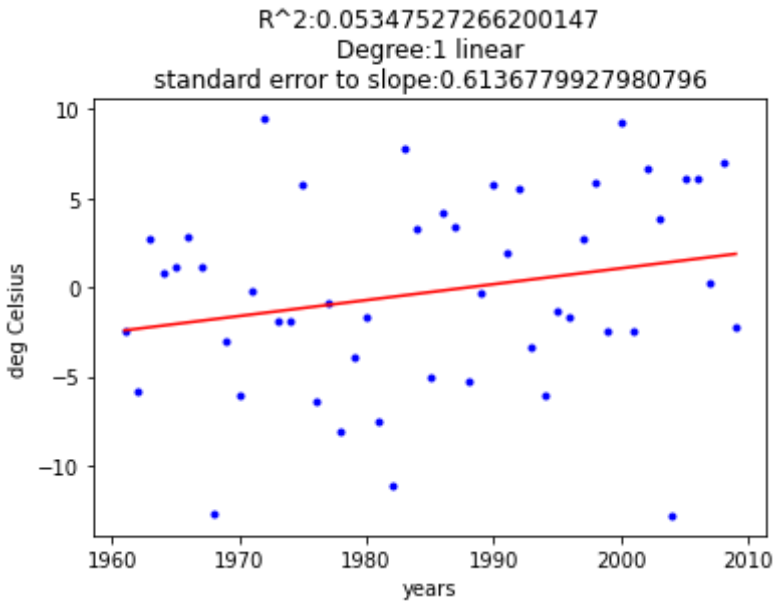
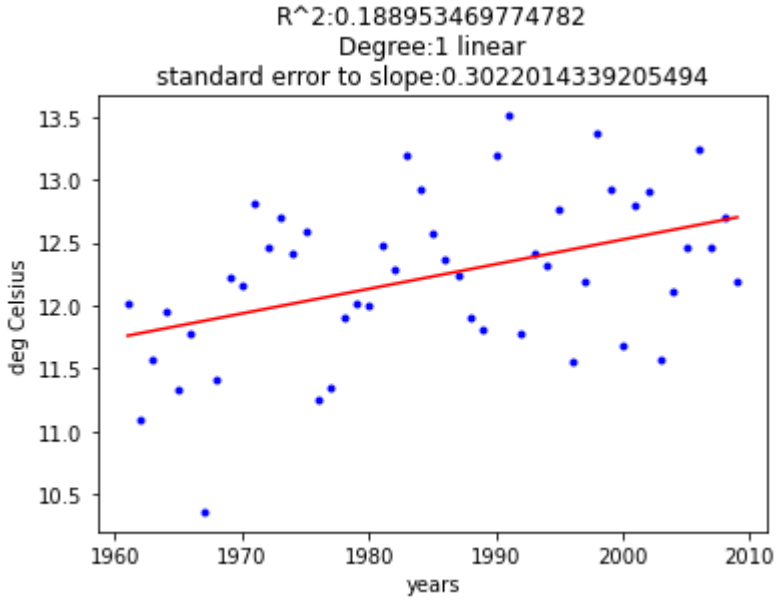


