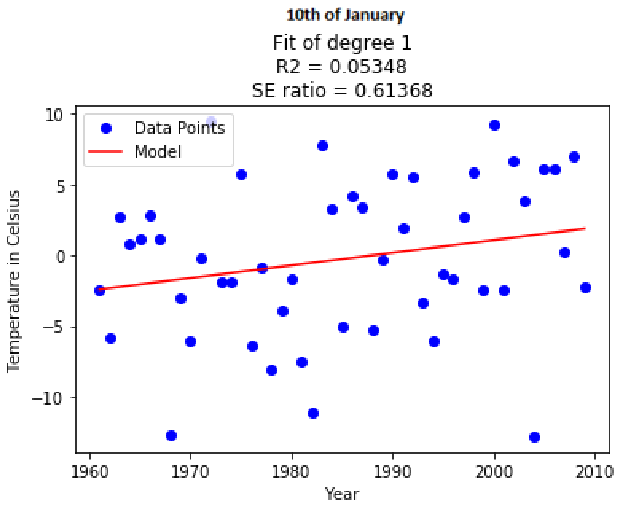
PART A

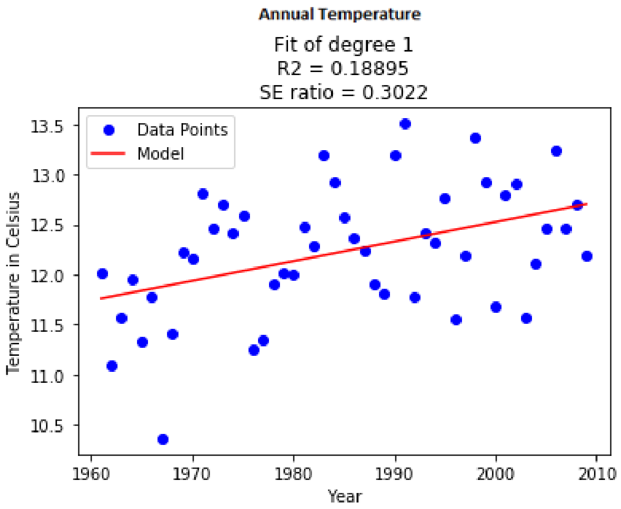


10th of January

Fit of degree 1

R2 = 0.05348

SE ratio = 0.61368



Annual Temperature

Fit of degree 1

R2 = 0.18895

SE ratio = 0.3022

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.

A.

Case	1: 10th January	2: Annual
R2	0.05348	0.18895
SE ratio	0.61368	0.3022

By comparing R2 from both cases, although case 2 presents a better R2 and SE ratio value than case 1, it appears neither captures the proportion of variability in the data set accounted for by the statistical model provided by the fit. The values are close to zero suggesting that there is very little relationship between the values predicted by the model and the actual data.

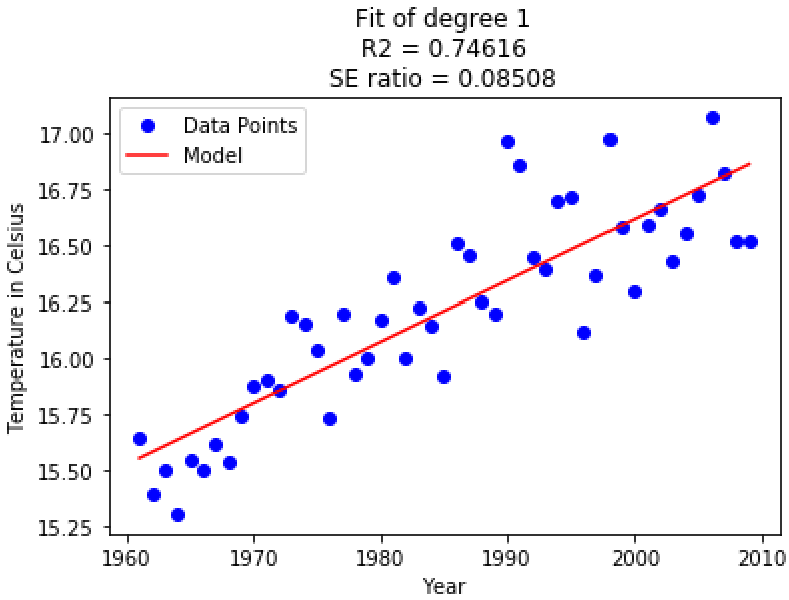
2. Why do you think these graphs are so noisy? Which one is noisier?

A. The data from a specific day is noisier than the data from a yearly average. Which suggests that the data gathered of a single day gives a not so good, small sample size to get a good approximation.

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.

