Lab 3: Exploring UBC and global temperature time series: trends, regression and causality.

In the upcoming lab you will use linear regression, correlations and multi-linear regression to ask the following three questions for the UBC weather station/global temperature time series:

-Is there a trend in the temperature time series?

-Are temperature time series correlated with external or internal factors (e.g. atmospheric CO2, El Nino oscillation)?

-Are your results compatible or not with the IPCC reports, or can they prove or contradict there is a global warming trend and that it is caused by anthropogenic CO2 emissions?

In this pre-lab assignment you will:

1. Review variance, covariance, correlation and linear regression
2. Review basic notions on climate forcing and El Nino Southern Oscillation

Assessment

1. Online pre-lab quiz on *Connect*

1. Variance, Covariance, and Correlation

The ***variance*** of a variable is the average of the square of the difference between each individual data point and the mean of all the data. It is a measure of the spread of the data. Large variances imply a large spread in the data points, and small variances imply the data points are all close to the mean.

The variance V of a variable *x with N datapoints* is defined as:

$$V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

where $\bar{x}$ is the average of x. Notice that the square-root of the variance is the standard deviation.

Covariance is a measurement of the variance shared between two variables. When the covariance is positive, it implies that when one variable increases, the other tends to increase as well; when the covariance is negative, it implies that as one variable increases, the other decreases; and when the covariance is 0, it implies that there is no statistical relationship between the two variables.

The covariance between two variables x and y is defined as:

$$CV(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Notice how similar this is to the definition of the variance of one variable. You can see that the covariance of a variable with itself is just the variance of the variable.

Correlation is the covariance divided by the standard deviations of the two variables:

$$r(x,y) = \frac{CV(x,y)}{\sigma_x \sigma_y}$$

where σ_x and σ_y are the standard deviations of x and y.
Dividing the covariance by the standard deviations of the variables *normalizes* the covariance -- it transforms the covariance to a standard scale of measurement. The magnitude of a covariance depends on the magnitude of the variance of each variable, but the correlation is always a value between -1 and 1. This allows for fair comparisons of the relationships between variables with different units.

An important measure of the reliability of a correlation is the ***p-value***. The p-value of a correlation is the probability that two uncorrelated, random variables could have the measured correlation simply by chance. Generally, if the p-value of a correlation is less than 0.05, we say that "the correlation is significant at the 95% confidence level". In the lab, you will learn how to calculate both a correlation and its associated p-value using MATLAB.

Linear regression is the best fit linear relationship between a dependent variable y and one explanatory variables (or independent variables) x. In a linear regression, one assumes that there is a relationship between x and y of the form y=a.x+b and find the values of the coefficient a and b that minimize the least square error $\sum_{i=1}^{N}(y_i - a.x_i - b)^2$.

The goodness of fit of a linear regression can be represented by the coefficient of determination

$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - a.x_i - b)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$ which is the percentage of variance of y explained by the linear regression model.

Linear regression and the correlation coefficient are closely related. In particular, the slope a of a linear regression between y and x is related to the correlation coefficient between x and y via $a = r(x,y) * \sigma_{y/}\sigma_x$ and the coefficient of determination is the square of the correlation coefficient.

Multilinear regression is similar to the simple linear regression but uses several independent variables x_1, x_2, x_3,… The assumed relationship between the dependent variable y and independent variable is of the form $y=a+ b.x_1 +c.x_2 +d.x_3$…

2. Brief overview of climate forcing and internal variability

Climate forcing

The critical process influencing Earth temperature is the balance between the energy incoming from the Sun and the energy radiated out by Earth. This is called Earth radiative balance. Some factors can influence and modify this balance; these are called climate forcing. In the lab, we will explore the influence of the forcing below on our local and Earth global temperature:

-**Total solar irradiance** is the flux of energy received by Earth from the Sun, at the top of Earth's atmosphere. This flux is not constant due to variations in solar activity, or to changes in Earth's orbital parameter that impact the average Sun-Earth distance over a year. Do you think that Earth temperature increases or decreases when the total solar irradiance increases?

-**Stratospheric aerosols optical depth (volcanic forcing)** is a measure of how aerosols in the upper atmosphere prevent the sunlight from reaching Earth's surface. The main factor influencing these aerosols are volcanic eruptions. Explosive volcanic eruptions produce column of ashes and gases that can reach the upper

atmosphere. Volcanic sulfur gases produce sulfate aerosols – tiny droplets of sulfur and water – which reflect some of the incoming sunlight back to space. When these aerosols settle in the lower part of the atmosphere, water can easily condense on them resulting in enhanced cloud formation. Do you think that Earth temperature increases or decreases when there is a large volcanic eruption?

