Problem Set 5: Modeling Global Warming

Part A: Creating Models

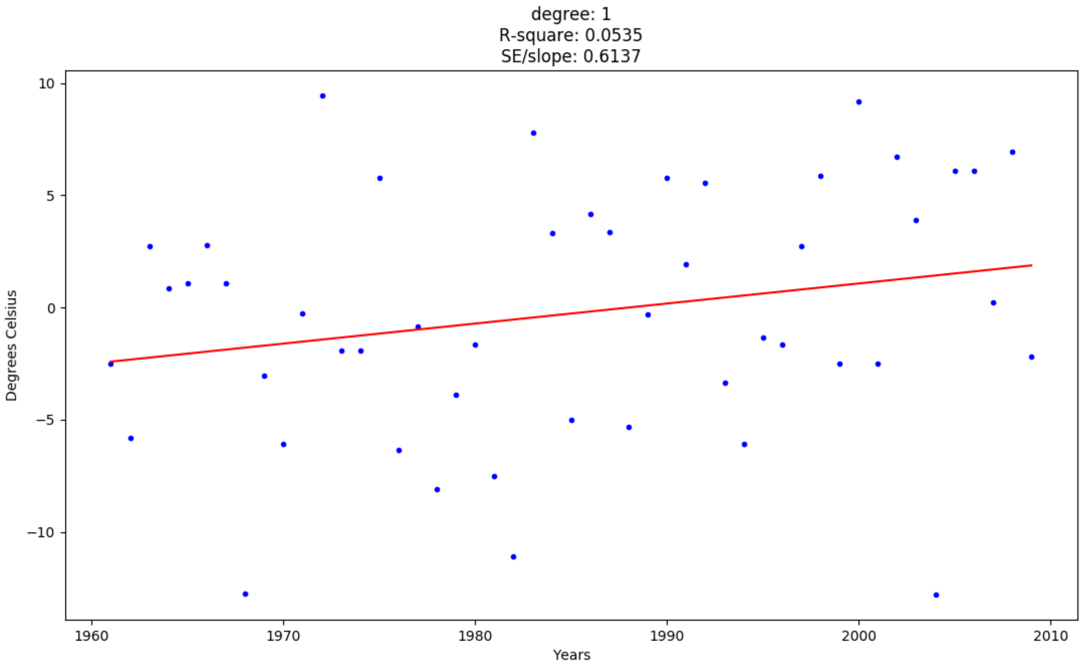
Questions:

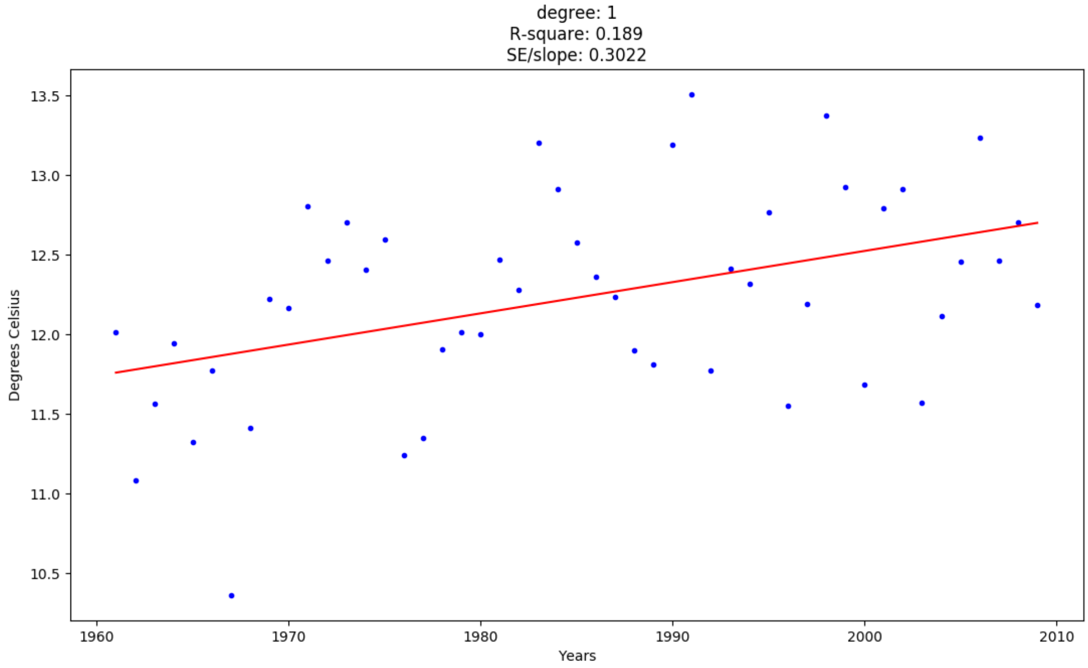
Include the plots for A4.I and A4.II in a document called **ps5_writeup.pdf**. Make sure each plot has appropriately labeled axes, and is titled according to the type of model(e.g. linear, quadratic, etc.), the R^2 value, and the standard error-to-slope ratio.

