PS5 writeup

Figure A: Highest daily temperature of New York from 1960 to 2010

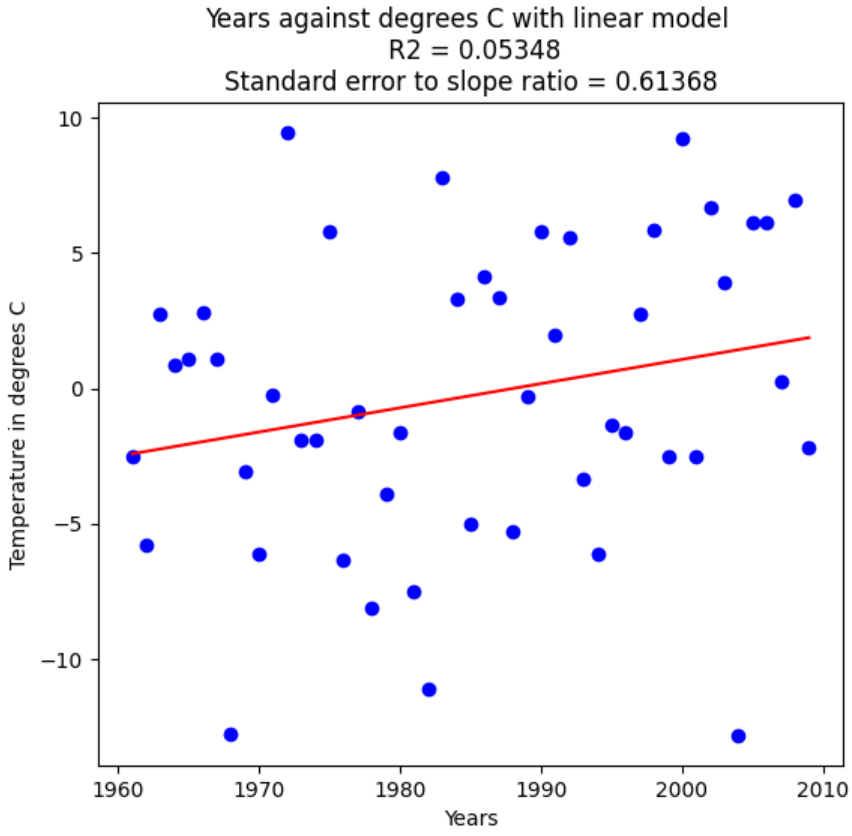
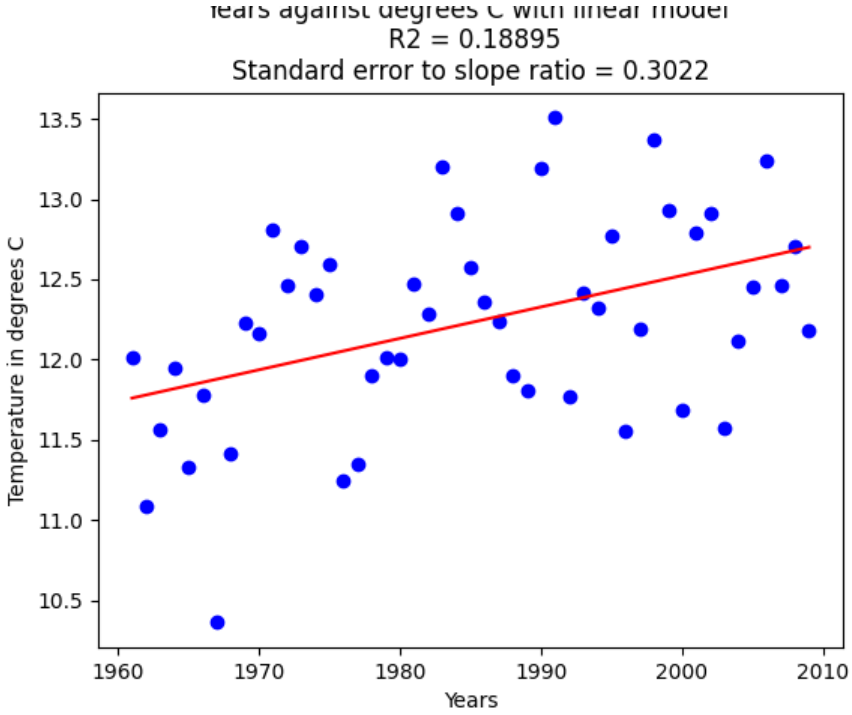


