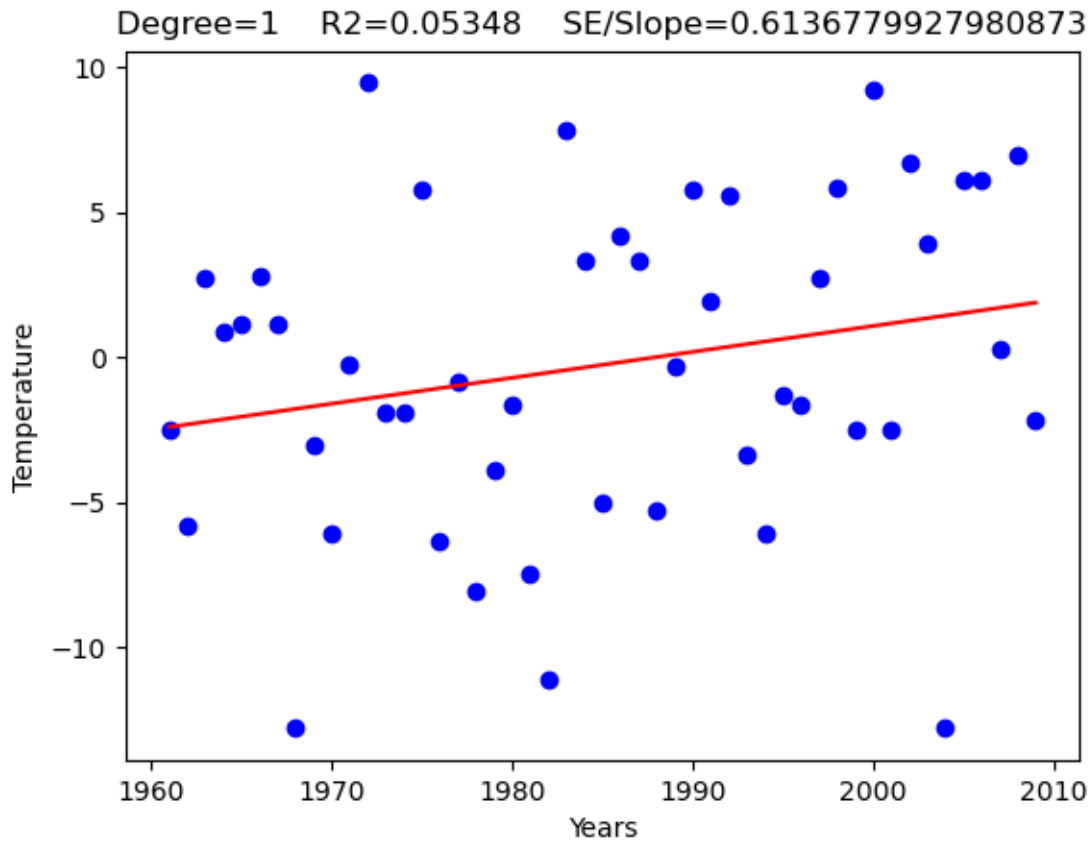
Student: Hayk Stepanyan

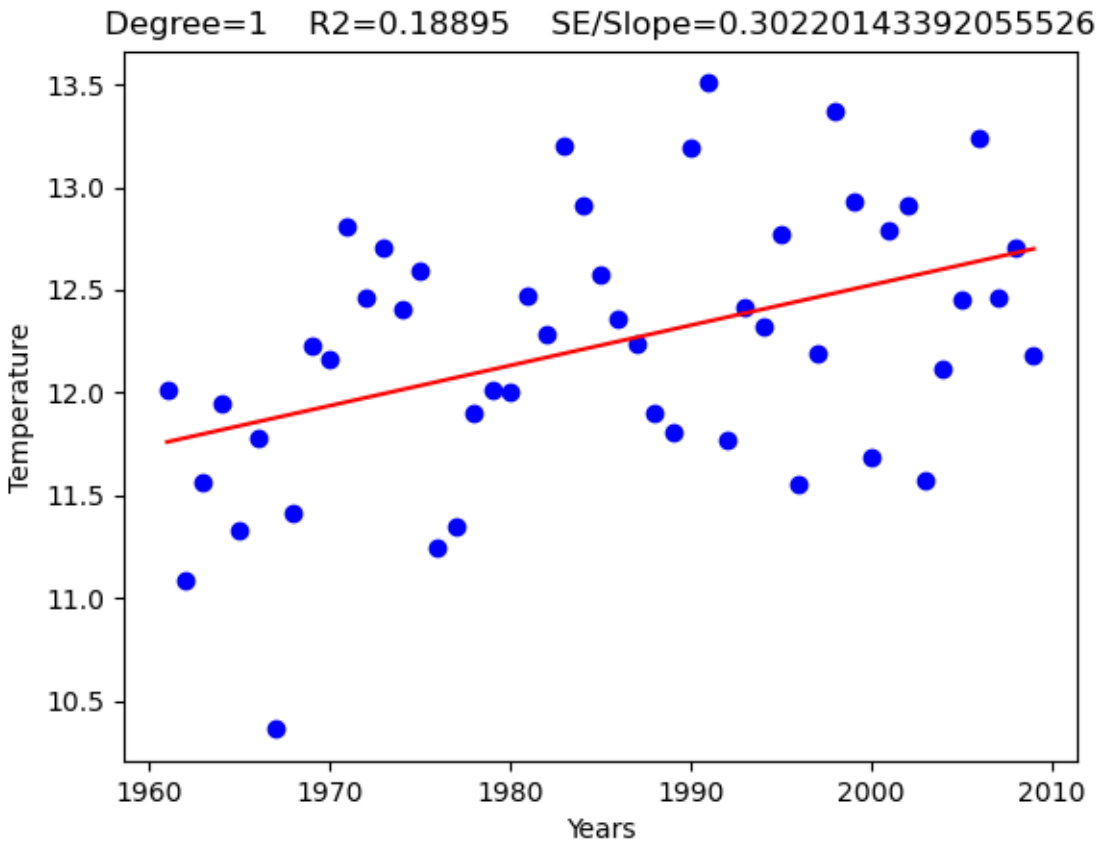
Created on July 23, 2020

Part A

Temperature on Jan 10 in New York during 1961-2009 years



Temperature in New York during 1961-2009 years



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.

Choosing the yearly average makes R2 become larger and SE/slope become smaller, which means that the model representing the yearly average fits degree-one form better.