You also need to answer the following questions with a short paragraph in **ps5_writeup.pdf**.

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

Plots:





Answers:

- The first plot is for January 10th and the second is for yearly average. Apparntly, the yearly average plot gets a better R^2 value than January 10th and a better fitting curve because the yearly average plot uses more data and thus represents the change better than just selecting one day to explore the trend.
- I think the reason for noisy graphs is that we only use small set of data , so it is more likely to have some outliers and go up-and-downs which makes the graphs noisy. The first plot is more noisy.
- the slopes of the fitting lines for both graphs are positive and the SE/slope error in second plot is small, which indicates that these graphs support the claim that global warming is leading to an increase in temperature.

Part B: Incorporating More Data

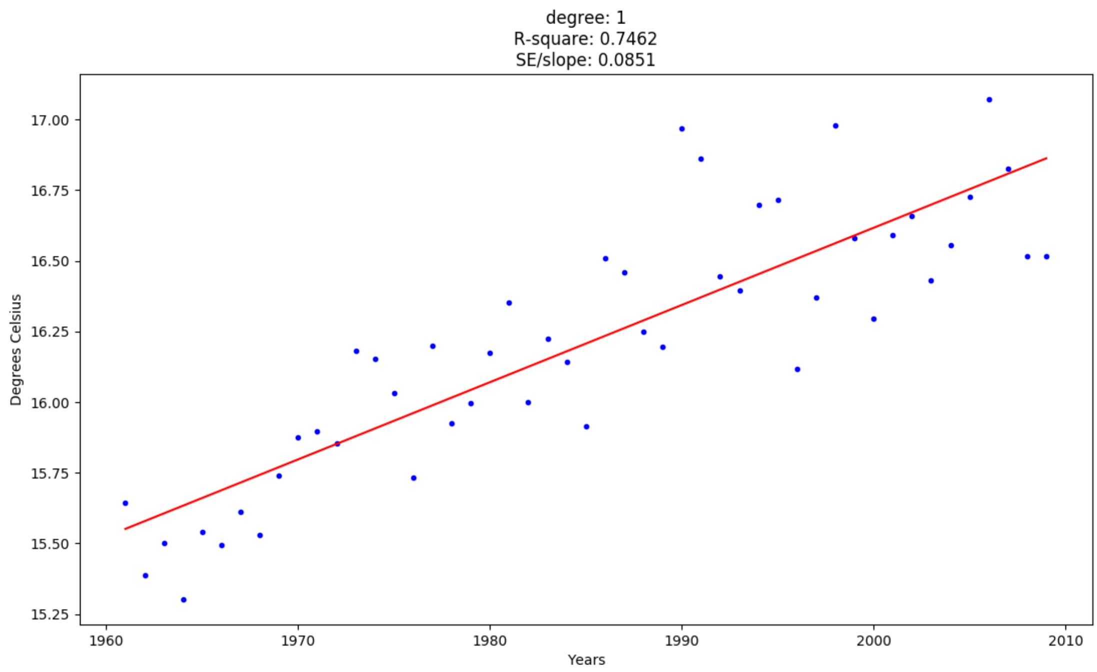
Questions:

Plot the results and include it in **ps5_writeup.pdf**.

