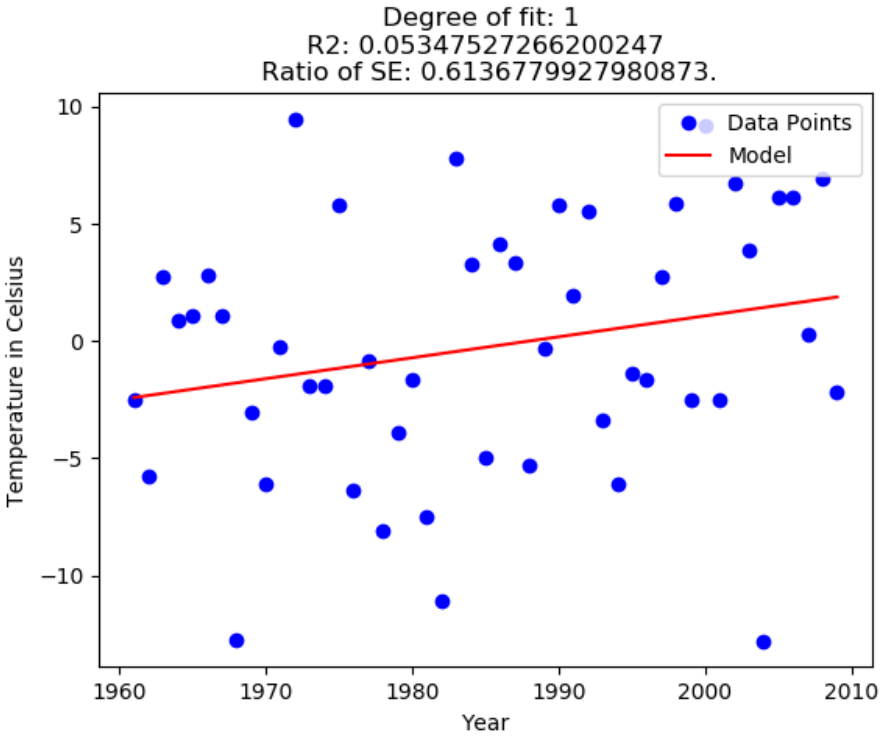
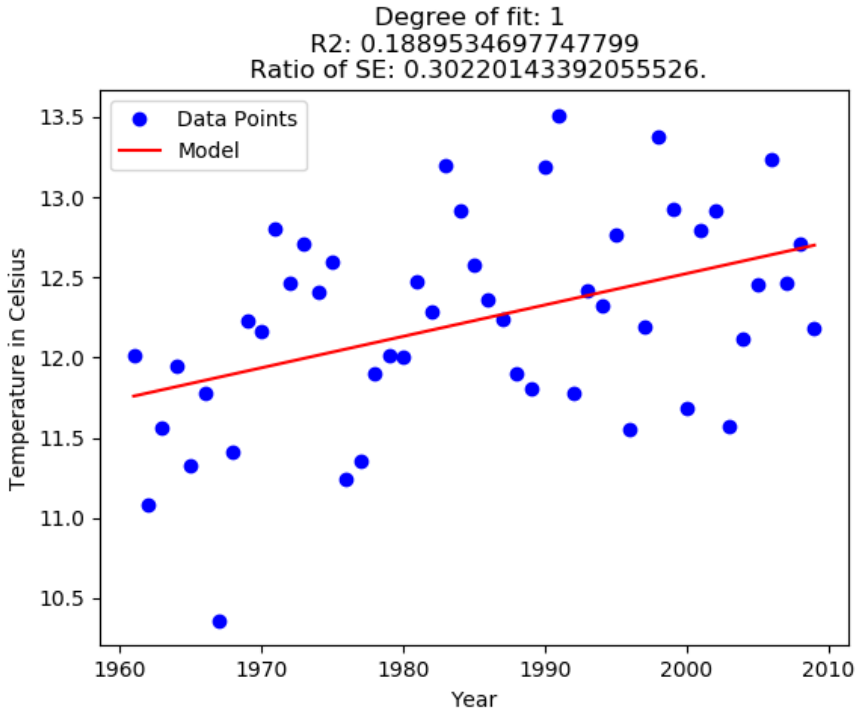
1. *What difference does choosing a specific day to plot the data for versus calculating the yearly average have on our graphs (i.e., in terms of the R2 values and the fit of the resulting curves)? Interpret the results.*