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

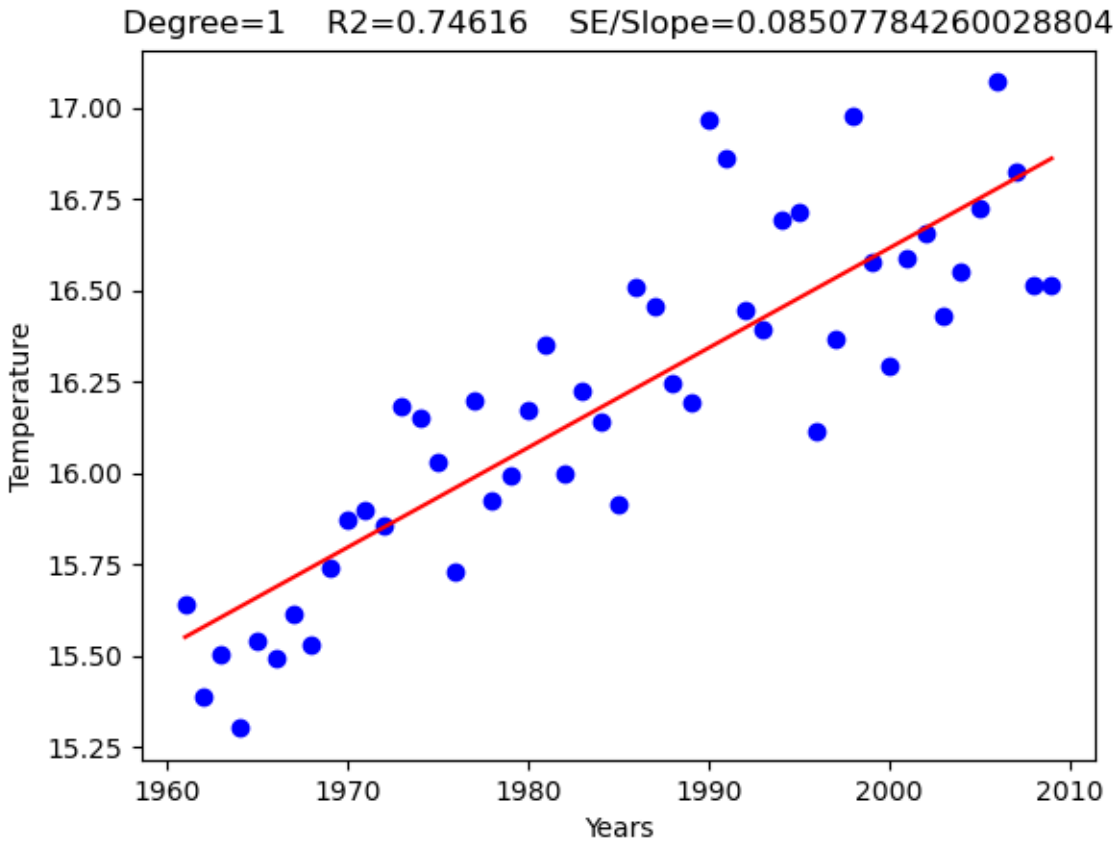
The second graph is noisy because it does not represent the whole country only New York is taken into consideration. Moreover, the first graph is more more noisy, because despite taking data from New York only, it takes only from a single day rather the average from the whole year.

