UNST 124g fall 2011
historical climate records

1 introduction

Informally, we can define climate as the typical weather experienced in a particular region. More rigorously, climate is a statistical description of weather variables over a specified period of time. The statistical description would include the mean but also variation around the mean, and perhaps also something about cycles and trends over time periods long enough to make them statistically meaningful. The standard period for calculating the climate-related means is 30 years, defined by the World Meteorological Organization. The years 1951 to 1980 are used to define the mean in most analyses you will see that deal with global warming and climate change. Some common climate variables are temperature, precipitation, atmospheric pressure, humidity, and wind measured at Earth's surface.

The Dust Bowl Era, from 1931 to 1939, was defined by a series of deep droughts and extensive soil erosion in the southern Great Plains. Coinciding and resonating with a global decline in wheat prices and the economic effects of the 1929 stock market "crash," the Dust Bowl Era was a time of great economic and social dislocation throughout the United States.

Drought is a recurrent feature of global climate, though its features vary from region to region. A simple, conversational, definition of drought is "an extended period of lower-than-average precipitation, often coincident with higher-than-average temperatures." Drought represents variation from the mean state in a region (and is thus different from aridity). Drought is a progressive phenomenon, with meteorological, agricultural, and hydrological aspects. Meteorological drought is defined according to the degree of dryness compared to average conditions and the duration of the dry period. Agricultural drought links various meteorological drought to agricultural impacts. Hydrological drought links meteorological drought to declines in surface or subsurface water reservoirs. The longer the drought period, the greater impact on its agricultural and hydrological components.

It this exercise, we will examine temperature and precipitation at two of the towns in *The Worst Hard Time* (Egan, 2006) and in McMinnville, Oregon, as well as global mean annual temperature. You will be asked to use data arranged for you in an Excel workbook to perform some simple calculations and make graphs of the data. At the end of this document you will find numbered questions that are intended to assist you in studying these data sets. You should write answers the questions using complete sentences and short paragraphs. Where the question asks for a plot from your data analysis, print the graph, number it, and refer to that number in your answer.

It is important in this analysis to remember the difference between climate and weather, as well as the spatial variability in climate that we have discussed in lecture. It is often useful to compare the trends you discover at a particular site to regional or global trends.

2 the data

Climate records of one sort or another have probably been kept as long as people have been farming, perhaps longer. Such information forms the knowledge base upon which decisions about planting and harvesting can be made. "Modern," instrumental records have been kept since the middle of the 17th century, although very few are continuous before the 18th century. By the start of the 19h century, reliable and continuous records were being kept in Europe and East Asia. More widely available and reliable records pick up the mid 19th century (North et al., 2006, Chapter 3). In North America, reliable climate records begin in the late 19th century.

When you open the Excel workbook, you will see four tabbed worksheet pages. The tab labels tell you what is stored in each worksheet. Column headings at the top of each sheet indicate the data type and measurement units. The records do not all start in the same year and not all of the records are complete. This is the nature of any observational data set.

2.1 station data

Mean monthly and seasonal temperature and precipitation records for Boise City, Oklahoma (36.7 N latitude, 102.5 W longitude), Liberal, Kansas (37.0 N latitude, 100.9 W longitude), and McMinnville, Oregon (45.2 N latitude, 123.1 W longitude), have been assembled for you into an Excel workbook, using the temperature station data archive maintained by NASA's Goddard Institute for Space Studies and precipitation data from the United States Historical Climatology Network project at Oak Ridge National Laboratory, and the Desert Research Institute. You can download the workbook at the class website. Both of these long-standing records are collected at regional airports. You can read more about the stations and see pictures of the station sites at `http://weather.gladstonefamily.net/site/KMMV`, `http://weather.gladstonefamily.net/site/KLBL`, and `http://weather.gladstonefamily.net/site/K17K`.