A.1: temp on jan 4



A.2: annual temp

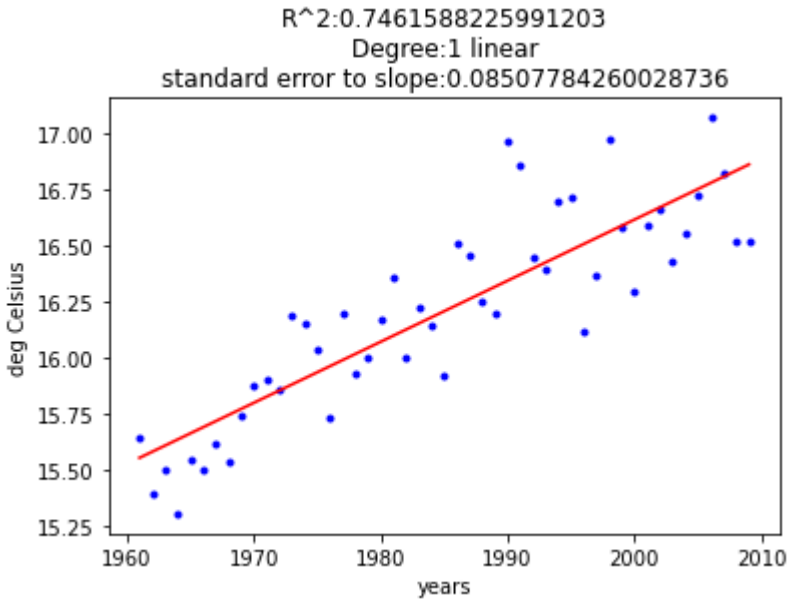
What difference does choosing a specific day to plot the data for versus calculating the yearly average have on our graphs (i.e., in terms of the R2 values and the fit of the resulting curves)? Interpret the results.

The R^2 value for the annual graph is much higher than that of the daily graph, and the line clearly is a less precise fit. This means that there is more noise and variation in the data for the single day. This makes sense, as for the daily graph, each data point on the graph corresponds to only 1 temperature measurment, while for the annual graph, each point represents the average of the entire years temperatures. Thus, due to Bernoulli's law, they will tend to have less random vairence

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

The daily graph is much more noisy, as its temperature plot points come from only a single measurement, while the annul graphs come from an entire years average. The reason both of the graphs are so noisy though is that neither have a large data set. They are only measuring the temperatures of a single city, and due to the unpredictability of weather, this causes a large amount of randomness.

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.
While they both do show an upwards trend with both graphs having a positive slope, the daily graph is far to noisy and has a much too high standard error of slope for its results to really say anything significant.. The annual graph is better, but even it suffers from a high amount of noise and a relatively high SES.



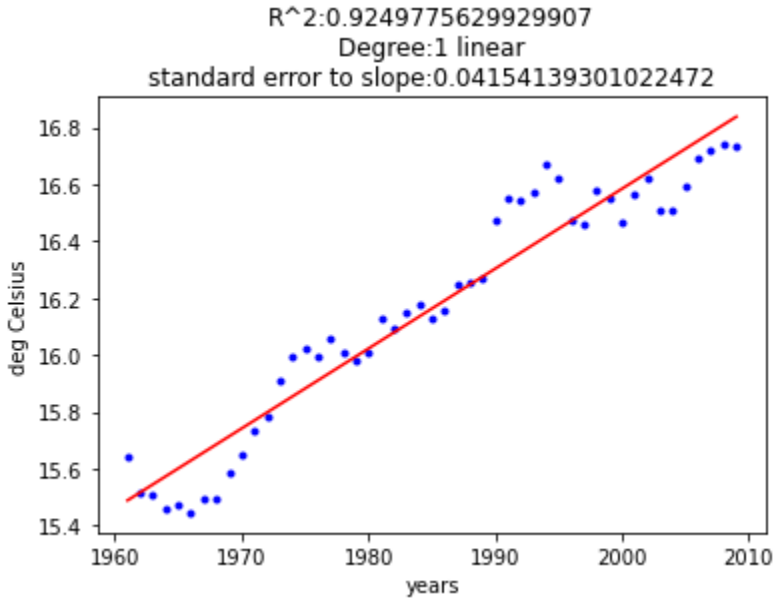
B:national national temp
How does this graph compare to the graphs from part A (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
The R^2 value is significantly higher than either of the other graphs, and the SES is much lower. This model shows a very clear and indisputable upwards trend in temperature, supporting global warming.

Why do you think this is the case?
This is the case because this time we had a much larger, more varied data set, with 21 cities from all over the country, instead of just 1. Thus, there were many more data points, and because of this (predicted bu Bernoulli's law), the overall noise is much lower.