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

3. *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.*





@1 The ratio of standard error of curve slope is according to our measurements over twice more reliable with data where we used mean of yearly temperature (Ratio of SE = 0.3022~) compared to analysis made only with 10th January of each year. Coefficient of determination is higher for analysis with mean of yearly temperatures as well, which suggest that this model will approximate the real data(future) points better, although it is still relatively low.

It all makes sense, since a specific day in a year may not be that sensitive even if there is trend of changes in temperatures, taking average temperatures over the year should be much more reliable.

@2 The more noisy one is the graph containing data collected only from 10th January of each year. Which can be observed by lower density of points close to the representation of our model. I am not sure at that point why data with yearly temperatures is so noisy though…

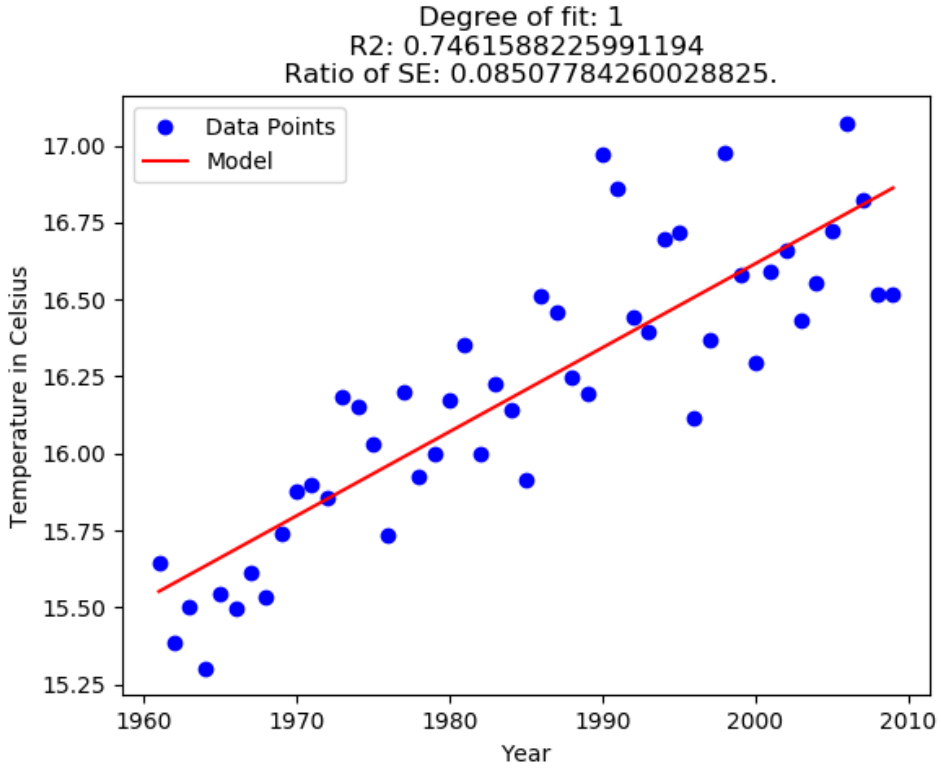
@3 Graph with only daily temperatures over the years does not give us any answer on that. Graph which includes average yearly temperatures is supporting that global warming is leading to increase of temperatures to the higher degree, although the model is still far from being reliable due to the low coefficient of determination (0.18).

1. How does this graph compare to the graphs from part A (i.e., in terms of the R2 values, t he fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.

