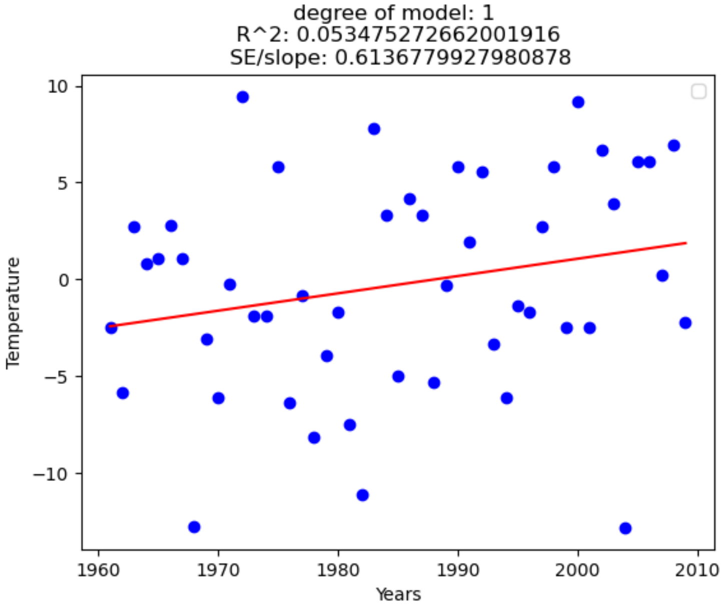
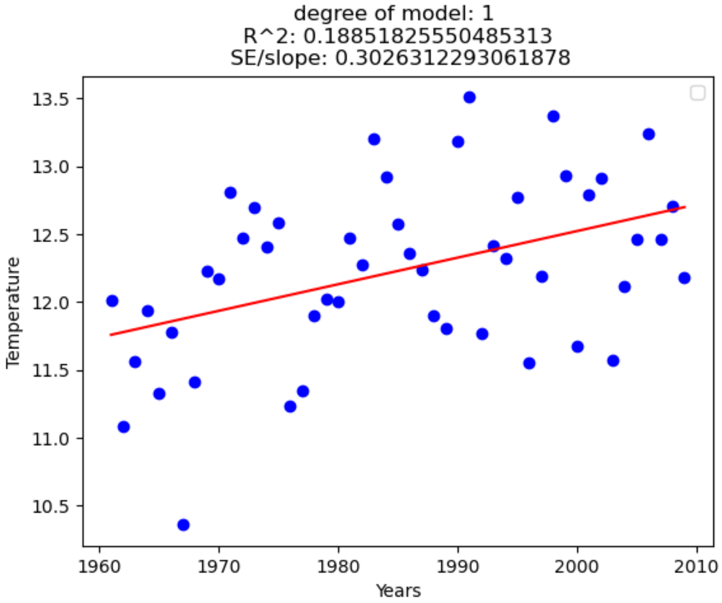
Alexander Kahanek – INFO 4501 – 4/13/2021