How would we expect the results to differ if we used 3 different cities? What about 100 different cities?
With 3 cities, there would be much more noise and a lower R^2, although still higher than the first 2 graphs. With 100, R^2 would be even higher and SES lower.

How would the results have changed if all 21 cities were in the same region of the United States (for ex., New England)?
The results would have still been better than the first 2 graphs, but they would have been much worse than how it was now. This is because temperature in nearby regions tends to be similar, so if tehre was a particularly hot or cold year for that area of the country, that would effect all data points, instead of that outlier being normalised by the other points like it is in the current model

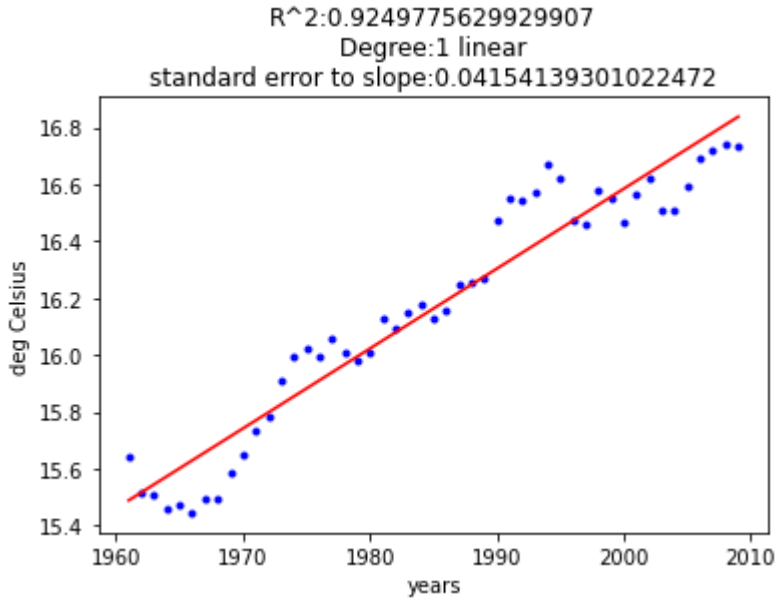
C:national annual temp moving avg



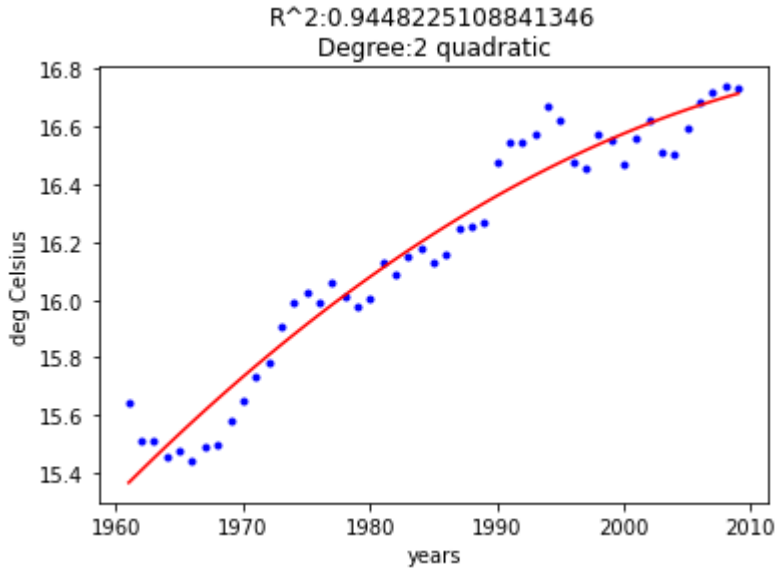
How does this graph compare to the graphs from part A and B (i.e., in terms of the R2 values, the fit of the resulting curves, and whether the graph supports/contradicts our claim about global warming)? Interpret the results.
This graph has the highest R^2 value yet, implying an even more accurate model, and it similar to part B shows a very clear upwards trend in the temperature, supporting global warming.