Figure B: Average annual temperature of New York from 1960 to 2010



The average annual temperature (figure 2) has less variation, due to larger sample size and regression towards the mean. This leads to a better fit of the linear model, the R squared value for figure B is 3 times higher, however, they are both very low, with figure A at 0.05 and figure B at 0.18, so most of the variance is not predicted by the model. Both graphs support claim that global warming is occurring, with both showing an increase in temperature. Figure A has a larger increase in temperature, from around –2.5C to 0C, with figure B having an increase from 11.75 to 12.5C. Since they are both over the same period figure A's model has a steeper gradient. However, Figure B provides more convincing evidence as it has a lower ratio of the standard error to slope, of 0.3 to 0.6, which makes it less probable to be luck.

Both graphs are very noisy, this is because there are many factors that influence the weather, from pollution to wind patterns and speed, therefore it is extremely hard to find a model that takes these variations and factors into account, especially without having data about level of factors such as windspeed. However, there is less variation in the annual mean, as it is an average of many days, so a larger sample size, which has a higher probability to represent the total population. In short figure B is more likely to portray the real increase in global temperatures.

Figure C: Average annual temperature of 21 cities in USA from 1960 to 2010

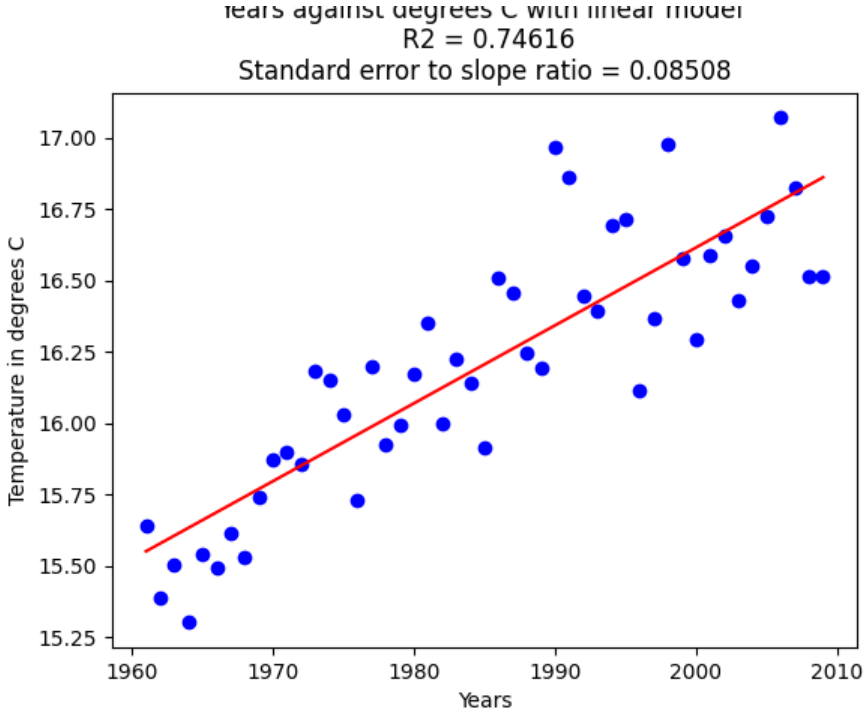


Figure C has less variation and noise than figure B and A, this is because it is a larger sample size, therefore more representative of the total population, likely closer to the mean of the total population. Less variation leads to a very high R squared value of 0.74, so most of the variance is accounted for. Furthermore, the very low standard error to slope ratio suggests the trend is not luck and there is indeed global warming occurring. The model shows a 1.25C increase in temperature, very few anomallies, all the points support the claim that global warming is occuring.