Answer the following questions with a short paragraph in **ps5_writeup.pdf**.

- 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.
- Why do you think this is the case?
- How would we expect the results to differ if we used 3 different cities? What about 100 different cities?
- How would the results have changed if all 21 cities were in the same region of the United States(for ex., New England)?

Plots:



Answers:

- Obviously, this graph does a better job comparing to the graphs from part A for a higher R^2 value and a more fitting curve. The higher R^2 value and smaller SE/slope value shows that the data is more close to the fitting line, and the positive slope supports our claim about global warming.
- Because we use data from more than just one city so the data is more convincible.
- We can expect that more cities we use, more convincible and truthful data turns out to be.
- The results will probably not be as good as before because the temperature for 21 cities in the same region cannot speak for the temperature for the whole America.

Part C: 5-year Moving Average

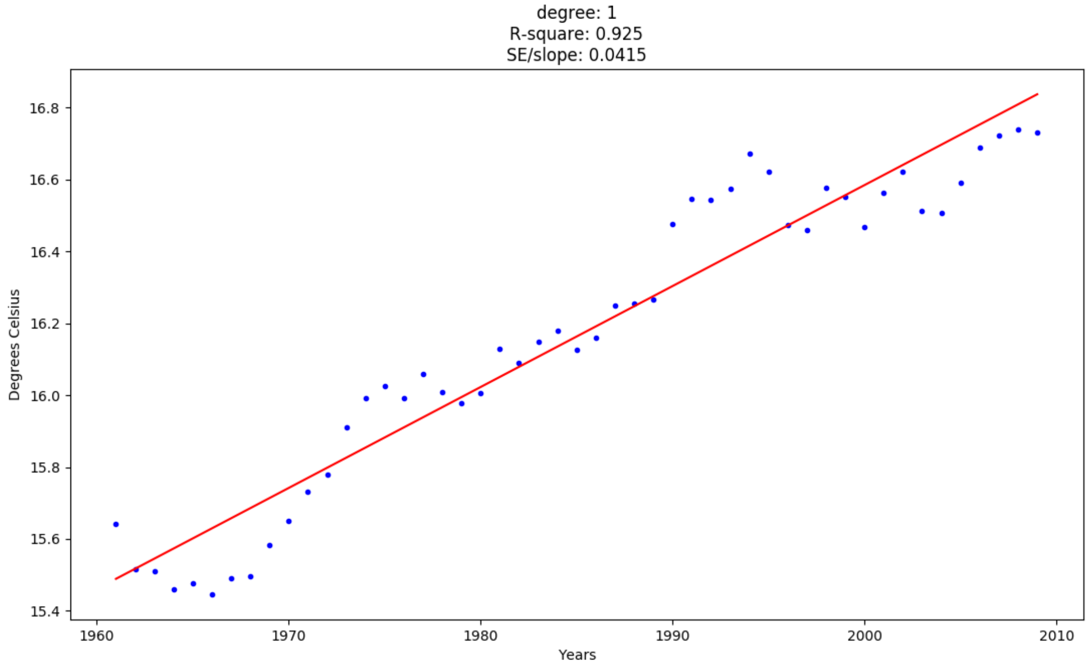
Questions:

Plot the results and include it in **ps5_writeup.pdf**.

Answer the following questions with a short paragraph in **ps5_writeup.pdf**.

- 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.
- Why do you think this is the case?