PART A

Plot for part A.4, sec: 1

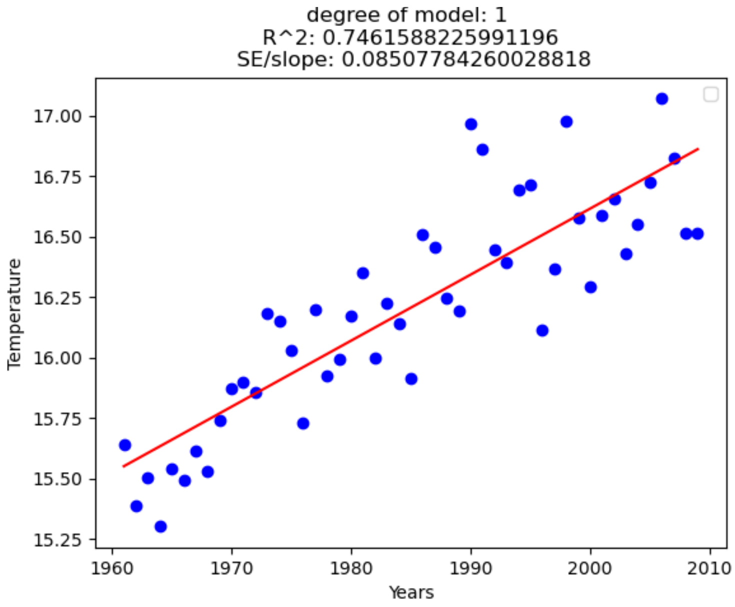


Plot for part A.4, sec: 2



- 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 R2 values and the fit of the resulting curves)? Interpret the results.
    - The difference between the two methods are there are more daily variations, as opposed to yearly variations. This leads to a better R2 value for the yealy average, as it lowers the overall variance, which is what R2 measures.
- Why do you think these graphs are so noisy? Which one is more noisy?
    - These graphs are noisy because there is a lot of variance in the temperatures. Especially when looking solely at one day, the temperature can vary much more than the general yearly average temperature. Thus, the graph for January 10th is nosier.
- 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.
    - Both graphs suggest that the earth is getting warmer as time goes on. However, as global warming is a much more complex issue, this alone does not tell us that the increasing temperatures are due to outside sources, nor does it give inclinations on if the earth will start to decline in temperature. However, with the plot for A.4 section 2, we can claim that about 18% of our yearly average temperature are positively affected by the years. i.e., we have a loose assumption that the yearly temperature is rising from 1961 to 2009 in New York.

PART B

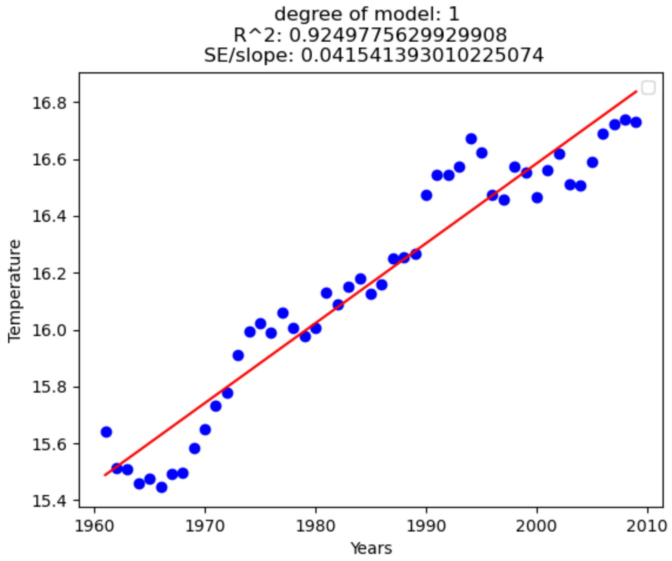

degree of model: 1
R^2: 0.7461588225991196
SE/slope: 0.08507784260028818

- How does this graph compare to the graphs from part A (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
    - The R2 value for this graph is much higher. This tells us that the data has much less variance around its mean. We can attribute about 74% of the increasing temperature

averages to the increasing years. This graph helps support the claim of rising temperatures from 1961 to 2009 in the given 21 cities.

- Why do you think this is the case?
  - We added 21 cities to the yearly temperature averages. This helped to further reduce variance across our data by aggregating many results.
- How would we expect the results to differ if we used 3 different cities? What about 100 different cities?
  - Based on these results, if we used 3 cities, we would expect the R2 value to drop, and conversely if we used 100 cities we would expect it to rise. However, this would heavily depend on which cities we are adding. Especially as these chosen cities are very cherry picked, to give specific results.
- How would the results have changed if all 21 cities were in the same region of the United States (for ex., New England)?
  - It completely depends on the temperatures of the area. If we assume the rising temperatures is a constant, then it would just shift our curve up or down, depending on if the average temperature is colder or warmer in that area. However, if our R2 value changed, it would weaken the claim as it shows there is bias in the areas chosen between specific regions and the current cities.