Why do you think this is the case?
This is because by using a moving average we are able to reduce the effect of outliers even more. This is because now, each data point is not just one day for once city, or a year for once city, or even a year for 21 cities, but instead 5 years for 21 cities.

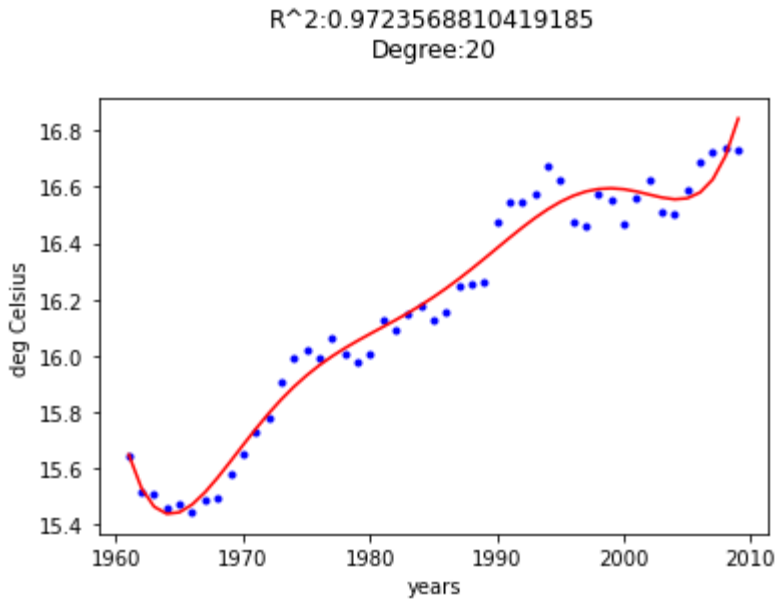
D.1 moving avg training data:



Deg1:



Deg2:



Deg20:

How do these models compare to each other?

As the degree increases, so does the R^2 and how accurately they fit the data.

Which one has the best R 2 ? Why?

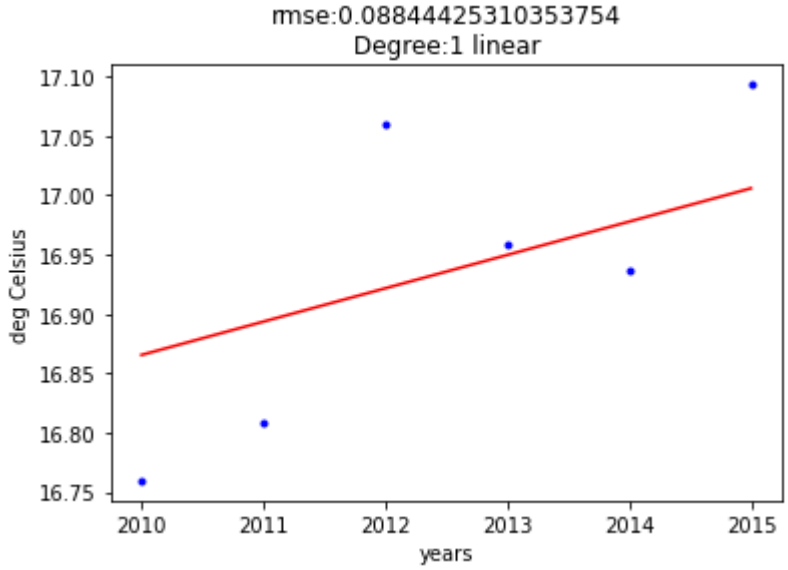
The degree 20 model has the best R^2. this is because it is able to not only model the overall trend of the data, but also fit itself to the noise in the data.

Which model best fits the data? Why?