A. In our case, if the absolute value of the SE ratio is less than 0.5 the trend is significant. Otherwise, the more likely it is that the trend is by chance. In case 1, SE ratio of 0.61368 we can’t be sure if the upward trend shown is by chance. But in case 2, where SE ratio is 0.3022 temperatures are going upwards supporting the claim that global warming is leading to an increase in temperatures.

PART B



1. 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.

A. In part A, R2 was close to zero indicating there is almost no relationship between the values predicted by the model and the actual data. In this part, R2 = 0.74616, we can say the model explains 74,62% of the variability in the data. The SE ratio of 0.08508 is also an indicative that the upward trend predicted by the model is significant, not by chance. We also observe there is a higher density of data points to the fitted curve showing a less noisy graph. All that supports the claim about global warming.

2. Why do you think this is the case?

A. As explained in part A, the sample size was too small to get a good approximation. By incorporating data from more than one city we have a larger sample and a tighter bound.

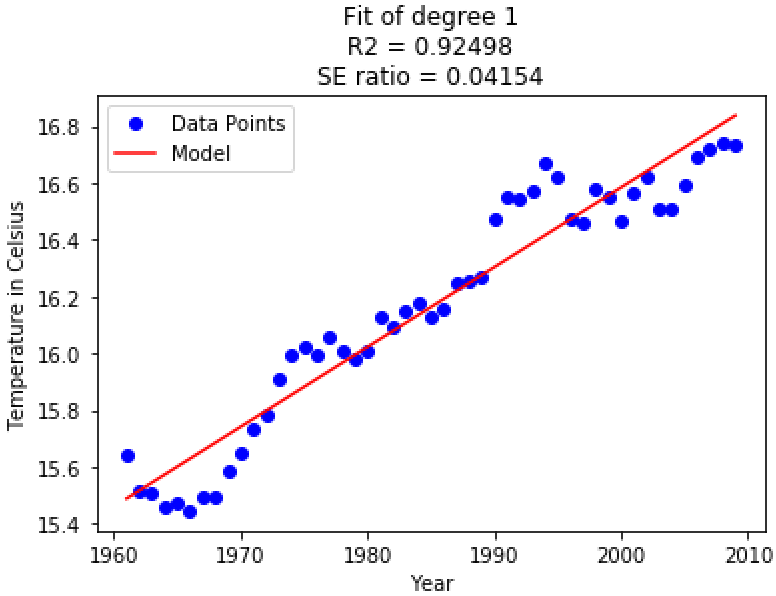
3. How would we expect the results to differ if we used 3 different cities? What about 100 different cities?

A. Although the upward trend may be true for both situations, we would expect data points to be closer and closer to the predicted model as more cities are considered. Se ratio would be higher and R2 lower, indicating the trend predicted would have less relation to actual data. In case 100 different cities are considered, the ratio would go closer to 1 explaining most of the variability in data and SE ratio would go closer to zero suggesting the likelihood the upward trend is significant. But can't be ignored the risk of overfitting the model and that should be handled carefully.

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

A. Although data points may be more dense, we might not be analysing global warming correctly as its effects are all over the world and not only in a very specific region. Diversity of data is important for this case.

PART C



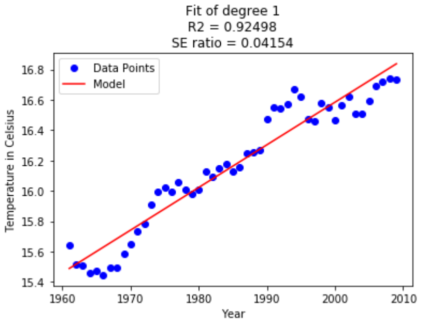
1. How does this graph compare to the graphs from part A and B (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.

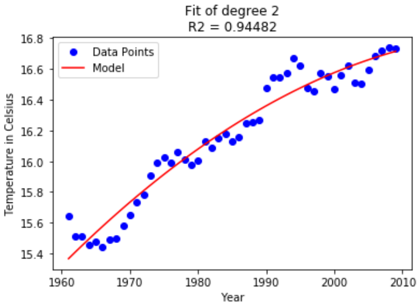
A. R2 value is much closer to 1 in this graph, indicating close relationship between the curve and the data. The SE ratio also got closer to zero than the graphs from part A and B, indicating that the trend is significant, not by chance. Both results, supports again the claim about global warming.

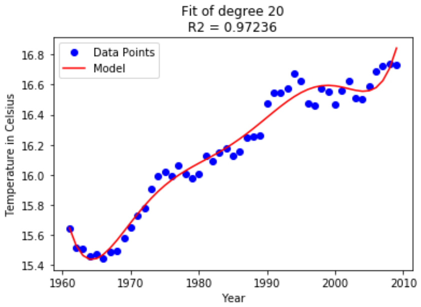
2. Why do you think this is the case?