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.

These graphs partially support the claim, because the slope is positive (meaning that temperature is directly proportional to the year), however, it isn't big enough to make final statement. Unlike the first graph, error-to-slope value for the second graph is acceptable (<0.5).

Part B

Yearly average temperatures over 21 cities during 1961-2009 years



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.

This graph is a better fit for degree-one model compared to the ones in part A, because the R2 value is bigger and and SE/slope is smaller. The graph supports our claim about global warming.

Why do you think this is the case?
The slope of the graph is big enough to state that there is a direct relationship between years and temperature. Hence, as the year becomes bigger the average temperature rises, which is what we claimed initially.

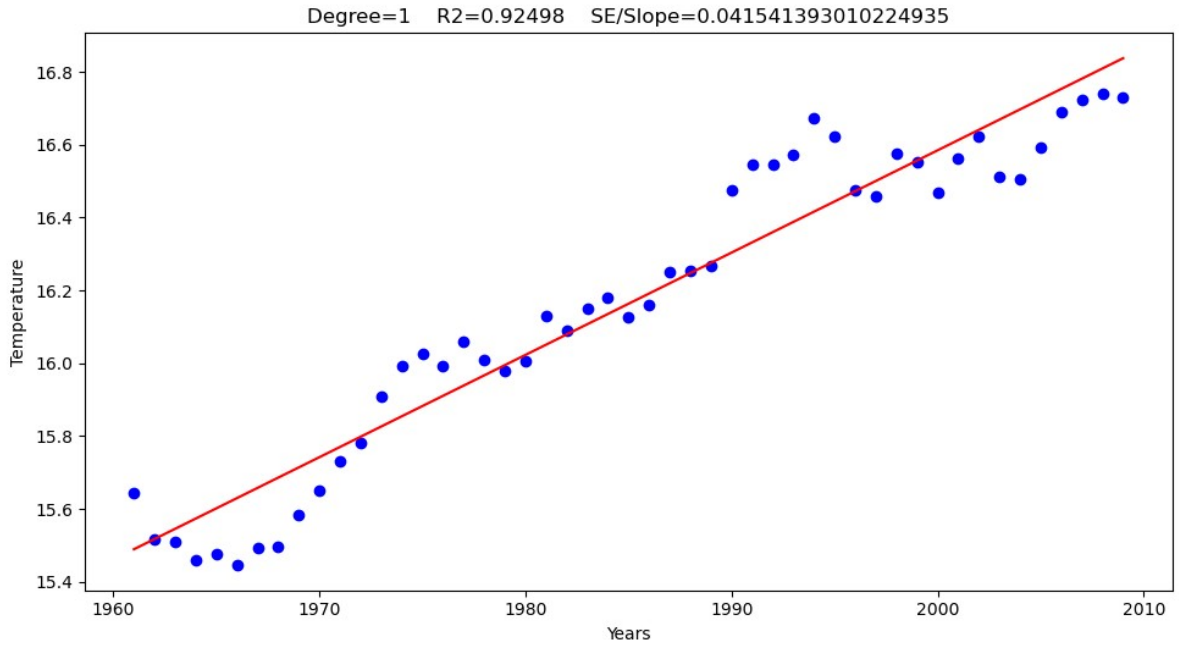
How would we expect the results to differ if we used 3 different cities? What about 100 different cities?
Choosing 3 different cities will make the graph noisy, because that would not be able to provide efficient data for modeling temperature throughout the country. In contrast, 100 different cities will add data, which will make the graph smoother and less noisy.

How would the results have changed if all 21 cities were in the same region

of the United States (for ex., New England)?
The graph would show only the result of a particular region, which will not be a prediction for the whole US. Also, having less data will result noisy graph.