Plots:



Answers:

- We can easily conclude that this graph looks better than graphs from part A and B. It has an excellent R^2 value and a very low SE/slope. The positive slope for the fitting curve indicates that this graph supports our claim about global warming.
- The `moving_average` method is quite like what we use in image noise reduction. we can reduce the noise in our dataset by replacing one data point with the average value around this data point. And in this way we can explore the internal pattern in the dataset better.

Part D: Predicting the Future

Questions:

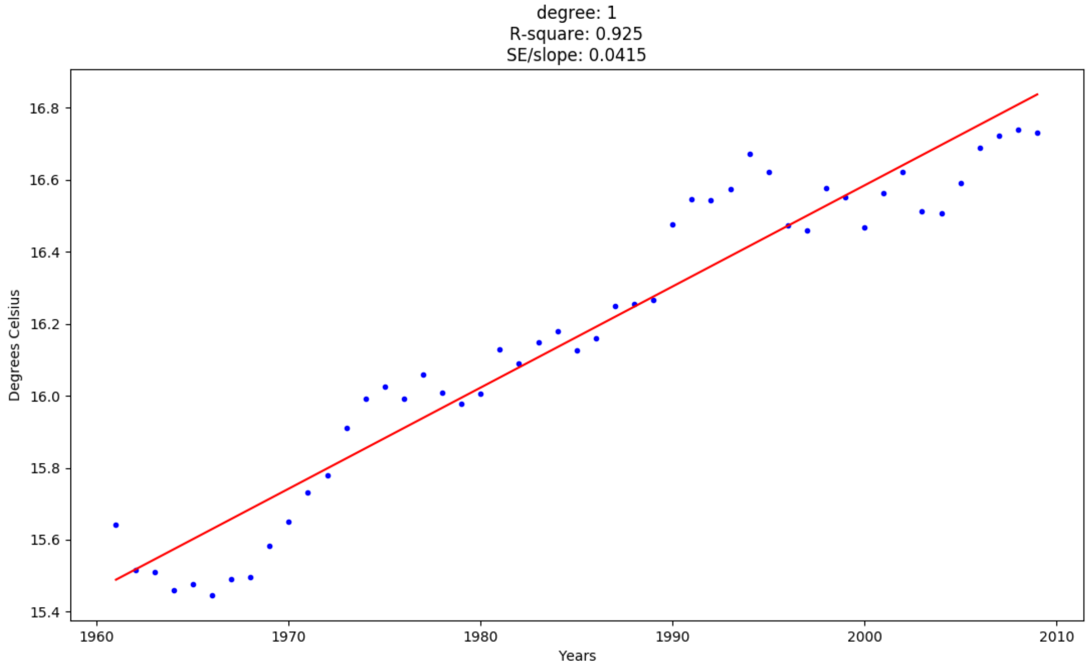
Problem 2.I Generate more models

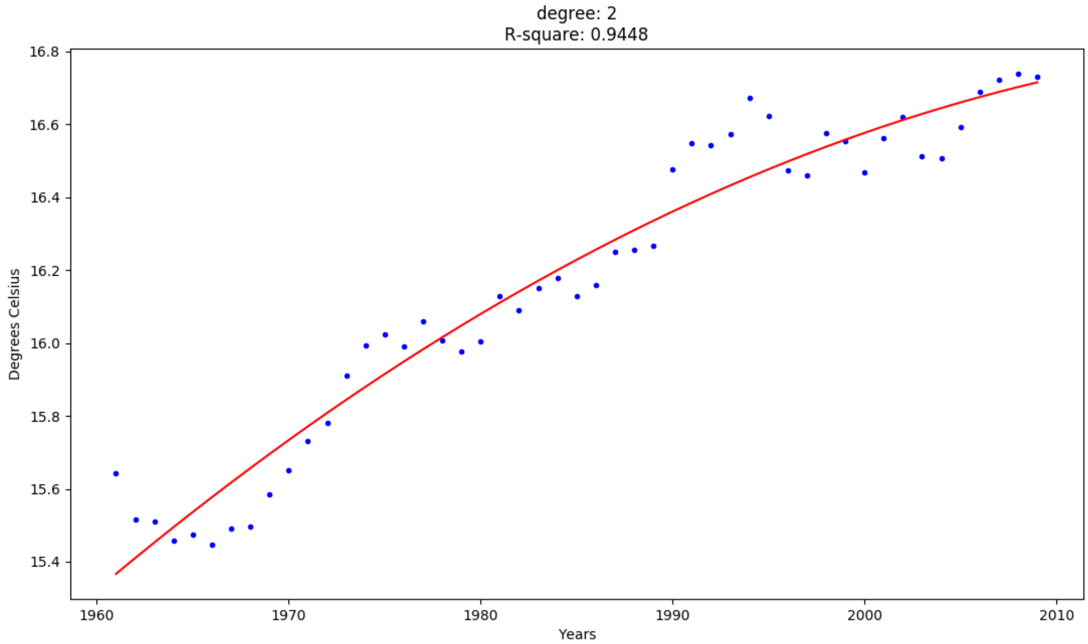
Plot the results and include it in **ps5_writeup.pdf**.