A. By modelling the curve by taking a moving average over 5 years of data, it allows o emphasize the general trend over local fluctuations.

PART D.2.I







1. How do these models compare to each other?

A. All three models, of degree 1, 2 and 20, represents a polynomial to a predicted model. The higher the degree the better R2 value is observed.

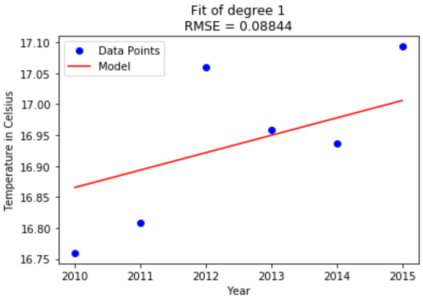
2. Which one has the best *R*2? Why?

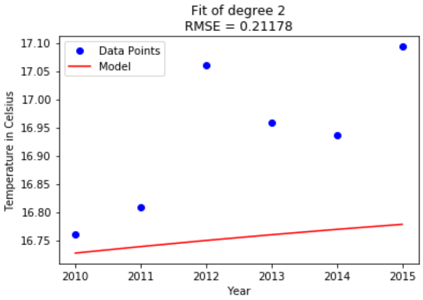
A. The fit of degree 20 has the best R2, but the overly-complex model may lead to overfitting future data.

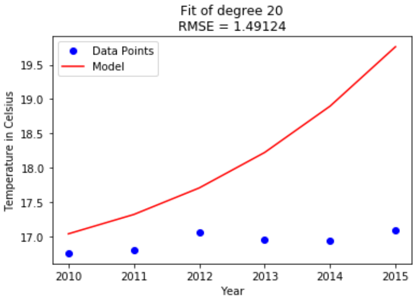
3. Which model best fits the data? Why?

A. The best is the fit of degree 1, taking in consideration that we don't want to overfit our data and R2 value in fit of degree 2 is just slightly higher than fit of degree 1. If we ignore the issue of overfitting, fit of degree 20 is the one.

PART D.2.II







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

A. The fit of degree 1 had the best RMSE value, of 0.08844, compared to degree 2 and 20. Fit of degree 2, presenting a RMSE of 0.21178, not so good performance. And fit of degree 20 had the poorest work on the data, with an RMSE value of 1.4089 with a steep curve.

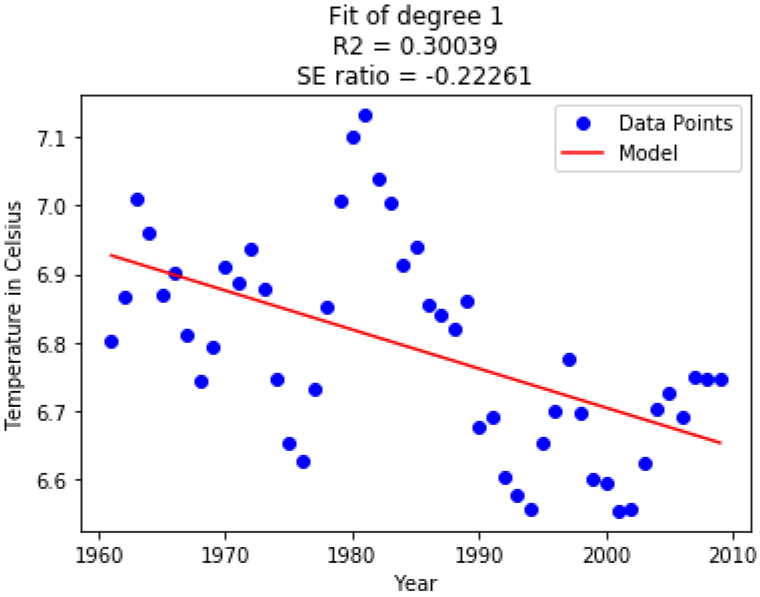
2. Which model performed the best? Which model performed the worst? Are they the same as those in part D.2.I? Why?

A. Fit of degree 1 performs the best meanwhile fit of degree 20 works poorly. When the same models from the training data are applied to the testing data to predict future temperatures choosing an overly-complex model lead to overfitting to the training data. Increasing the risk of a model to work poorly on data not included in the training set, which is exactly what happened.

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?

A. It wouldn't give us a good model, once the models from A.4.II had already R2 < 0.2 which is a strong indicator that the model is not suitable.

PART E



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

A. The result does not match the claim.

2. Can you think of ways to improve our analysis?

A. There may be other factors to be considered to model the global warming but we could consider higher degree models and improve R2 for example.