-**Anthropogenic CO2 emissions:** atmospheric CO_2 is a gas transparent to the shortwave radiation of the Sun, but opaque to the longwave radiation of Earth. Thus, when atmospheric CO2 concentration increases, less radiation is emitted to space by Earth. In this lab, we will use atmospheric CO2 concentrations which are mostly explained – for last century – by anthropogenic CO2 emissions. Note that many natural factors, e.g. marine biology or volcanic eruptions, impact the CO2 cycle. Do you think that Earth temperature increases or decreases when atmospheric CO2 concentration increases?

-**Anthropogenic aerosols** are tiny droplets formed by some pollutants released by human activities, such as sulfur dioxide or black carbon. Their impact on the radiative balance depends on their exact composition but overall, their effects are similar to volcanic aerosols: they make the atmosphere more opaque to Sunlight and favor the formation of clouds. Note that there are many natural source of aerosols, like terrestrial and marine biology or volcanic eruptions. Do you think that Earth temperature increases or decreases when human aerosols emissions increases?

Internal variability

Even if the radiative balance is a key control on climate, important variations in climate can happen without any change in the radiative balance. For example, local changes in ocean or atmosphere circulation can have global effects and impact climate. They are refer to as internal climate variability because they result of processes specific to ocean and/or atmosphere dynamics.

In lab 3, we will mostly explore the relationship between the El Niño-Southern Oscillation (ENSO) and temperature, as it is one of the mode of climate variability impacting the most Earth global temperature.

The El Niño-Southern Oscillation (ENSO) is a pattern of climate variability that affects the climate of the whole planet. It is composed of two components: El Niño in the ocean, and the Southern Oscillation in the atmosphere.

El Niño is a quasi-periodic oscillation in the surface temperature of the eastern equatorial Pacific. During an El Niño event, the surface water in the eastern Pacific is warmer than normal; eventually this pattern reverses and becomes a La Niña event, during which the surface water is cooler than normal (Figure 1).

The Southern Oscillation is an oscillation in the atmospheric pressure over the eastern Pacific. Higher than average atmospheric pressure causes easterly (east-to-west) winds to blow along the equator, and lower than average pressure creates westerly winds.

However, as you know the two phenomena are linked to each other. The Southern Oscillations wind patterns affects the ocean's surface temperature along the equator, while El Niño's temperature patterns affect the air pressure, causing El Niño and the Southern Oscillation to be *coupled.* This means that the two oscillations strongly affect each other. As you're probably also aware, the ENSO cycle strongly affects climate patterns in many regions of the world including north-west coast of North America.

There are several ways to measure the state of ENSO. In this lab we will be working with the MEI, the Multivariate ENSO Index which you already loaded and plotted in Lab 1. The MEI uses measurements of sea-level pressure, zonal (east-west) and meridional (north-south) surface wind, sea surface temperature, surface air temperature, and cloud cover to create an estimate of the ENSO state. When the MEI is positive, the ocean is in an El Niño state and the equatorial winds are eastward; when the MEI is negative, the ocean is in a La Niña state and the equatorial winds are westward.

The MEI is an anomaly index, which means that it measures ***deviations from a background state***. The definition of a "background state" is a little bit arbitrary, but it is usually just the monthly or annual average of the entire available range of data. Keep in mind that an anomaly time series is always an anomaly from some calculated average, so that when you create an anomaly time series, you must mention somewhere what you are using as a background state. In the case of the MEI, the background state is the average ENSO conditions between 1950 and 1993.

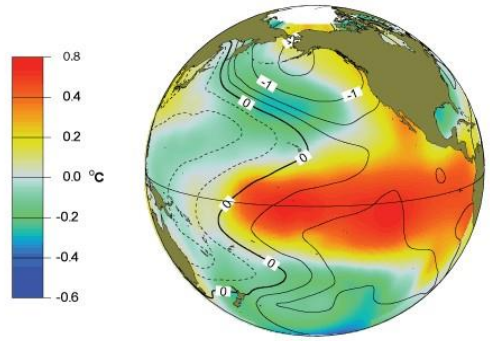

Figure 1. Warm Phase ENSO (El Niño). The spatial pattern of anomalies in sea surface temperature (shading, degrees Celsius) and sea level pressure (contours) associated with the warm phase of ENSO for the period 1900-1992. Contour interval is 1 millibar, with additional contours drawn for +0.25 and 0.5 millibar. Positive (negative) contours are dashed (solid). Figure from http://cses.washington.edu/cig/pnwc/aboutenso.shtml