Part C

Moving average over 21 cities during 1961-2009 years



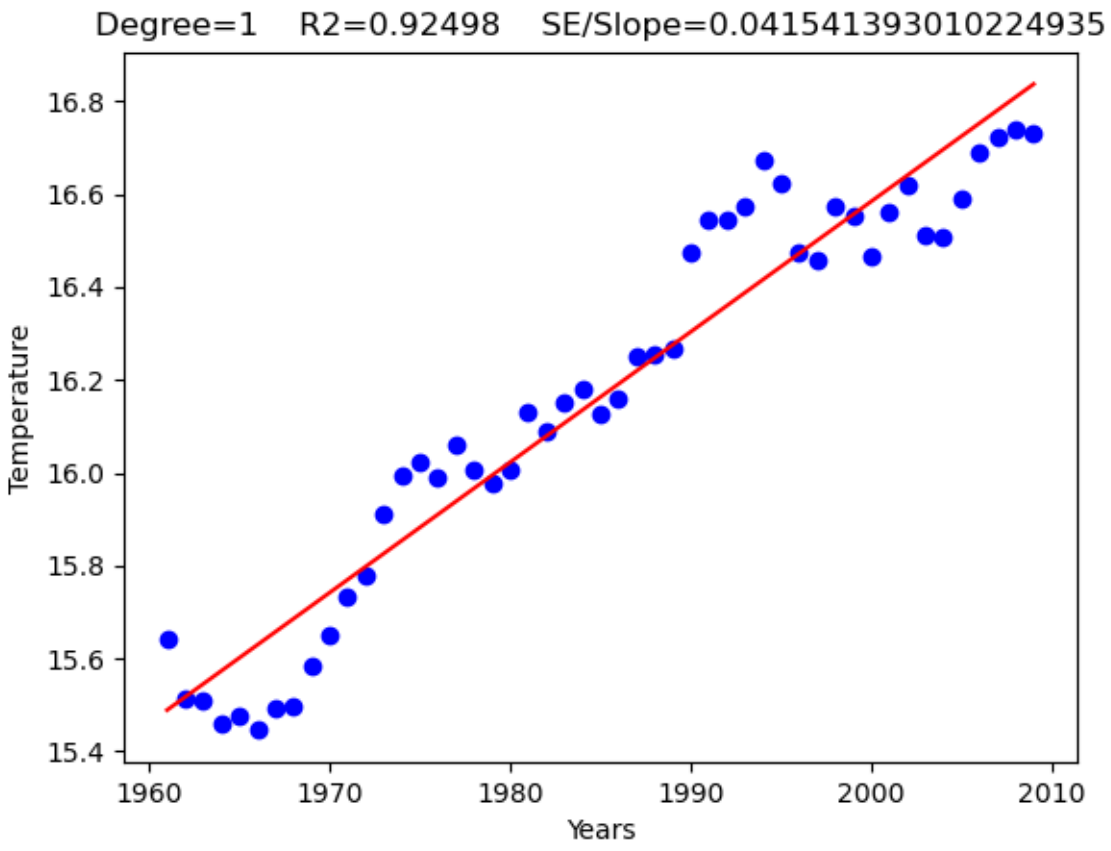
How does this graph compare to the graphs from part A and B (i.e., in terms of the R 2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
The results of the graph in part C are much better than those in part A and B. R2 is significantly larger and the SE/slope is very small, which leads to a almost-perfect degree-one feet. Hence, the graph supports our claim about global warming.

Why do you think this is the case?
Taking 5 year average globalizes the data over yearly fluctuations, which shows the data over 21 cities in a more general way. This is why the linear relationship and our claim is approved.