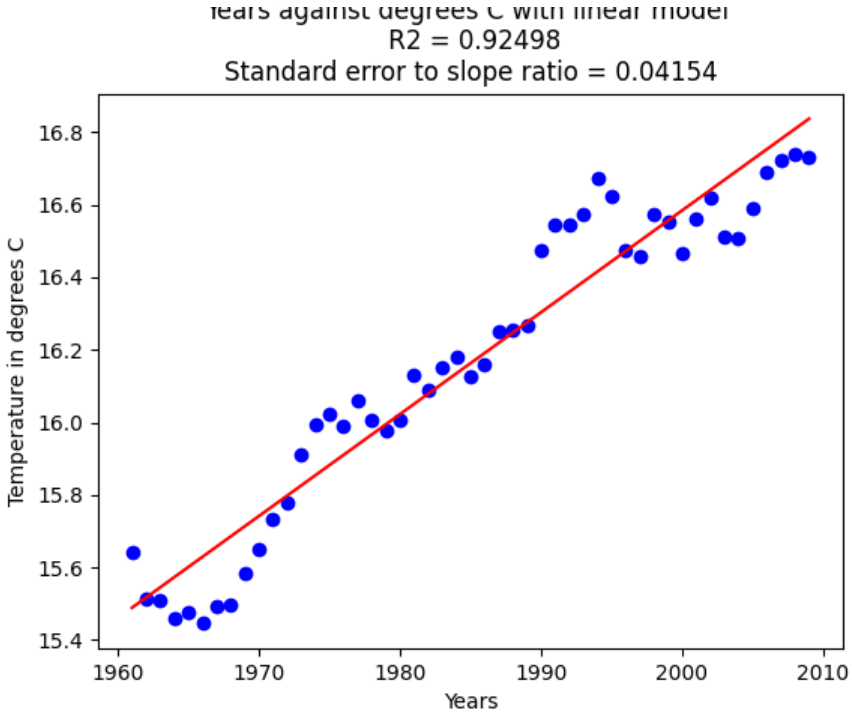
This is due to the very lare sample size, average of all daily temperatures to make yearly temperature for 21 different cities, then the averae of those 21 different cities. This means the sample will reflect the population better, there will be a smaller standard error.

If we used 3 different cities there would be more variation and noise as the sample size would be smaller, this would lead to a worse model, at least in terms of both R squared and standard error to slope ratio. Using 100 different cities would lead to less variation as there is a bigger sample size, therefore the mean of the sample will be more likely to be closer to the mean of the actual population.

If all the cities were in the same region there would be a bias in the sample chosen, therefore they might not represent the total population. They may also have a higher variance and more anomalies as factors such as cloud

amount or sun coverage would have an influence on all the cities, so would have a disproportionate effect on the cities and therefore mean temperature. The trend would not be as clear nor would our confidence in the trend be as certain.

Figure D: 5-year average of temperature of 21 cities in USA from 1960 to 2010



It has less variation, as it is a bigger sample size therefore closer to the mean of the actual population, any anomalies are averaged out, therefore given less weight. This makes the linear model a better fit, accounting for 92% of the variation, with very little chance of this being by accident. Therefore, it strongly supports the claim that global warming is occurring. Theres is a 1.3C increase in temperature according to the model.

Figure E: 5-year average of temperature of 21 cities in USA from 1960 to 2010 with quadratic model