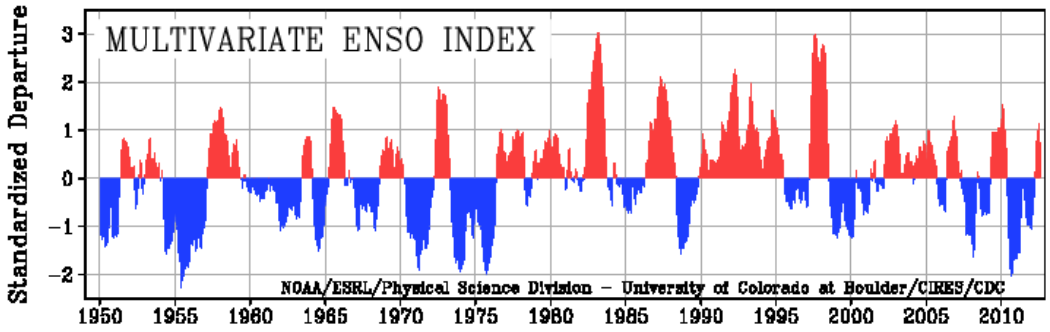

Figure 2. MEI index time series from 1950-2012. Positive values indicate El Niño events, negative values indicate La Niña events.

3. MATLAB Skills

Correlation coefficient

To calculate the correlation between two variable x and y, you can use the function `corrcoef()`. The result of running this command is a little complicated. The command `corrcoef(x,y)` returns a 2x2 matrix containing four different correlations:

- corr(x, x) in the top left
- corr(x, y) in the top right
- corr(y, x) in the bottom left
- corr(y, y) in the bottom right

The correlations in the top left and bottom right are both unity (1). The values in the top right and bottom left are non-zero, but the same. This value is the one in which you're interested.

The corrcoef function does not remove missing values automatically so you will have to do it if your dataset contains any missing value. Since the locations of the missing values differ in the two time series x and y, *you have to create a mask that identifies* `NaN` *values in x OR in y* and remove the values identified by your mask from both arrays.

You can get `corrcoef()` to return the p-value of the correlation if you give the function another variable to place the p-value into:

```
[R, p] = corrcoef(x, y)
```

In this form, the variable `R` contains the 2x2 correlation matrix as before, while the variable `p` contains the probabilities that the correlation could have happened by random chance. Low p-values means it is unlikely that the correlation happened randomly. Generally we want a p- value to be less than 0.05 (a 5% chance of this correlation happening randomly) to declare the correlation statistically significant.

Multilinear regression

To perform a linear regression in matlab, use:

```
[coef,bint,r,rint,stats] = regress(y,X)
```

The output of the function regress will be stored in b, bint, r, rint and stats and will be described below. Note that you can use any name you like for the outputs! The inputs are the dependent variable, y, which should be an *n*-by-1 column vector of observed responses. For example, if you are trying to explain temperature variations based on various climate forcing, y should be the observed temperature. The second output, X, should be a matrix in which each column is one of your independent variable, for example climate forcings. The first column should just be filled with ones if you want the intercept to be non-zero.

Concretely, let's say that you want to regress the temperature T against the climate forcing F_1, F_2 and F_3. First, make sure that T, F_1, F_2 and F_3 are column vectors of the same size. Then, use

```
[coef,bint,r,rint,stats] = regress(T,[ones(size(F1)) F1 F2 F3])
```

Where [ones(size(F_1)) F_1 F_2 F_3] is the input matrix with the first column filled with 1.

In this case, the fitting relationship would be $T_{pred}=a+b.F_1+c.F_2+d.F_3$ where T_{pred} is the temperature predicted by the multilinear regression model. Note that a will be equal to 0 if you don't add the columns with 1 in the independent variable matrix.

The main outputs will be:

-The vector coef, which contains the coefficients a, b, c and d
-The matrix bint, which contains 95% confidence intervals on the coefficients a, b, c and d
- The vector stats, which first element is the coefficient of determination and third element the p-value testing the strength of the assumed linear relationship

The temperature predicted by your model is then $T_{pred}=X*coef$ where X=[ones(size(F_1)) F_1 F_2 F_3].

Last remark: to perform a simple linear regression, you can use the exact same function. Your vector X will just contain 2 columns: one filled with 1 and one containing your independent variable.