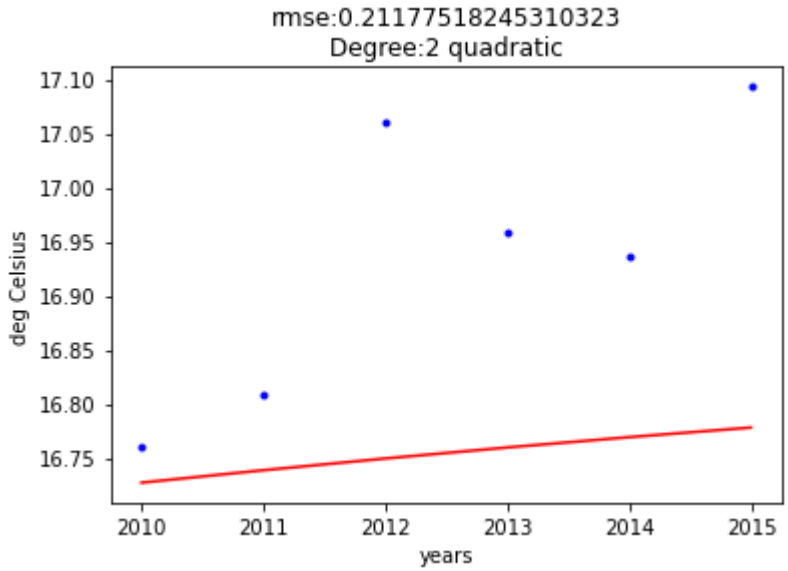
(by fit, I assume you want me to predict which one will perform the best at predicting new data, as otherwise this just seems like the same question as which has the best R^2).

Most likely the linear model, although possibly the quadratic. Most things in the real world will have less than a quartic degree. That is simply the nature of things. And because the degree 20 model has trained itself with a data set that has noise, it will try to predict that same noise pattern for future data, which will end up making it less accurate.

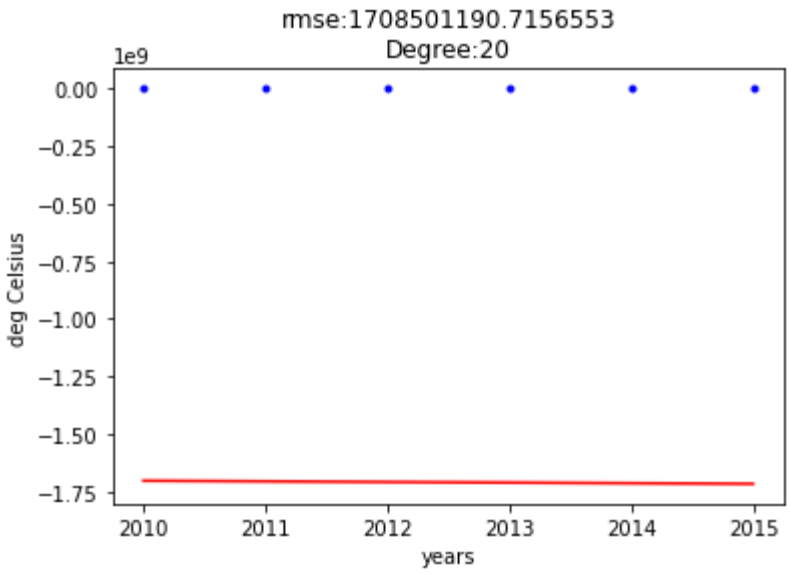
D.2: predicting future data:



Deg1:



Deg2:



Deg20:

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

The linear had the lowest RMSE and thus performed the best, the quadratic had a high RMSE, and the degree 20 had and RMSE 11 orders of magnitude higher than that.

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 linear and the worst was the degree 20. This is reversed from what it was previously. This is because the actual relationship between temperature and time is a linear one, so the other models overtrained and tried to include the noise they were trained on in there predictions.

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 positions for worst and best would remain the same, as the relationship is still linear, but they all would have been worse as the training data was less accurate.