2. Why do you think this is the case?

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

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



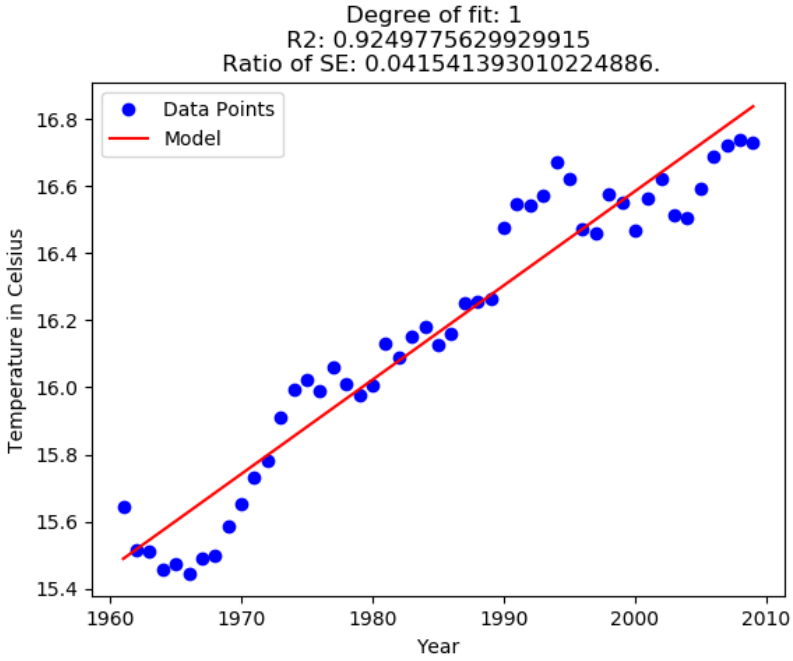
@1 R2 value is much higher, Ratio of SE is much lower which means our model is much more reliable and data is much more dense around regression line. We can now clearly say that there is strong positive correlation between time (years) and temperature which supports global warming claim.

@2 It looks like we did not have enough data before to produce strong model and with the increased amount of data from different cities we are able to make much reliable analysis.

@3 I suspect model will be more reliable with increased amount of cities used in analysis and a lot less reliable and more noisy with only three cities, although there is always risk of overfitting which needs to be considered with 100 cities or even larger amount.

@4 I suspect that data would be more dense, although there would be a risk that we are not really answering the question we started with in the best possible way, since we would like to know if global warming effect is present all over the world and not only in specific region. I believe that the diversity in this case is better than uniformity.

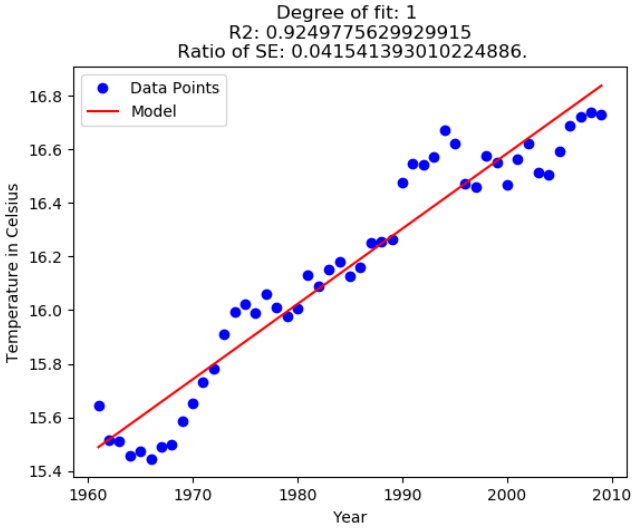
1. How does this graph compare to the graphs from part A and B (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
2. Why do you think this is the case?

