Skills development: data analysis

Project 1:

Consider the following data:
Timeseries of global annual temperature anomalies for 1880-2017 (reconstructed from climate models and observations) – plotted in Figure 1
Timeseries of measured annual CO2 concentrations in the atmosphere from Mauna Loa observatory for 1959-2017 – plotted in Figure 2

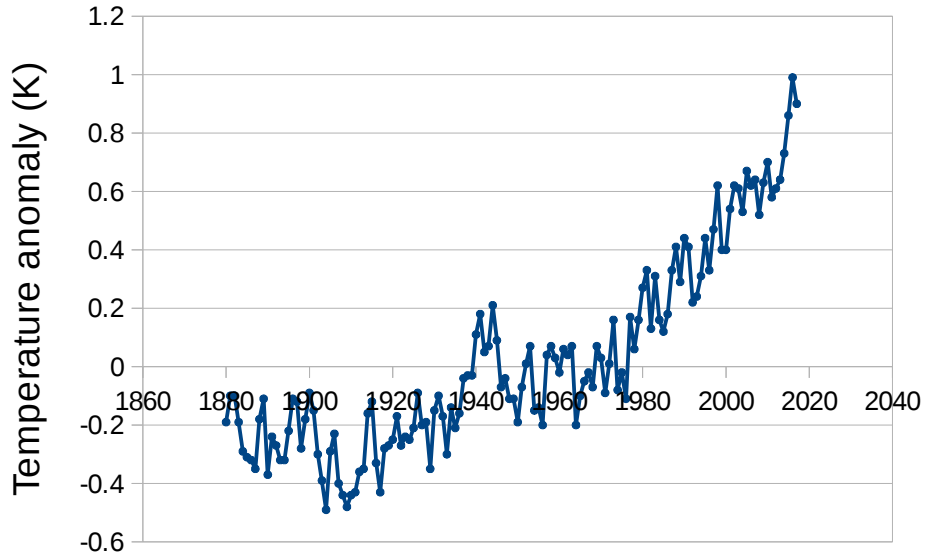

Figure 1

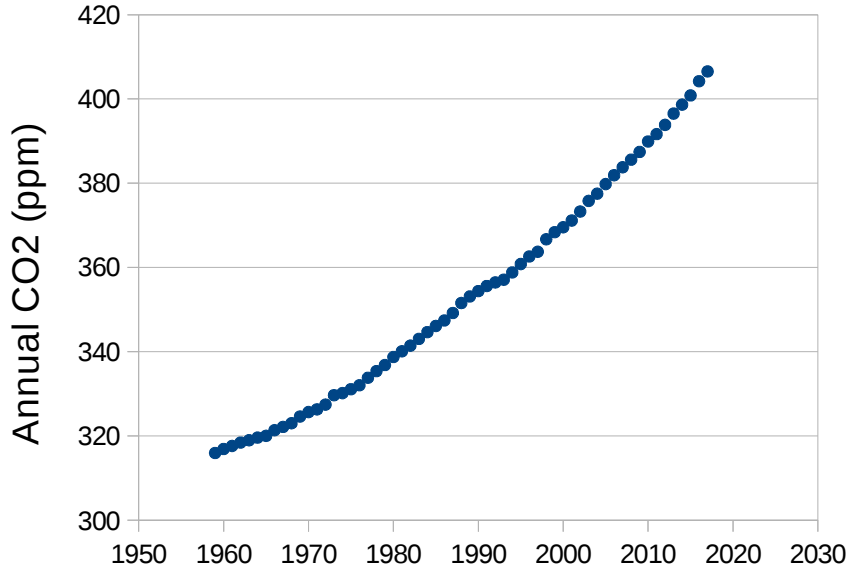

Figure 2

Research questions/objectives:

1) Is there a long-term trend in the global temperature data and is the trend significant? How different is the trend over the last 50 years in comparison to the overall trend (1880-2017)?

2) Is there a long-term trend in CO2 data and is the trend significant?

3) How strongly are temperature and CO2 timeseries linearly related and what is the linear relation between the two? How good is the linear model that relates temperature to CO2 (i.e. how much variance in temperature is explained by this model)?

4) If CO2 is twice its 2017 value, how much the global temperature would change according to the linear model? If CO2 concentration continues to rise with the same rate per year until 2100 what would be the global temperature in 2100 according to the linear regression model?

5) What would be the CO2 concentration in year 1880 according to the linear regression model between CO2 and temperature?

6) Is there a relationship between year-to-year variability in global temperature data and year-to-year variability in CO2 data?

Instructions and guidelines:

Feel free to use any computing environment/software (e.g. Excel, Matlab, R, Python, ...) to perform the data analysis. The data is given in **Data.xls** file. In the worksheet labelled as 'Data', the fist column contain years, the second column contains temperature and the third contains CO2. The second sheet, labelled as 'Temperature', gives the solution to question 1: the trend calculation and its significance (T-statistics).

Start by opening the data (or loading it in the computing software of your choice) and plotting it. See if you can reproduce the Figures 1 and 2.

Specific instructions for each question + guidelines for students using Excel:

1) Use the linear regression (ordinary least square method) to calculate the trend. See the guidelines in '*Stats_recap.pdf*' presentation from Day 5 (on Canvas) on how to derive the slope and intercept. Plot the trend-line on top of the original timeseries. Calculate the T-statistics and determine whether the trend is significantly different from zero at your choice of the confidence level (note: 5% and 1% are the most commonly used confidence levels).

2) Same steps as in (1) but for CO2 timeseries.