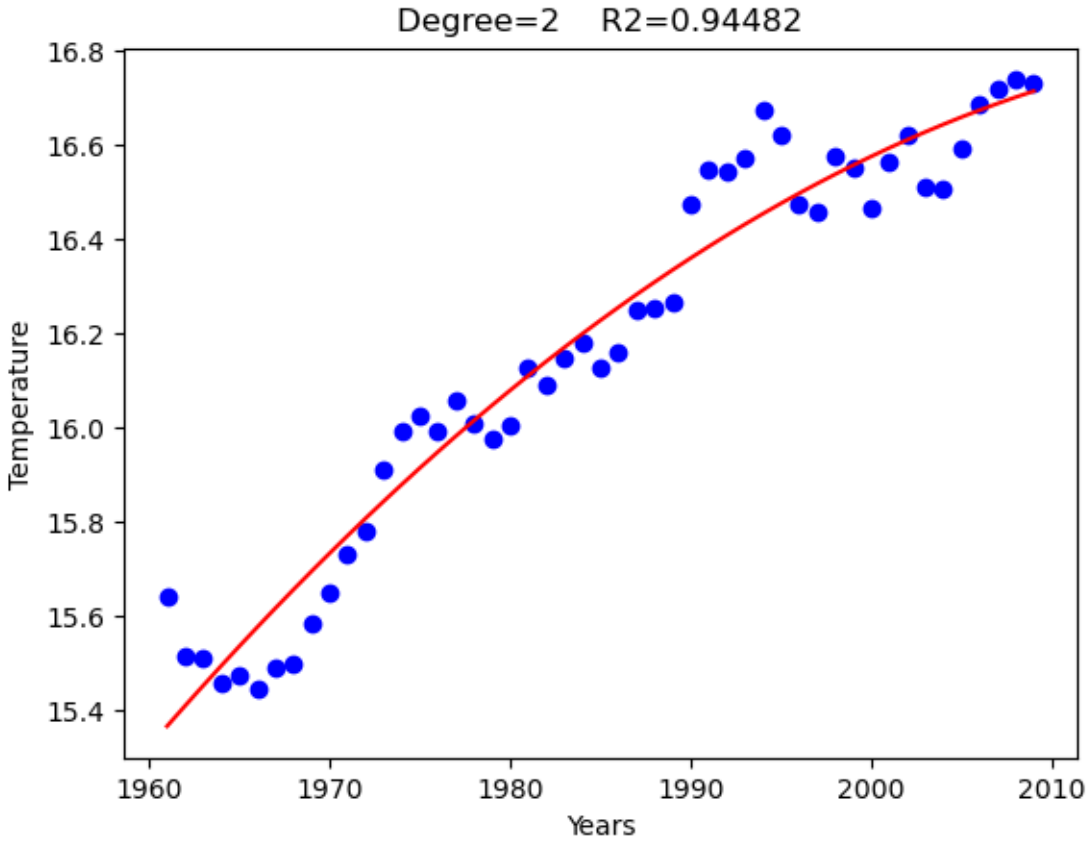
Part D

Generating more models

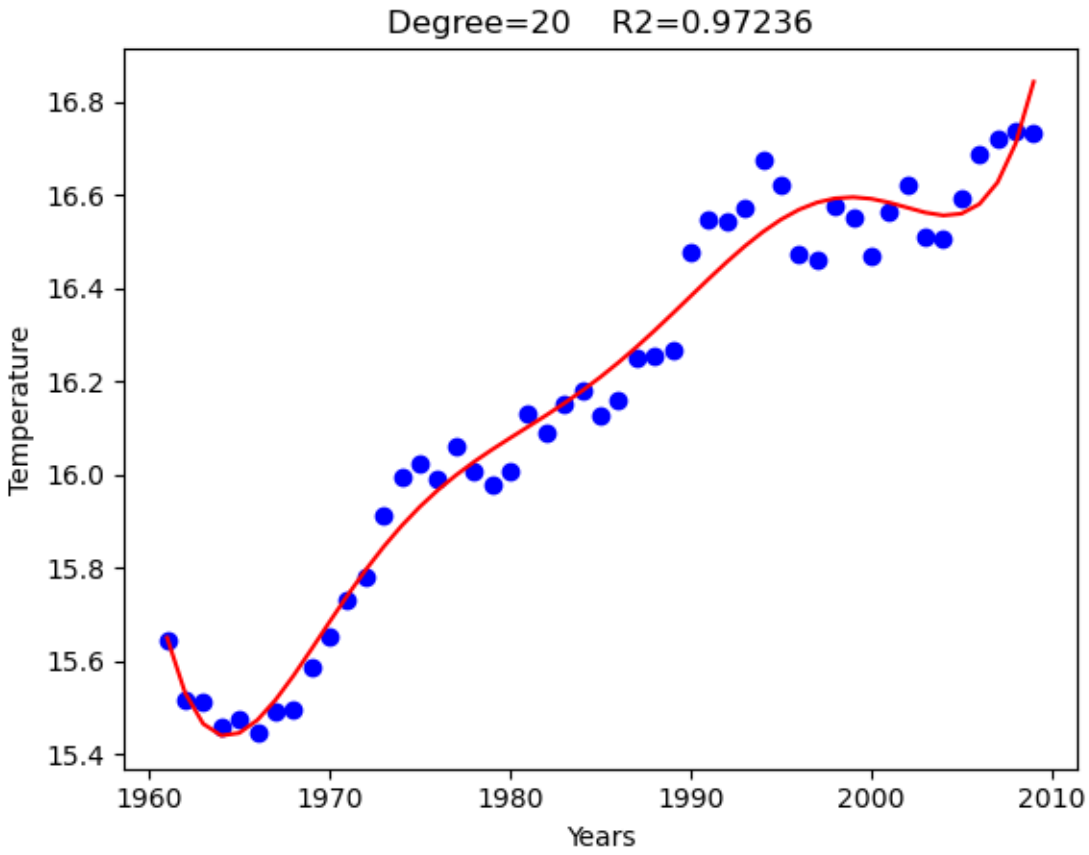
Moving average over 21 cities during 1961-2009 years: degree 1