PART C

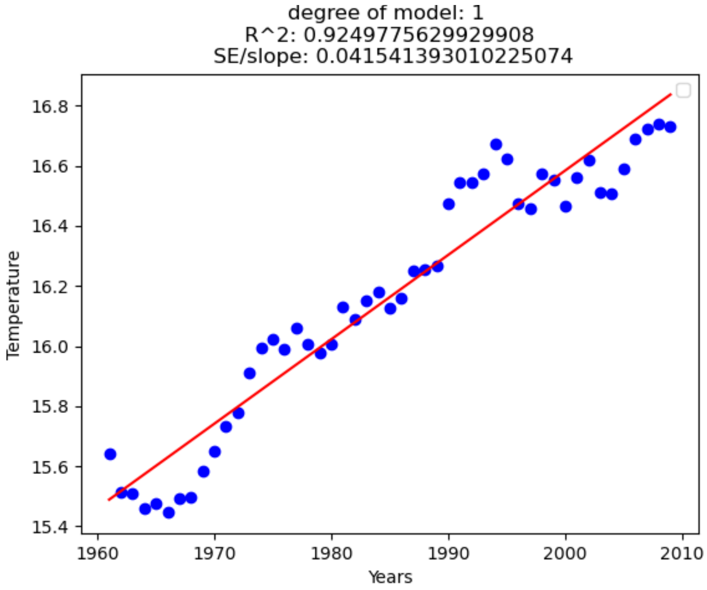


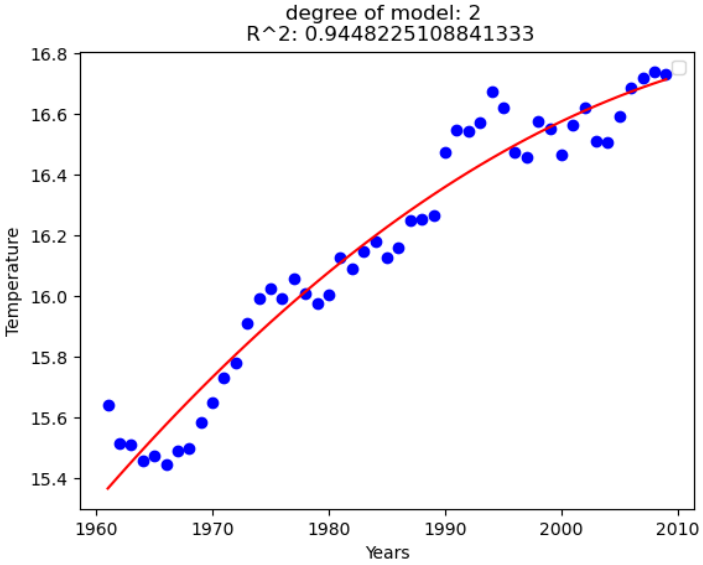
- How does this graph compare to the graphs from part A and B (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
  - This graph has much less variance in its temperatures, which is expected as we are using a rolling average. The R2 value of this is much higher, showing that 92% of our rolling yearly temperature averages, with a window of 5 years, are increasingly affected by the increasing years. Again, this helps support a trend of rising temperatures.
- Why do you think this is the case?