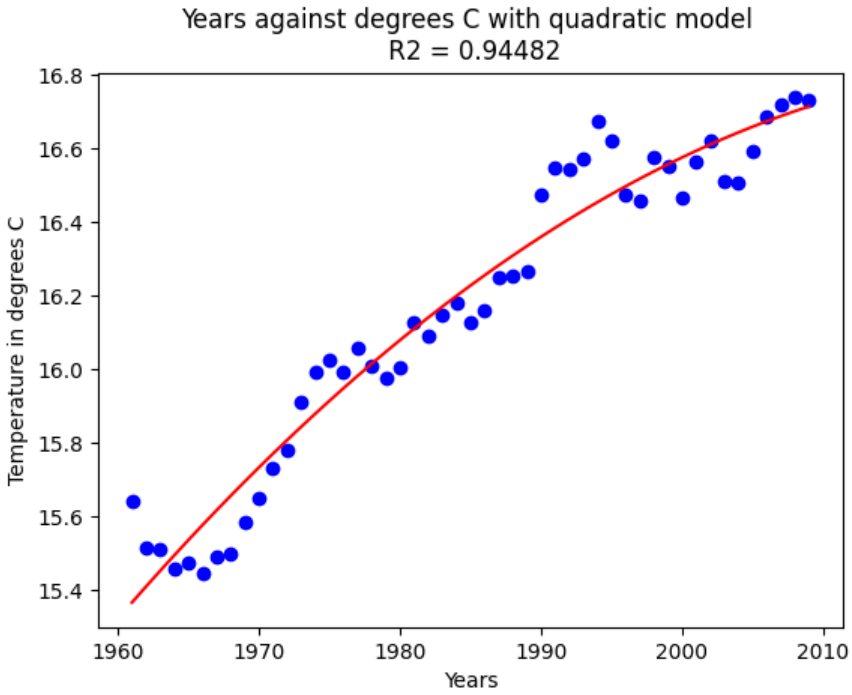
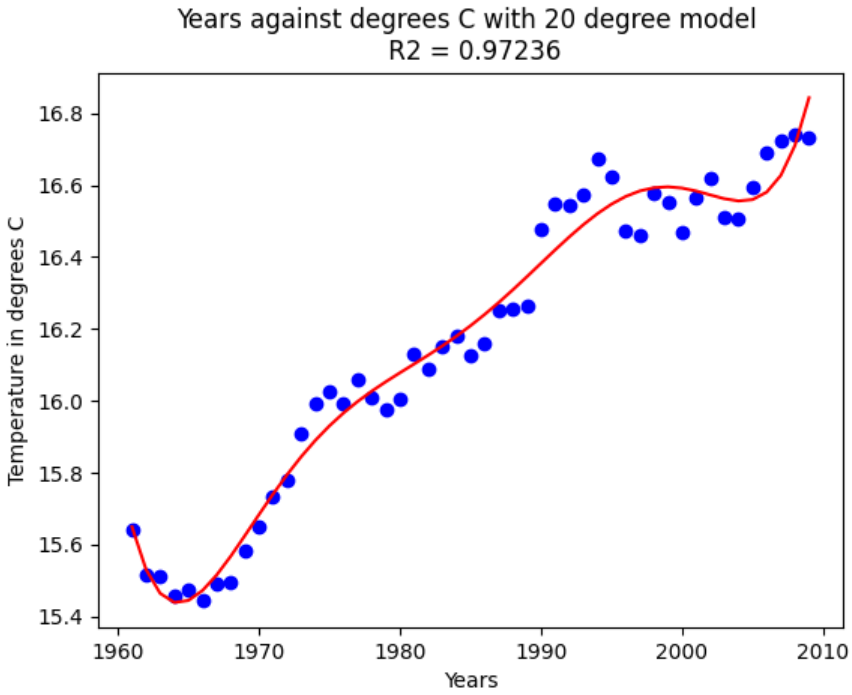


Figure F: 5-year average of temperature of 21 cities in USA from 1960 to 2010 with 20-degree model



All models fit the data very well, with the worst fit being the linear model, which accounts for 92% of the variation, there is a slight improvement with the quadratic model to 94% of the variation and the 20degree model fits 97% of

the variation. However, by the 20-degree model the curve seems to be fitting the noise as it is doubtful that the underlying temperature trend is so curved. The quadratic model is probably a better predictor of temperature as it goes through the middle of most points while ignoring anomalous increases in temperature.

The 20-degree model has the best R squared value as it fits the best to the data points as it has more coefficients that can be changed to reflect the variance in the data. This model also fits the data the best for the same reason, if a coefficient isn't needed then it will go to 0, therefore the higher degree model will always have the best fit to the data.

Predicting data

Figure G: 5-year average of temperature of 21 cities in USA from 2010 to 2015 with linear model from testing data set

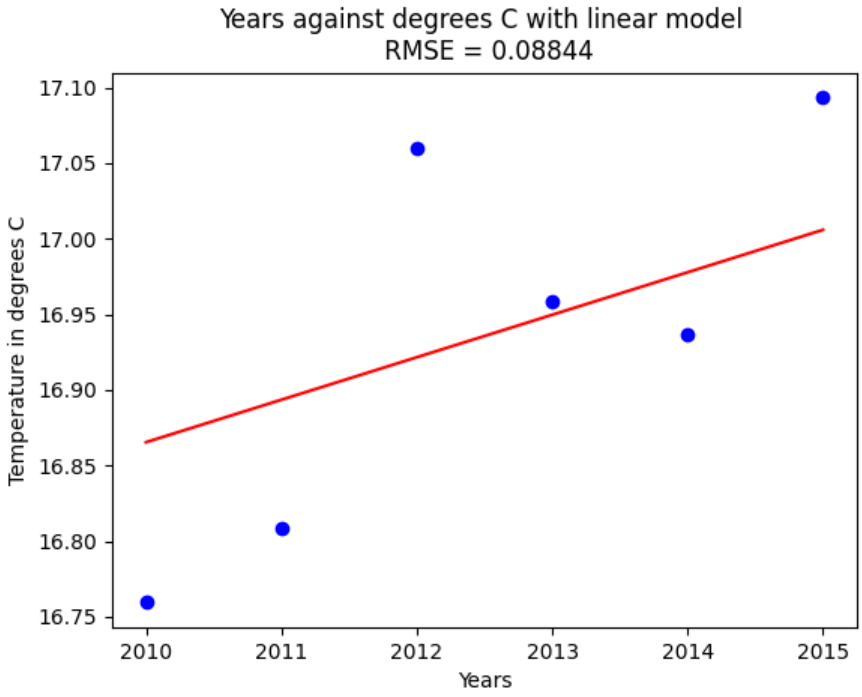


Figure H: 5-year average of temperature of 21 cities in USA from 2010 to 2015 with quadratic model from testing data set