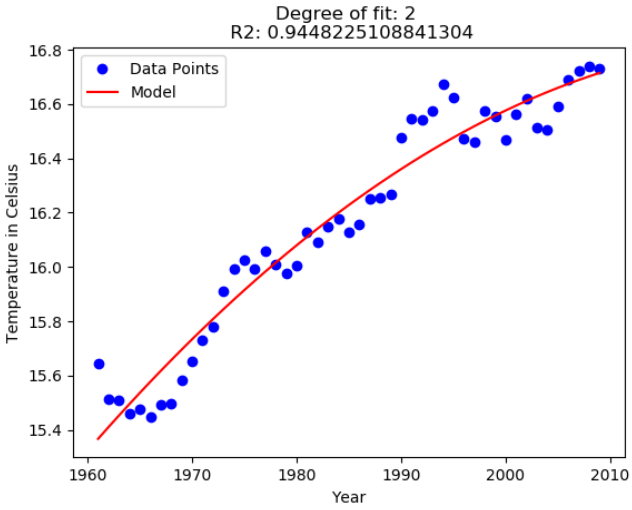


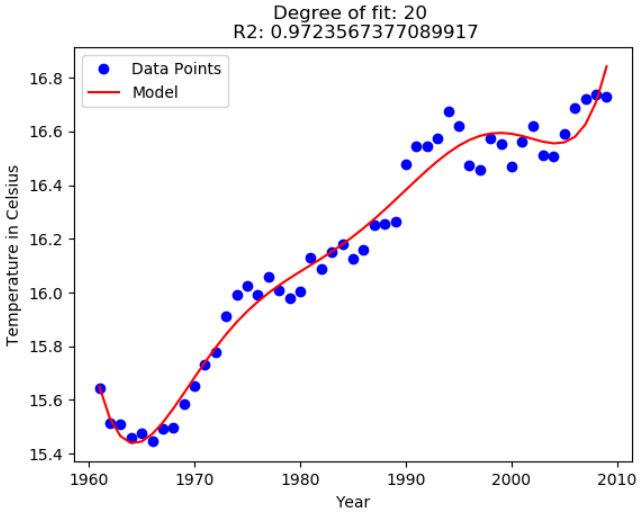
@1 Graph is even more accurate and data is more dense around regression line. Model is more reliable and there are less outliers.

@2 By getting mean of each 5 years we are minimizing the impact of potential outliers.

1. How do these models compare to each other?
2. Which one has the best R2 ? Why?
3. Which model best fits the data? Why?







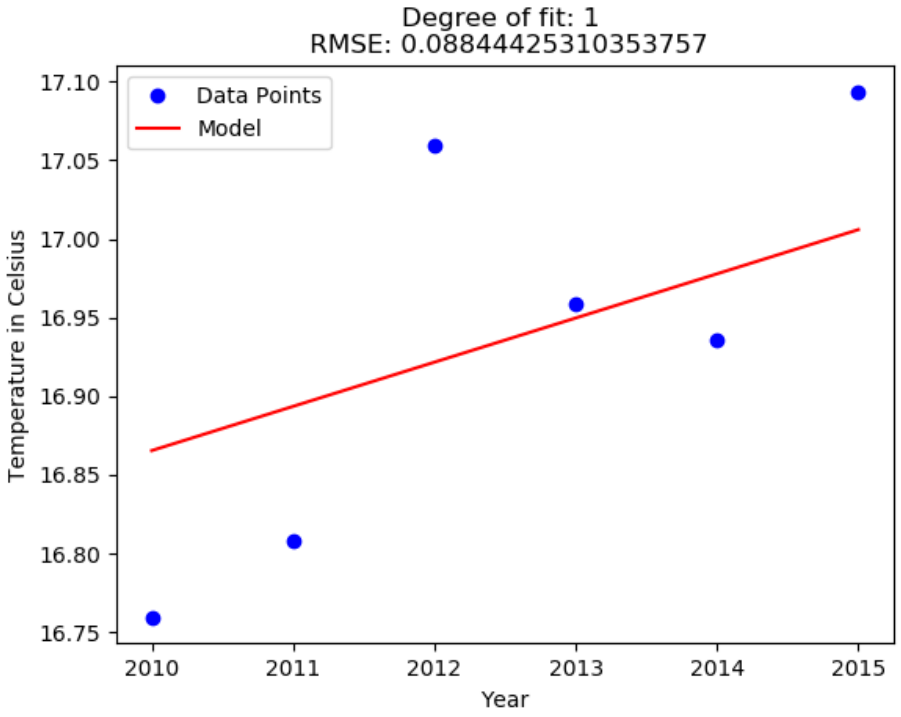
@1 They are all quite accurate. The accuracy of model is increasing while there are added higher degrees of fit.