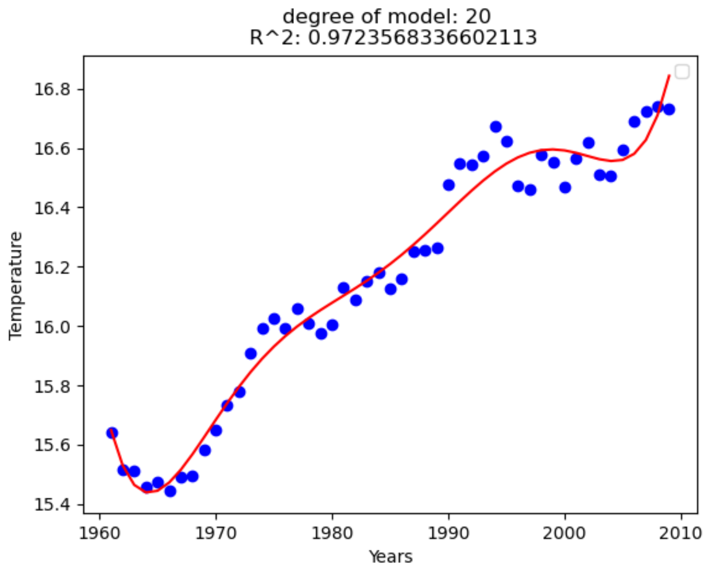
- Again, a rolling average was used, reducing the variance drastically. Effectively, you are averaging the past 5 years for each given year, further reducing the amount of variance that can occur. This method is quite biased to be using R values, in my opinion.