Moving average over 21 cities during 1961-2009 years: degree 2



Moving average over 21 cities during 1961-2009 years: degree 20

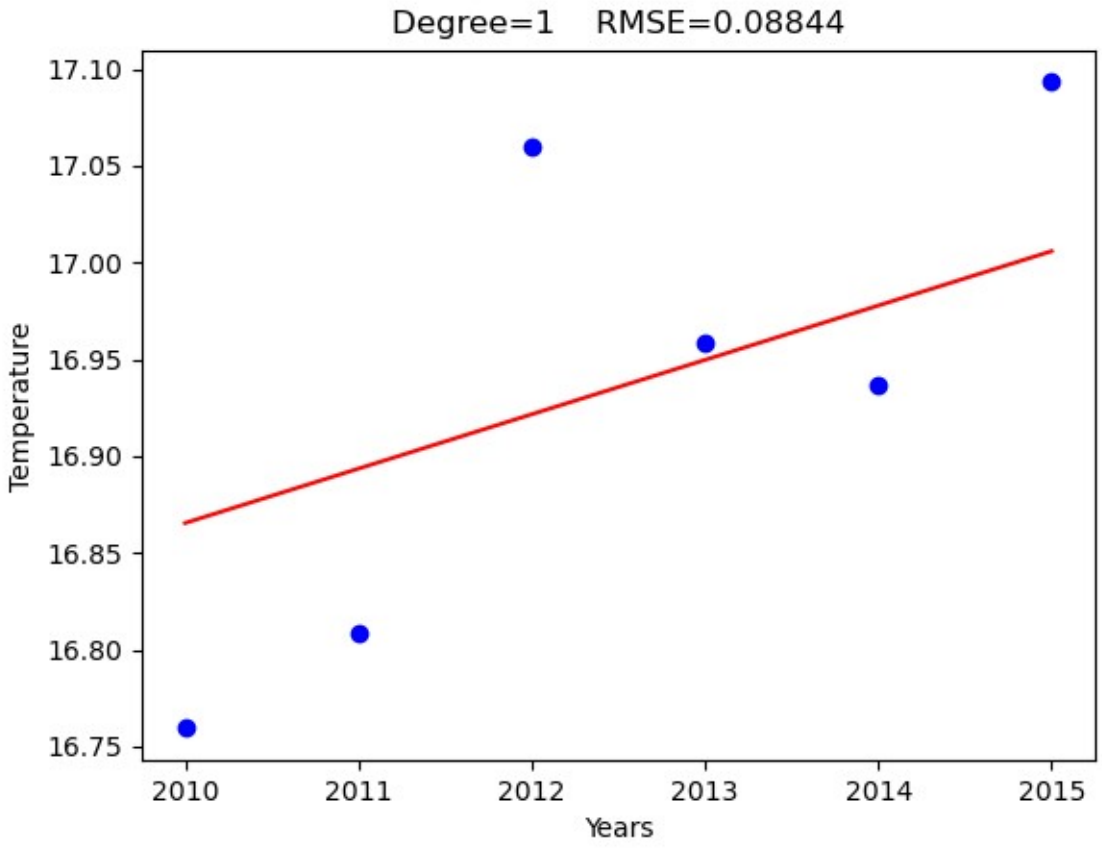


How do these models compare to each other? Which one has the best R 2 ? Why? Which model best fits the data? Why?