Seasonal values are calculated: D-J-F (December-January-February) for the northern hemisphere winter season; M-A-M (March-April-May) for spring; J-J-A (June-July-August) for summer; and S-O-N (September-October-November) for fall. You will notice that some of the spreadsheet cells are blank. This indicates that no data is available for that observation interval. Record keeping can be variable for a number of reasons. Maybe the thermometer broke and there were insufficient funds to replace it or perhaps the pages on which the data were written have been lost.

2.2 global surface temperature analysis

The NASA Goddard Institute for Space Studies makes all the data they use in their global temperature anomaly analyses available online in a variety of formats. The homepage for everything they produce is `http://data.giss.nasa.gov/gistemp/`. Many details about their long-standing temperature analysis program can be found at the website. We will use only the mean annual global land and sea surface temperature anomaly record created using weather station, ship-board, and satellite data. Details of the analysis are presented in Hansen et al. (2010). The combination of land and ocean surface observations produces a slightly different anomaly than would land surface alone due to the large heat capacity of water. The anomaly is calculated relative to a base period from 1951 to 1980.

3 data analysis

3.1 mean values

The mean value for a set of numbers is the sum of the individual values y_i divided by the number of values n. A bar over a variable name usually indicates a mean value, for example

$$\bar{y} \;=\; \frac{\sum_{i=1}^{n} y_i}{n} \tag{1}$$

In this project you will use seasonal means, computed using sets of months representing the seasons, and *running* means (also called a moving average). A running mean is a series of mean values calculated along the data set using subsets of the data. Running means are helpful here because they smooth out some of the year-to-year noise that can make plots of annual data hard to interpret.

3.2 trend analysis

Trend estimation is a statistical method often used as a first step in studying the relationship between two variables. In our case, we are interested in the relationship between time (years in our observation record) and air temperature. Trend analysis tells us *if* a relationship exists but it doesn't tell us *why* it exists. For that, we need to know something about the physical processes involved in the system we are studying.

Trend analysis is performed on data sets for which we expect a meaningful relationship because we understand at least a little bit about the system they represent. For example, we expect that the temperature we measure today is related to whatever the temperature was yesterday. We must, however, be prepared for the case when no relationship exists, that is, the data are random. A random quantity may change in an unpredictable way every time it is measured. There should be no discernible relationship—no trend—in a series of measurements of a random variable. Statisticians have formal mathematical rules for determining the sample size required for meaningful analyses and the confidence that may be placed on the result.

In this exercise you are asked to examine the relationship between mean annual temperature and time in the temperature time series data. The Excel Trend Line tool computes this by finding the best fit of a line through the data. That tool allows you to fit a straight line—a line with constant slope—through the data (or something more complicated). We will use the Linear Regression tool to produce a line with constant slope. The *best fit* is one that minimizes the differences between the actual data points and the trend line for all the measurements in the time series.

The equation of a line made of of points (x, y) is

$$y \;=\; mx + b \tag{2}$$

where m is the slope of the line and b tells you where the line intersects the y-axis. The slope is the change in y over some interval of x, in our case the rate of temperature change in °C per year. The goal in linear regression is to find a single equation of a line that best represents all of the data in our series. For all of the x in the data set, we want to find the combination of slope m and y-intercept

b that gives the best *prediction* of the real values of y. This is accomplished using a mathematical formula—a linear regression—that minimizes the differences between the real and predicted y's for all of the x's. The differences between the real data values and values that lie exactly along the line can be used to calculate a single number that represents the quality of the linear fit to the data, often called the *R-squared* value.

In our example, the x's are years and the y's are temperature anomalies. The slope m is thus the rate of temperature change over the entire interval from the first to the last year in the record. The R-squared value tells us something about how much variation year-to-year there is around the linear trend. As you will discover by inspecting the data plots you produce in this project, a straight line is considerably more simple than the real time series. However, the linear best fit does provide a useful summary of the data.

