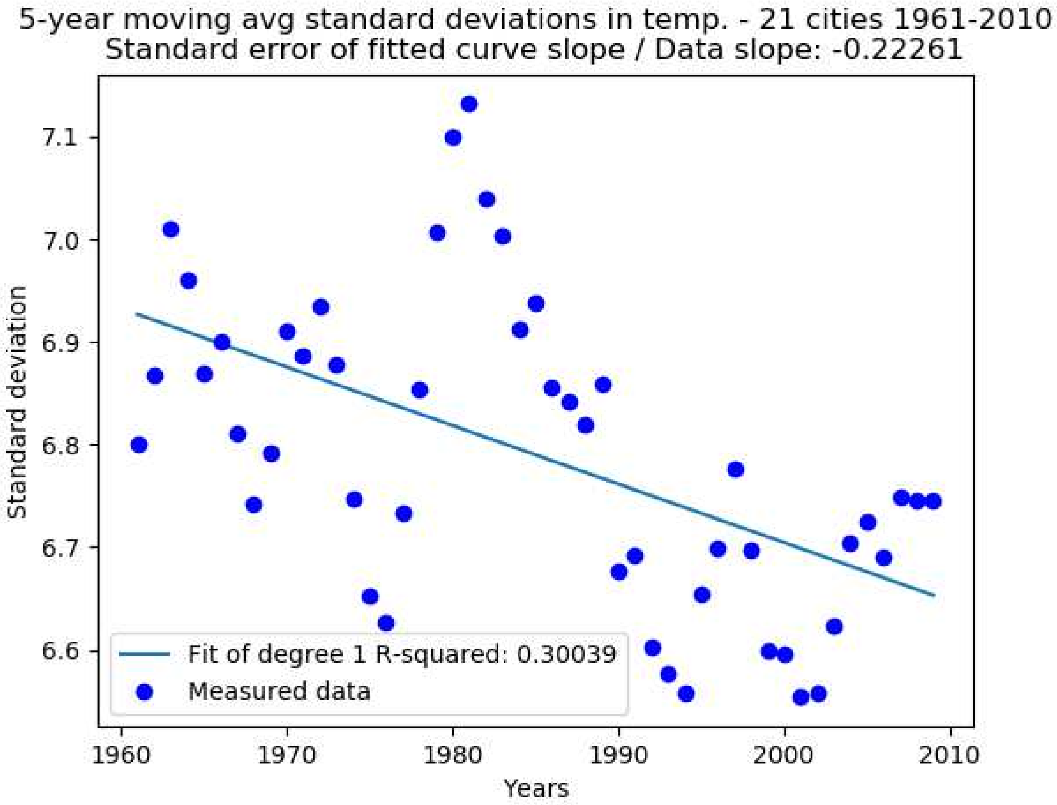
Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?

The degree 1 model clearly performed the best – its RMSE was the lowest, and its predicted values were closest to the observed data. The degree 20 model was the worst. I would argue that the best-performing models from D.2.1 are the same (i.e., the degree 1 model is the best), simply because it avoided overfitting, despite having the lowest R^2 *value.*

If we had generated the models using the A.4.II data (i.e. average annual temperature of New York City) instead of the 5-year moving average over 22 cities, how would the prediction results 2010-2015 have changed?

There would be more variance in the models' predictions because there was more noise in the training data. As a result, the model would most likely perform less well, predicting values farther from the data and having worse RMSE values.

E – extreme temperatures



E – writeup

Does the result match our claim (i.e., temperature variation is getting larger over these years)?

No. The model shows a distinct downward slope, although its R^2 value is abysmally low. This may be due to noise in the data.

Can you think of ways to improve our analysis?

Yes. Since we're looking for extremes of temperature, averaging the temperatures for a given day will tend to smooth out the extremes we're looking for. We might see better results if, instead of averaging the temperatures for a given day across all cities, we computed the standard deviation in temperature for a given day across all cities. Then, we could store the average standard deviation value for a given year and generate models based on those data points. This might show us the variance in the data more clearly.