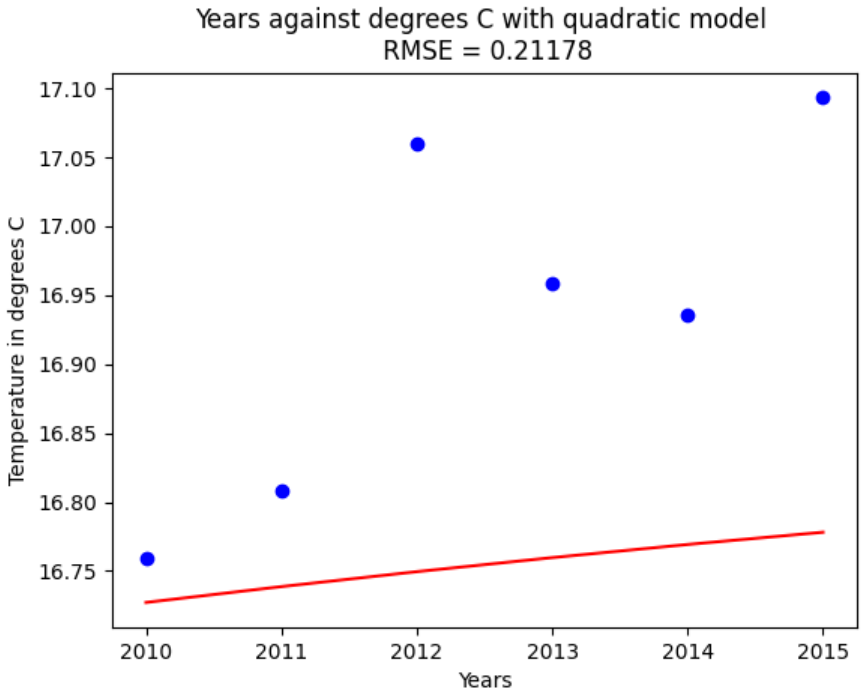
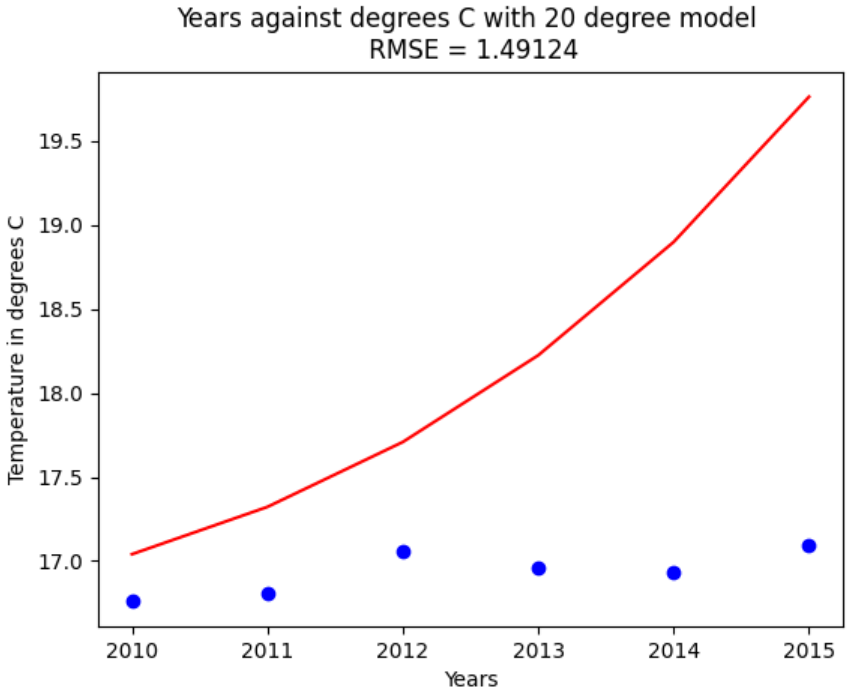