The linear fit metric, R-squared, tells us something about how close the original data are to the line we fit through them but it does not tell us anything about the slope of that line itself. It does not tell us if the slope is *meaningful*, that is, if it is distinguishable from no slope at all. For that, we need a statistical test. A common test used for this purpose is the "Student's t-test," developed in 1908 by a chemist working for the Guinness brewery in Dublin, Ireland. The chemist was interested in calculating how different individual batches of stout were from one another so that he could monitor product quality.

Student's t-test compares two data sets, asking the question if they are distinguishable from each other. Here, the test we want to perform is to ask if the slope of the linear trend we calculated for the global temperature anomalies is different from zero. In the language of statistical tests, we would say that the null hypothesis is that the slope is not distinguishable from zero and the alternative hypothesis is that the slope is different from zero. In a nutshell, measures of variance in the data sets are used to compute a statistic—in this case the t test statistic—that can be compared against standard values for the number of measurements and desired level of confidence. In general, larger the desired level of confidence, the more measurements you need to have. The test values are complied in tables we can find in statistics textbooks and similar references.

To perform this test we need the slope calculated for us by Excel, and two measures of the variance of the data—the original data and the linear prediction of the data. We will use the variable name y for our original *dependent* data and $\hat{y}$ for the values predicted by the linear regression formula. When we call y the dependent variable we mean that its value goes with some other quantity—the *independent* variable—in this case x. For our problem, the year number is the independent variable and the temperature anomaly is the dependent variable. We also need to know the number of samples n and a quantity called the *degrees of freedom*. The degrees of freedom are just the number of independent measurements we have, in this case $n - 2$ because we already used our data to compute two variables, the slope and the intercept.

The predicted values $\hat{y}$ are calculated using the equation for a line (equation (2) above), the slope m and intercept b from the regression, and the original *independent* data values x. In our case the x are the years in the time series. The predicted values are thus

$$\hat{y} \;=\; mx + b \tag{3}$$

Next, we use the original values and the predicted values to estimate the variance of the original values about the regression. This is a type of measure of the difference between the observed and

regression-predicted values. It is calculated

$$MS_D = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-2} \tag{4}$$

in which the summation goes over all the samples n and $n-2$ is the degrees of freedom. The subscript i indicates the sample number. We also need to know something about how much variance there is in the original data. We measure this using a quantity called the corrected sums of squares

$$SS_y = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{5}$$

Finally, the test statistic for the significance of the slope determined by the linear regression is calculated

$$t = \frac{m}{\sqrt{MS_D/SS_y}} \tag{6}$$

We then use a t-statistic table to evaluate the null hypothesis. If the value of t that we calculated is larger than the minimum in the table, the we reject the null hypothesis and if the value smaller than what appears in the table, we must accept the null hypothesis. In this case, the null hypothesis is that we can't tell the difference between the slope we calculated (the temperature trend) and zero.

Don't panic. Excel has several built in mathematical formulas that will help you complete this assignment. SUM, SQRT, SUMSQ, and COUNT can all be used to simplify your calculations. You have experience with some of these already from the food audit assignment. SUMSQ is new but it does exactly what the abbreviation sounds as if it does, it sums the squares of values. In this case you will want to use it for equation (4). The number of samples n can be measured using COUNT.

4 Questions

Note: Figures you produce should have labels for all three axes, a title, a legend, and should be easy to read.

1. Make plots of mean annual temperature and mean annual precipitation for Boise City (in Oklahoma, Boise is pronounced like voice), Liberal, and McMinnville using **XY (Scatter)** with both points and a straight line connecting the points. Please include all three cities on one chart for temperature and one for precipitation. You can include the charts in the main body of your text if is is large enough to read, otherwise print it separately. Be sure to take the time to make charts that are easy to read.

   What is the hottest year on record in each of the sites?