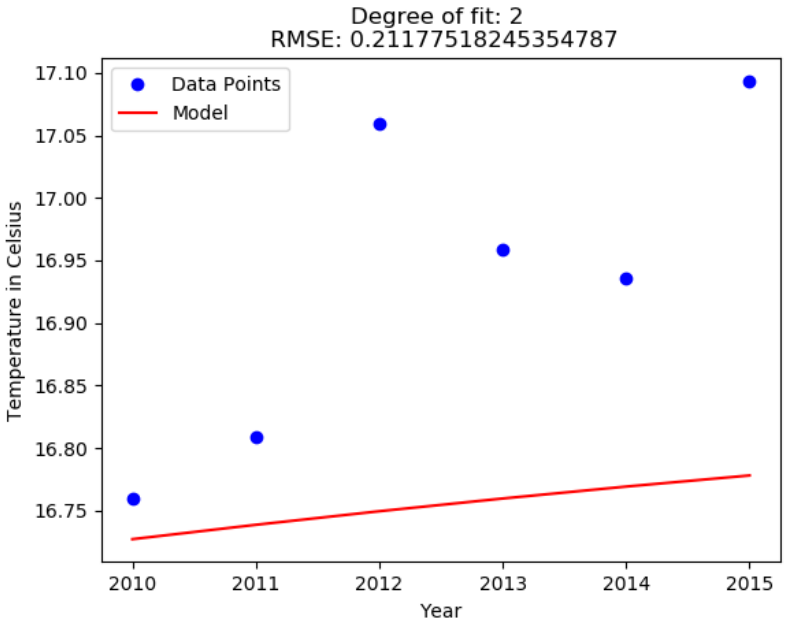
@2 Last model which is using degree of fit = 20 has the highest R2. Usually the higher degree model we make the more accurate it is, if the data we used to build the model is the same as the one we do prediction on, although there is a huge risk of overfitting in case the model will be used for predicting the future data.

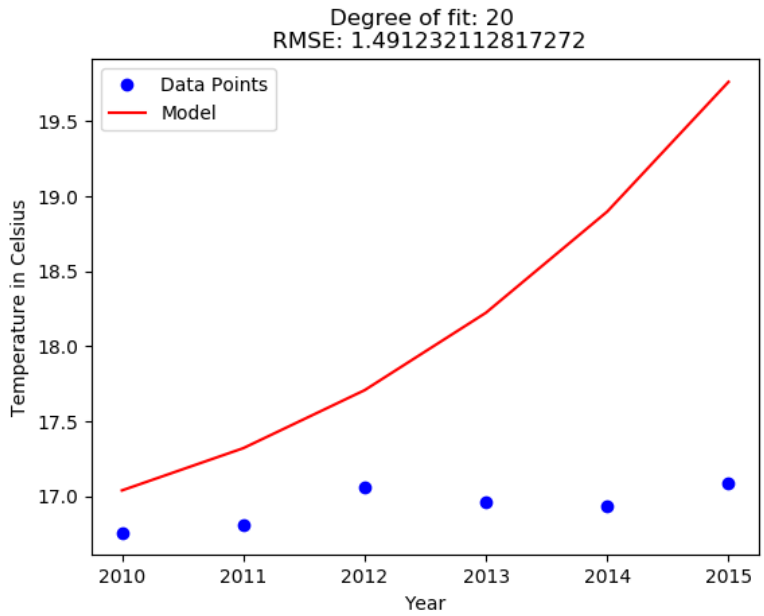
@3 Last mode fits data the best – the reasons are as mentioned above.