PART D

Plots for part D.2 sec: 1

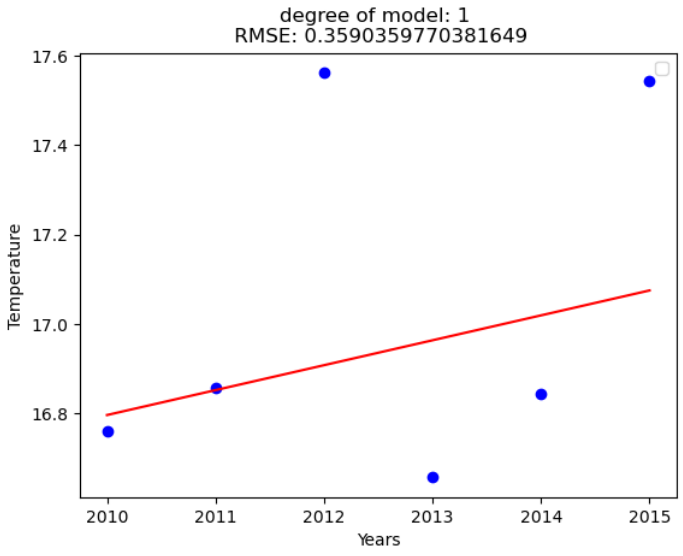


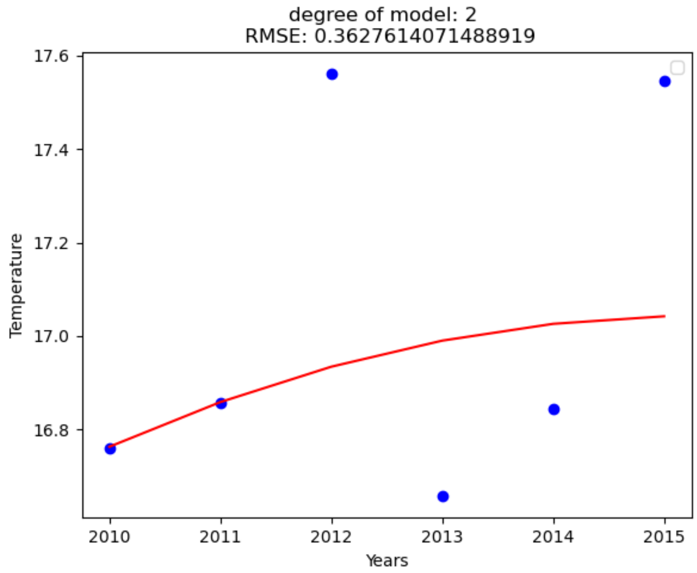


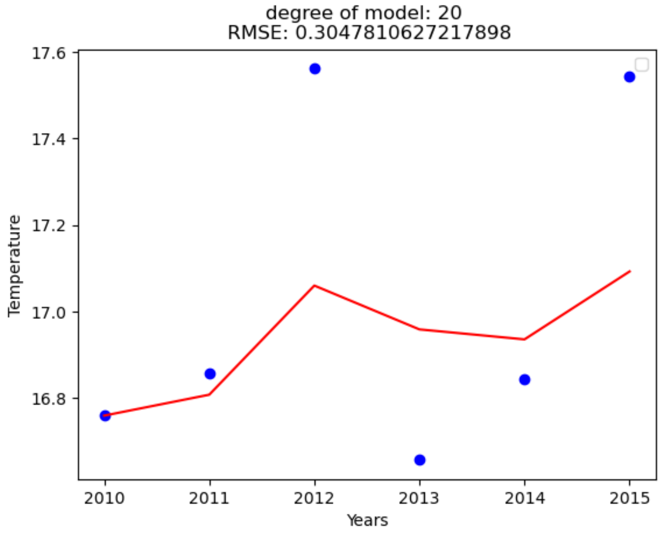


- How do these models compare to each other?
  - Here, the data does not change, but the fit of our curve does. As we add more degrees, we allow the line, or curve, to better fit our dataset of average temperatures. Thus, as we add more degrees, the more we fit our line to become closer to the actual data points, reducing our R2 value.
- Which one has the best R2? Why?
  - The degree 20 model has the best R2 values, for the testing data, because it allows the mean to stay closer to the actual datapoints.
- Which model best fits the data? Why?
  - The degree 20 model best fits the data, as it gives the best curve to fitting our training data.

Plots for part D.2 sec: 2



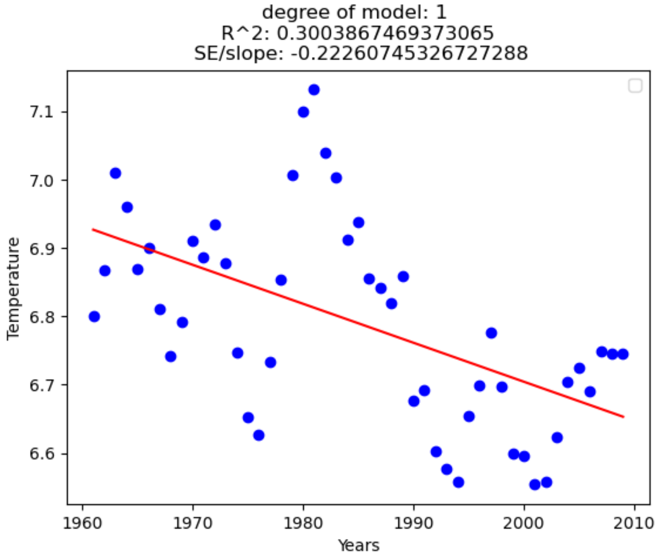




- How did the different models perform? How did their RMSEs compare?
  - These models well, with a RMSE consistently under 0.4. This means that our training data curve is a pretty good predictor for our testing data.
- Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?
  - The model that performed the best was the degree 20, while the worse was degree 2. These are not the same as the training data runs, as the R2 for the training models was ordered 20, 2, 1 from best to worst. While the testing models have RMSE values ordered as 20, 1, 2 from best to worst. This could be due to overfitting, or just that a degree 2 polynomial is not as strong as the degree 1. Or it could be due to random chance of life for our temperatures, as we do not much testing data.

- 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?
  - All these models would perform much worse, if we had used the average annual temperature of NY. This is because the overall variance was much higher for the NY temperatures, which would have made curve fitting really difficult. This would have in turn made the prediction results harder for 2010-2015, especially if they were not prediction on solely NY.

PART E



- Does the result match our claim (i.e., temperature variation is getting larger over these years)?
  - The result doesn’t support any claim of rising temperatures. All this graph shows is that the standard deviation is getting lower through the years. This effectively means that the data points are getting closer to their respective means, or that the variance of temperature tends to decrease throughout the years. This doesn’t tell us which direction the temperature is going, meaning we have no idea if the temperature is rising, lowering, or staying constant.
- Can you think of ways to improve our analysis?
  - I would add more cities. Cherry-picking 21 cities shows bias, especially when all the cities are US-based and we are attempting to make a claim on a global scale. It is inherently flawed.