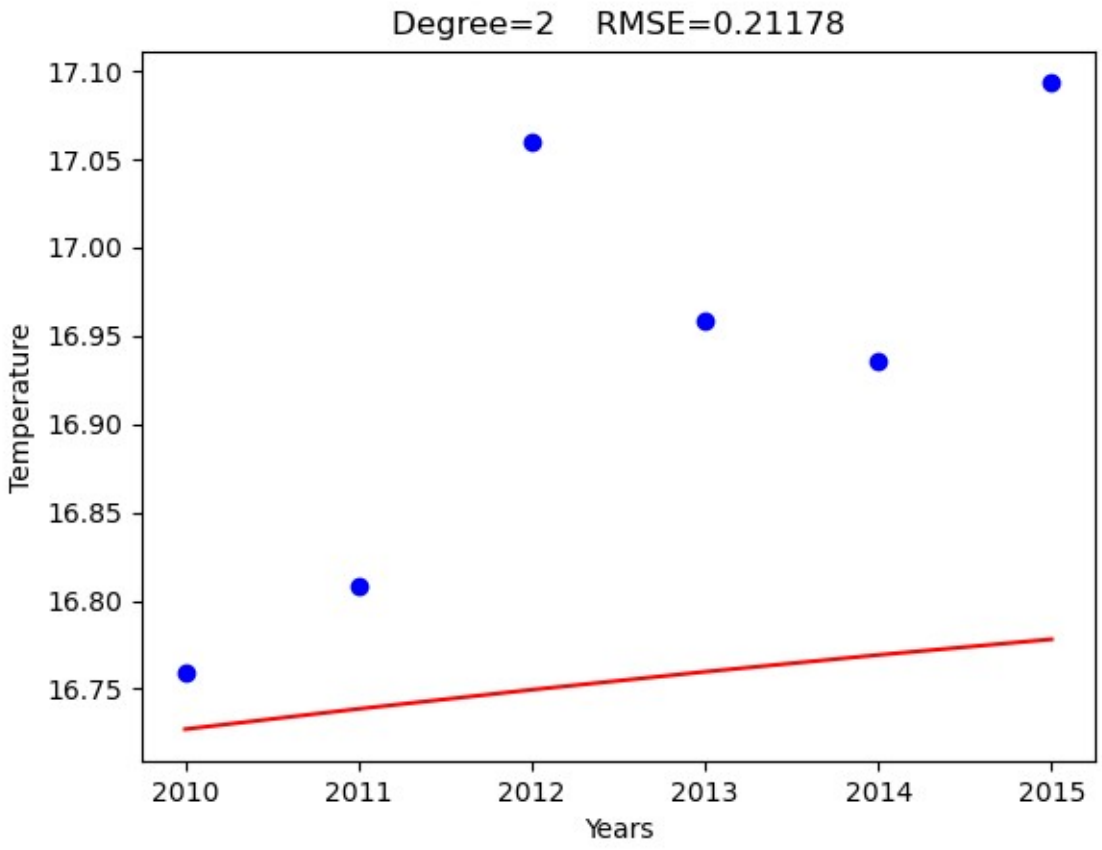
Degree 20 model has the best results R2 is larger and the slope error is smaller. In the same way the degree 2 model is slightly better than the degree-one model. Degree 20 model is the one that best fits the data, because it passes with almost every point on the graph.

Predicting Data

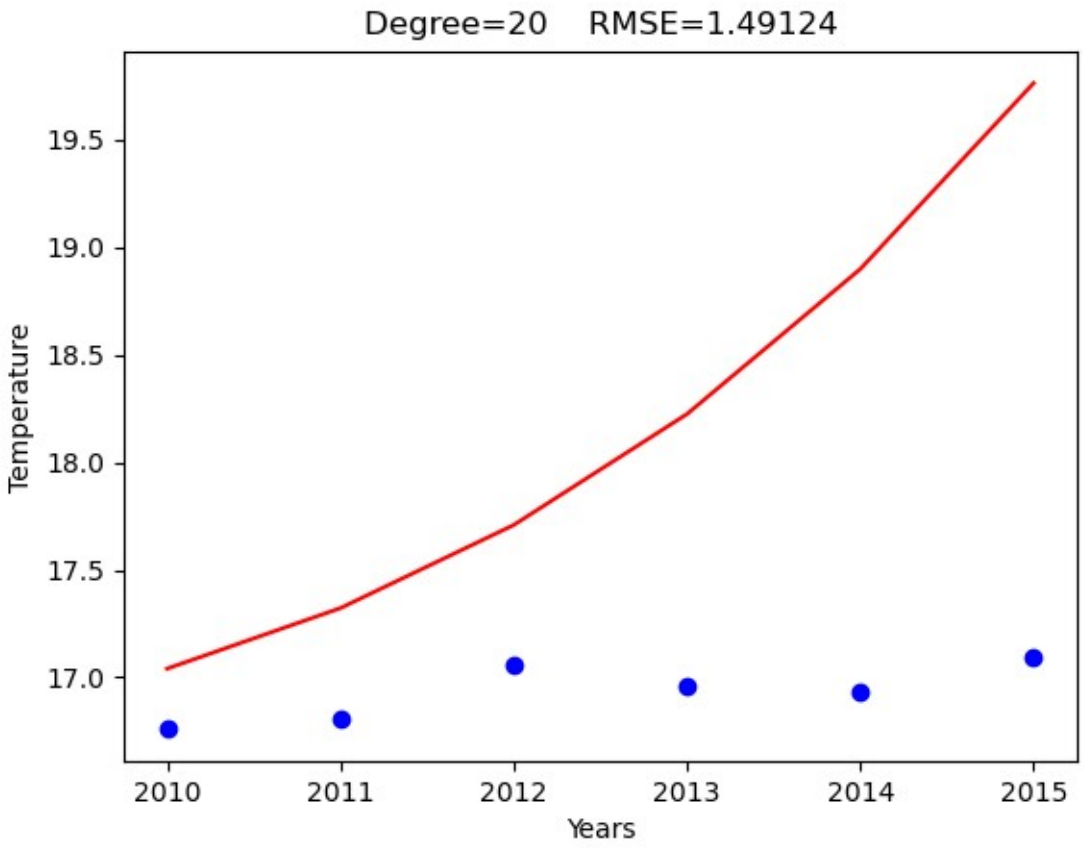
Predicting data for 2009-2016 years: degree 1