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

3. 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?





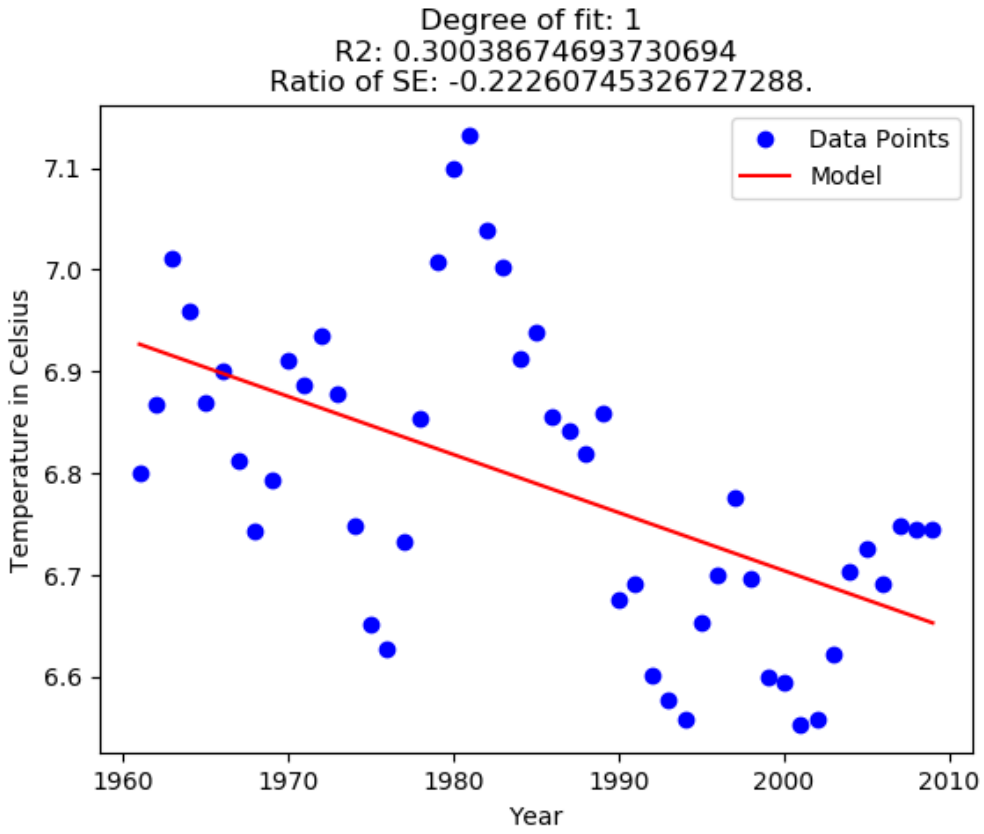


@1 Model with one degree has the lowest(best) RMSE, other model performed worse and the 20 degree model was overfitted which results in misleading prediction on the future data.

@2 One degree models performs tbest, due to the lowest(best) RMSE. 20 degree models is the worst when it comes to RMSE. They are not the same as with training data set, which shows that higher degree is not always more accurate when it comes to predicted data, although it can be very accurate with training which of it was built.

@3 It would be a lot less accurate and there would be a lot more noise in there – R2 of New York model was below 0.20 and R2 for average city temperature in certain year is 0.74 which make it far superior and we can conclude that the prediction based on all cities is more reliable than it would be only with New York.

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



@1 The result does not match that claim, actually there is much more to claim that there is negative correlation between time and temperature variation.

@2 We should be able to improve R2 which is quite low currently with higher degree model, although the correlation between time and temperature variation is not straightforward and it is not quite linear, so there might be other factors determining level of variation.