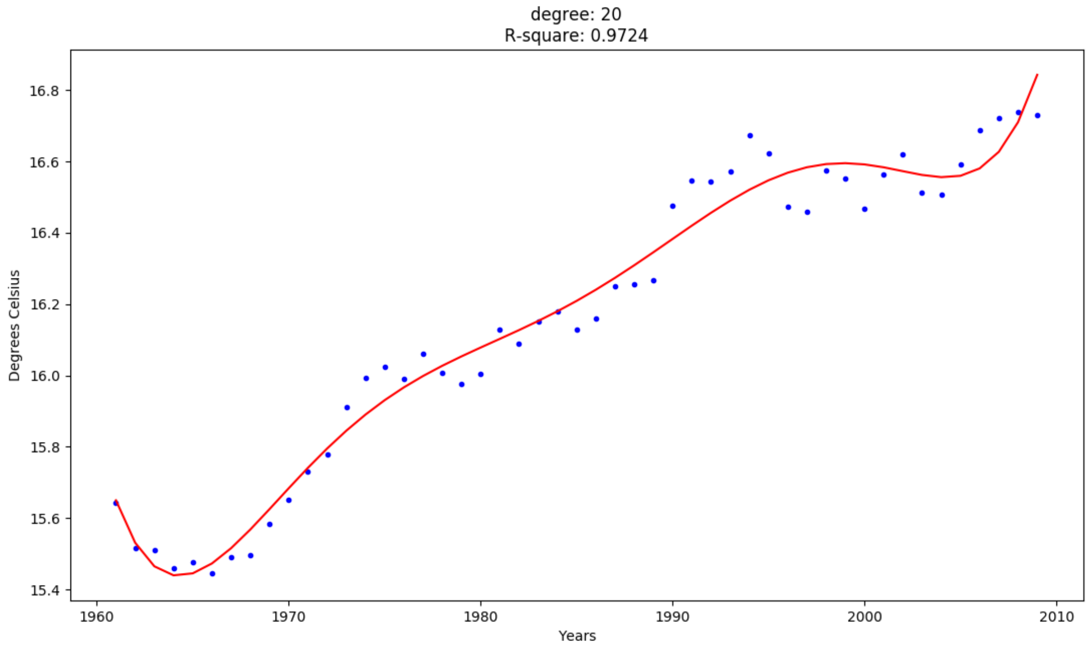
Answer the following questions with a short paragraph in **ps5_writeup.pdf**.