Predicting data for 2009-2016 years: degree 2



Predicting data for 2009-2016 years: degree 20



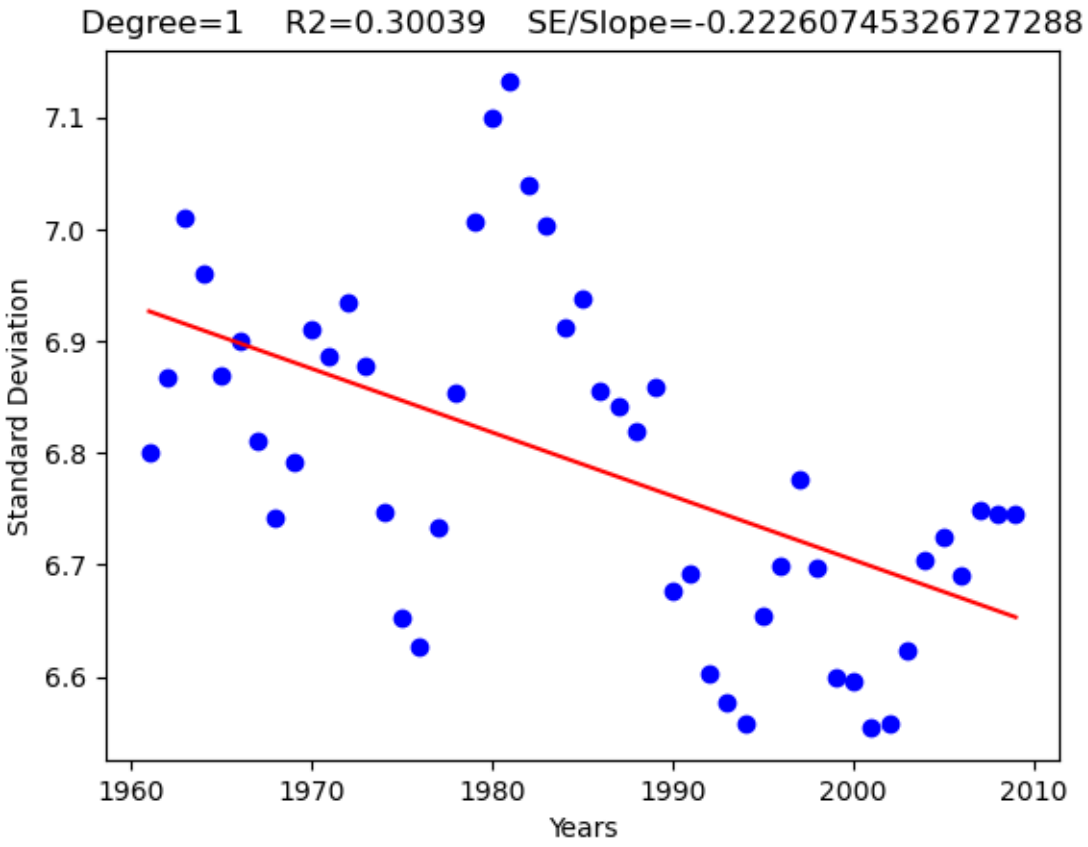
How did the different models perform? How did their RMSEs compare? Which model performed the best? Which model performed the worst?
The RMSE is the best (lowest) for degree-one model. Second-best for the degree-two model and worst for the degree-20 model. Hence, the degree-one model performs best.

Are they the same as those in part D.2.I? Why?
No they are not the same. The model that perfectly fits the training data, does not guarantee that it will fit the test data as well. Test data is noisy and choosing a big number of degree is not always good. Bias is big.

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?
The results will be worse, because the model which is trained only on New York City, cannot fit testing data representing the whole country.

Part E

Standard Deviation over 21 cities during 1961-2009 years



Does the result match our claim (i.e., temperature variation is getting larger over these years)?
No, it does not match. Over the years the standard deviation gets lower, which means that the variation in temperature becomes smaller.

Can you think of ways to improve our analysis?
Using data from more than 21 cities will improve the analysis.