3) In a new sheet copy the temperature and CO2 timeseries for the period they overlap (1959-2017). It's always a good practice to keep the raw data (the first worksheet) intact. In this new sheet, perform the correlation analysis between the two timeseries. Plot a scatter-plot with temperature on y-axis and CO2 on x-axis and perform a linear regression (y=ax+b) on this data. In a new column calculate the regressed temperature (y=ax+b). Plot the regressed temperature (trendline) on top of the scatter-plot. Calculate the correlation coefficient (r) between the regressed temperature and the original temperature -> r^2 tells you how much variance is explained by the linear regression model. You can also plot a scatter-plot with regressed temperature on y-axis and original temperate on x-axis to visually inspect the model performance.

4) Using the model from above (y=ax+b) try to answer the first question. For the second question: look back on the trend of CO2 from (2) and extrapolate that trendline further to 2100. What CO2 values do you get for year 2100? Once you estimate that value, use the model (y=ax +b) to determine the temperature for year 2100.

5) This question requires you to perform the linear regression between CO2 (y-axis) and temperature (x-axis). So you need to find a new linear regression model y=Ax+B. Once you calculate the values A (slope) and B (intercept), use the temperature values from year 1880 to determine CO2 value for that year.

6) Similarly to the question (3), in a new sheet copy the temperature and CO2 timeseries for the period they overlap (1959-2017). Observe that in both timeseries the long-term trend over the whole period is the most dominant signal. Therefore, we first need to get rid of the trend in order to see how the data fluctuates year-to-year. In other words, we need to perform de-trending of the original timeseries.
Steps for detrending (option 1):

- Calculate the trendline (linear regression) for both temperature and CO2 -> you have already done this in (1) and (2). Add a column that contains values of the trendline for temperature (T_trendline=trend_T*time+intercept_T) and add a column that contains values of the trendline for CO2 (CO2_trendline=trend_CO2*time+intercept_CO2).
- Subtract the T_trendline values from original temperature (T) values (for each year) and save the values in a new column -> this is the detrended temperature signal. Perform the same for CO2.
- Plot separately the detrended T and detrended CO2 timeseries.
- Perform the correlation analysis (calculate r) between the two detrended timeseries. You can also plot the two timeseries in a scatter-plot (detrended T versus detrended CO2).

Steps for detrending (option 2):
Another approach for detrending is by using a moving-average (running-mean). This approach is probably better for the CO2 data as the detending approach above did not really produce year-to-year fluctuations for CO2 (assumption of the normal distribution of residuals in the linear regression does not really apply in

this case).

- Start by calculating a 5-year moving average for temperature and save the values in the centre of the averaging window, i.e. find the average temp for 1959-1963 and save that value for the year 1961, then find the average temp for 1960-1964 and save that value for the year 1962, then find the average temp for 1961-1965 and save that value for the year 1963, and so on. Your last centred window will be year 2015 and it will contain the average CO2 value for 2013-2017. In this way the moving-average will not have the values for the first two years at the beginning of the period and the last two years at the end of the original period, i.e. the moving average is given for the period 1961-2015.
- Repeat the same procedure for the CO2 timeseries.
- Plot separately the moving-average timeseries for T and CO2.
- Similarity to the detrending in option 1, you need to remove the moving-average timeseries from the original timeseries (this can be only done for the overlapping period of 1961-2015). This subtraction will give you the residual timeseries for 1961-2015. Plot those residual timeseries.

If all goes well you should be getting the following plot for the residual CO2 timeseries:

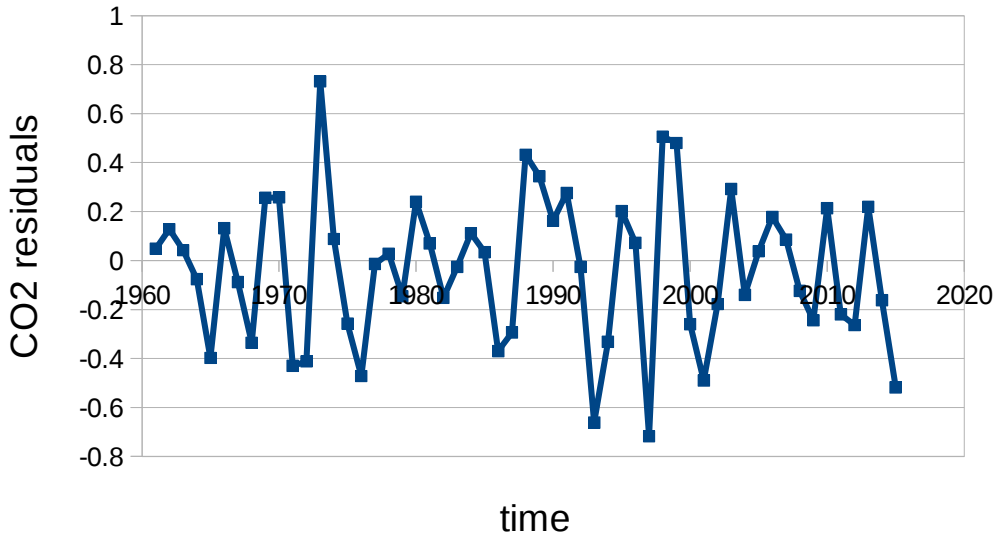

Figure 3: Residual CO2 timeseries (when the 5-yr moving average is removed from the original timeseries).

Finally, you can perform the correlation analysis between the detrended (residual) timeseries of T and CO2. Try also correlating the detrended temperature timeseries from option 1 with the residual CO2 timeseries from option 2. You can also test the sensitivity of your results (correlation coefficient) to the choice of the moving-averaging window (e.g. 3-yr, 7-yr moving average). Which moving-averaging window gives you the highest correlation coefficient between residual T and residual CO2 timeseries?