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

Plots:


degree: 1
R-square: 0.925
SE/slope: 0.0415
Degrees Celsius
Years
16.8
16.6
16.4
16.2
16.0
15.8
15.6
15.4
1960
1970
1980
1990
2000
2010


degree: 2
R-square: 0.9448
Degrees Celsius
Years
16.8
16.6
16.4
16.2
16.0
15.8
15.6
15.4
1960
1970
1980
1990
2000
2010



Answers:

- The higher the degree is , the better the model turns out to be.
- The graph for degree 20 has the best R^2, because it has higher polynomial order term which fits the data almost perfectly.
- The graph for degree 20 best fits the data. It is easy to conclude just from the graph without mentioning the best R^2 value.

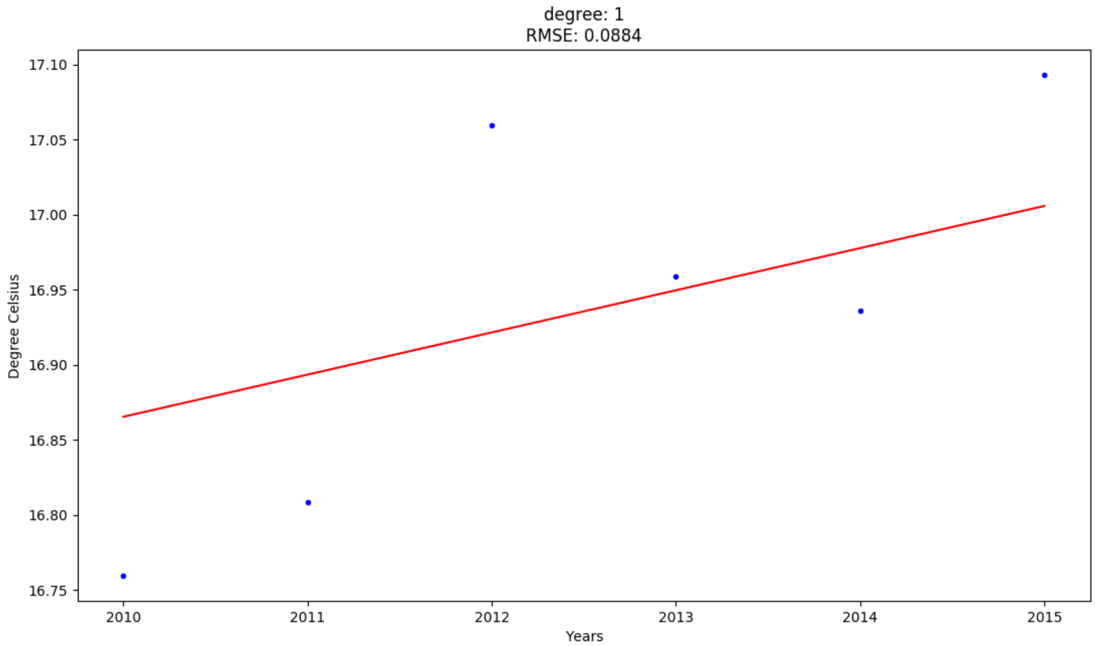
Problem 2.II Predict the results

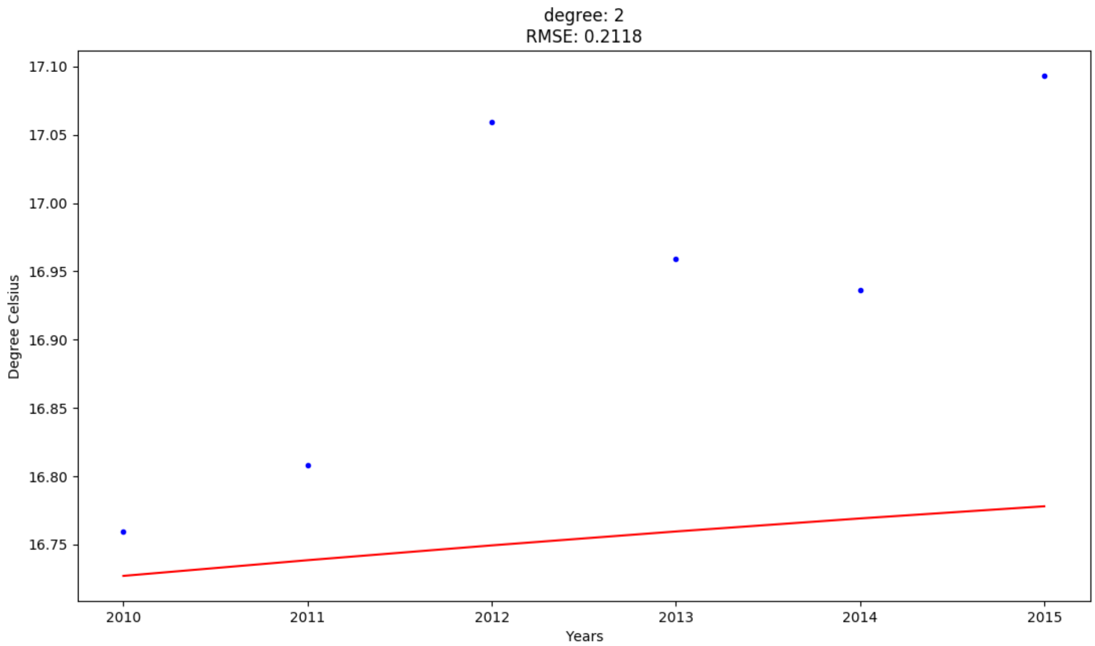
Plot the resulting graphs for degree = 1, 2, 20 models and include them in **ps5_writeup.pdf**.

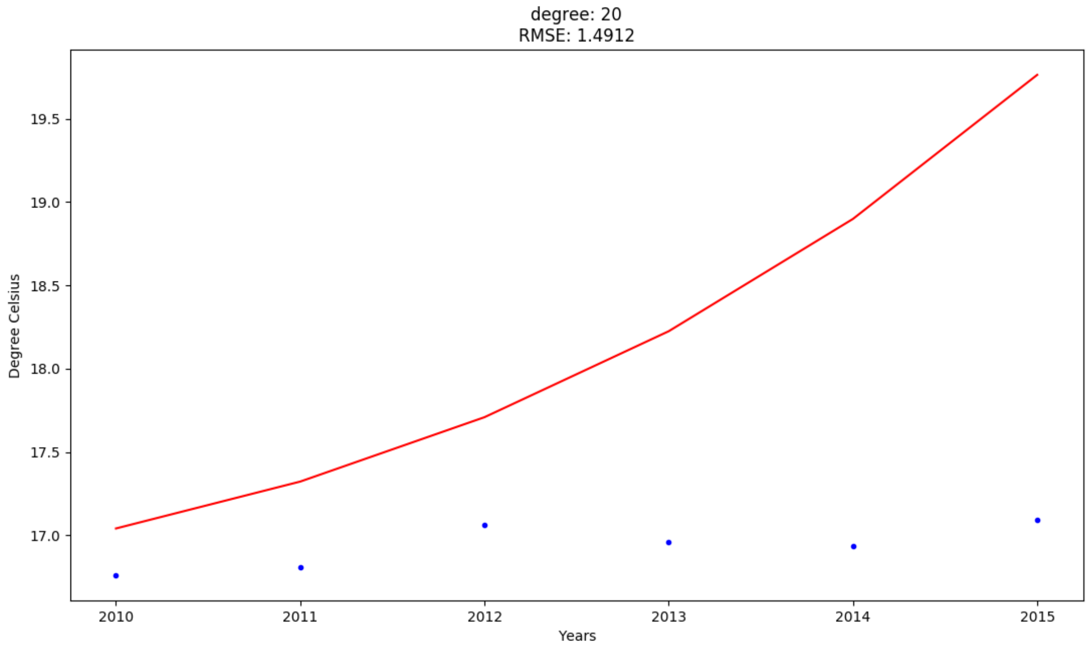
Answer the following questions with a short paragraph in **ps5_writeup.pdf**.