2. In *The Worst Hard Time*, Timothy Egan argues that in the southern plains, the years preceding the Dust Bowl Era (1931 to 1939) were unusually wet, giving settlers moving into the area a false impression about the suitability of the region for farming. You can evaluate this assertion using the precipitation data sets to compute a precipitation anomaly at each city.

   Computing precipitation anomalies requires you to first calculate an standard value against which the others are compared. In this case, you can calculate the average of the total annual precipitation values for each data set using Excel's AVERAGE function. Use all of the years in the data

set to compute the average. Next, calculate the difference between each individual year and the average. Set this up so you end up with positive values for the wetter years and negative values for the drier years.

(a) What is the average total annual precipitation in each of the locations? Give your answer to two decimal places.

(b) Was the decade preceding the Dust Bowl Era wetter than average for each location? There is more than one way to answer this question, please explain your method.

(c) Compare precipitation in the 1940s with the 1930s and 1920s at Boise City and Liberal. Which decade do the 1940s more resemble?

(d) The 1940s comparison is difficult for Boise City due to missing data in several of the months. How might that limitation be addressed?

(e) Is every Dust Bowl year anomalously dry in the Great Plains locations? Does that mean the drought ended in those years?

3. The data can also be used to examine how temperature has changed over time at the different locations. Make a mean annual temperature anomaly plot for the three towns, Boise City, Liberal, and McMinnville following the same method you used to make the precipitation anomaly plots. This time, use the standard baseline period of 1951 to 1980 to compute the mean against which the individual years may be compared. If you create the equation to calculate the anomaly in one cell and then copy it to others, be sure that you anchor the years you are using to compute the baseline mean. This can be done by adding $ signs to the equation, for example from my Boise City calculation, =R3-AVERAGE($R$46:$R$75) .

   The comparison may be easier if you plot all three time series on the same set of axes than if you make three different plots.

   (a) Compare the 1920s and 1930s temperature anomalies in the three cities. Were the 1930s anomalously warm compared to the 1920s in all three locations?

   (b) Compare the temperature anomalies since the 1950s in Liberal, KS and McMinnville, OR. *Do you see the same trend in the two locations?*

   (c) If you had only the Liberal, KS temperature data available to you, would you conclude that the Earth is experiencing a warming trend?

4. Comparing the three station temperature records supports an idea we have discussed in lecture, that not all places experience the same trends at the same time. The differences arise due to the coupled interactions of all the processes that define climate. The differences at single sites also suggest why these are not the best data for examining a global phenomenon such as global warming. For that, we need data averaged among many locations. One such data set is provided for you in the Excel workbook, a global temperature anomaly time series computed by climatologists at NASA Goddard Institute for Space Studies. The baseline period used to calculate the anomaly is the standard 1951 to 1980.

   (a) Plot the global temperature anomaly time series and compare it to the time series for the individual stations. Compare the global time series with the series for Boise City, OK. Are the trends the same?

(b) A meaningful answer to the second part of the last question requires statistical analysis. Fortunately, the test we need to perform is relatively simple. Begin by selecting the data set in a plot you have created and using Excel's **Add Trendline** tool to fit a linear trend line to the data. Excel does this using a least squares method. The options will allow you to add the equation for the line to your chart. What are the slopes of the lines for the global temperature anomaly and the Boise City, OK temperature anomaly?

(c) Are the temperature trends you computed for the global anomaly and for Boise City statistically significant? To answer this question you will need to write down the slope reported by Exel's **Add Trendline** and use it to complete the calculations described in section (3.2) of this handout.

(d) Would you answer to the last question have been different if you selected a different range of years (instead of all the years available)? Explain your answer.

5 references

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010, Global surface temperature change, *Reviews of Geophysics*, 48, RG4004, doi:10.1029/2010RG000345.

North, G.R. et al, 2006, Surface Temperature Reconstructions for the Last 2,000 Years, G.R. North, chair: Committee on Surface Temperature Reconstructions for the Last 2,000 Years, National Research Council, The National Academies Press, Washington, DC.