Figure I: 5-year average of temperature of 21 cities in USA from 2010 to 2015 with 20-degree model from testing data set



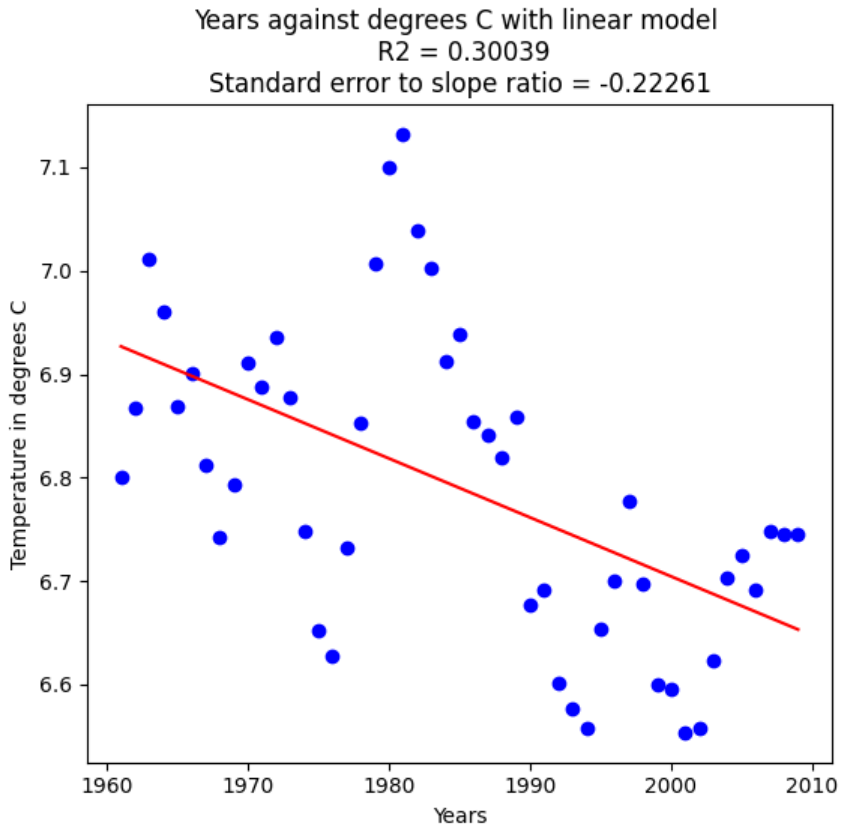
The linear model is by far the best model, with a very low RMSE of 0.09, which shows that most of the variance is accounted for by the model, it is also easy to see that the trend is clearly shown in the model and that it is roughly in the middle of the data. However, the quadratic model has overfit the training data and performs worse on the testing data, with an RMSE of 0.21. The 20-deree model performs even worse, with an RMSE of 1.49. Both the quadratic and 20-deree model overfit the training data and therefore do not generalise well

to new data and outside of their current range, therefore they are very poor predictors of actual temperature. The extra coefficients which helped them fit the training data the best let the model fit to the noise rather than the underlying signal or trend, while the linear model did fit the trend.

If we had used the data from New York's annual temperature instead of the 5-year moving average over 21 cities then the data would have had a lower signal to noise ratio, there would have been more noise. Therefore, the models would have had a less good fit and RMSE values as they would not have been able to predict the noise as it is random. Furthermore, with the quadratic and especially 20-deree polynomials the model would have overfit the data even more leading to even poorer predictions.

Part E: Modelling extreme temperatures

Figure J: 5-year average of standard deviation of temperature of 21 cities in USA from 1960 to 2010



The model does not match the claim that temperature variation is getting larger, it suggests there has been a decrease in temperature variation of around 0.2 standard deviations. This is further confirmed by the low Standard error to slope ratio of –0.22. However, the R squared value is low at 0.30, therefore most of the variation is not accounted for by the model, furthermore, the decrease in standard deviation is very small. Overall, the graph claims there has been a negligible decrease in variation of temperatures, as the low r squared value indicates there is too much noise or variation for the linear model to capture.

To improve the analysis, I would use a larger range, as this is clearly a very long-term phenomena and would also increase the sample size to include temperatures from around the world, as there is currently a bias towards the USA in the sample, so may not reflect the population. May want to average the standard deviation over a larger window, say 10 years, to reduce impact of anomalies, increasing signal to noise ratio. May want to generate different degree polynomial models and test them on the testing data to see which fits the data the best, from the best model can then have an approximation for change in variation of temperature.