- How did the difference models perform? How did their RMSEs compare?
- Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?
- 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?

Plots:


degree: 1
RMSE: 0.0884
17.10
17.05
17.00
16.95
16.90
16.85
16.80
16.75
Degree Celsius
2010
2011
2012
2013
2014
2015
Years


degree: 2
RMSE: 0.2118
17.10
17.05
17.00
16.95
16.90
16.85
16.80
16.75
Degree Celsius
2010
2011
2012
2013
2014
2015
Years



Answers:

- The higher the degree was, the higher RMSE they got, and the worse they performed.
- The degree-1 model performed best. The degree-20 model performed worst. No, they are not the same as those in part D.2.I because they used test data to evaluate their performance and it turnes out to be degree-20 is totally overfitting.
- I think the predicting results will be worse cause it is easier for the model to overfit the test data in just one city.

Part E: Modeling Extreme Temperatures

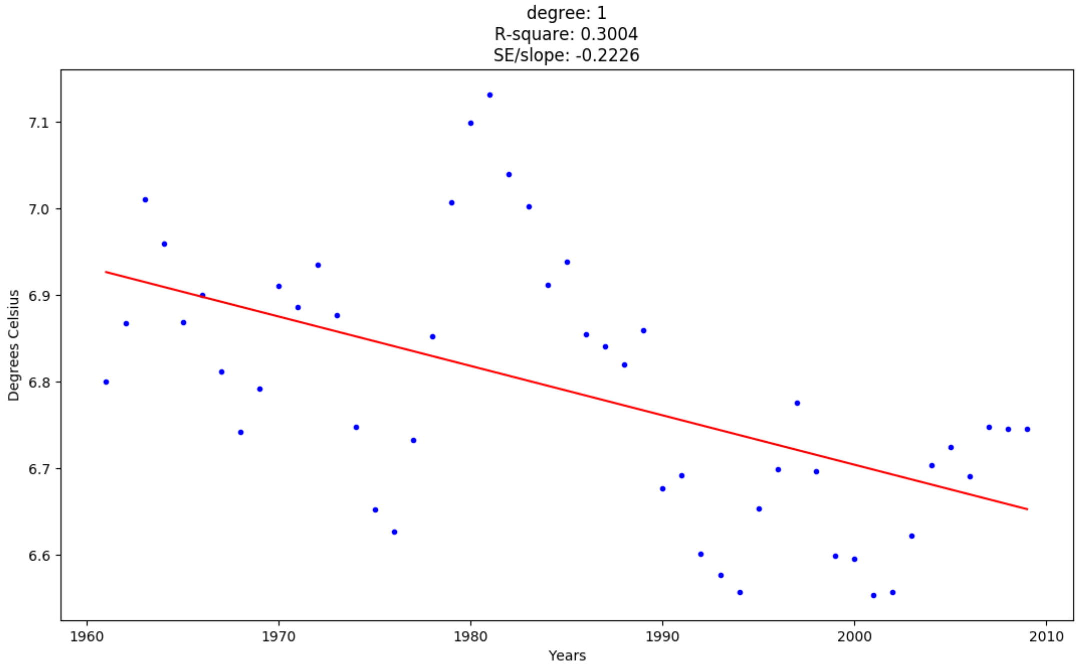
Questions:

Plot the results and include it in **ps5_writeup.pdf**.

Answer the following questions with a short paragraph in **ps5_writeup.pdf**.

- Does the result match our claim(i.e., temperature variation is getting larger over these years)?
- Can you think of ways to improve our analysis?

Plots:



Answers:

- No, it doesn't.
- One way to probably improve our analysis is to use a larger window length. But I want to use another way to find out the temperature variation. We can calculate the annual average temperature across the country and set a criteria that if the daily temperature is some value greater or less than the annual average temperature, it counts as an extreme temperature. We can calculate the number of days for extreme temperature for one year, and use that data to explore whether the temperature variation is getting larger over these years.