ATOC5860 - Homework 2 - due Thursday February 10, 2022

Please send your homework to Jen/Prof. Kay on Slack as a direct message.

Please Name Your Homework Files: "ATOC5860_HW2_LastName.pdf, .html, .ipynb"

Your submissions should include: 1) A .pdf document with responses to the questions below, 2) Your code in both .ipynb and .html format.

Show all work including the equations used (e.g., by referring to the Barnes Notes).
Write in complete, clear, and concise sentences.
Eliminate spelling/grammar mistakes.
Label all graph axes. Include units.

PICK EITHER PROBLEM #1 or PROBLEM #2 TO COMPLETE. ***ONLY DO ONE.***

1) In this problem, you will assess the influence of autocorrelation on basic statistics. Specifically, you will use Monte Carlo techniques to explore how autocorrelation in an AR1 time series influences estimates of the sample mean and sample standard deviation. (50 points total)

Generate four standardized red noise time series of length 5000. ***Remember: "standardized" data have a mean of zero and a standard deviation of 1.*** **Time series one (T_1) should have a lag-1 autocorrelation $\rho(1)=0$; T_2 should have a lag-1 autocorrelation $\rho(1)=0.25$; T_3 should have a lag-1 autocorrelation $\rho(1)=0.5$, and T_4 should have a lag-1 autocorrelation $\rho(1)=0.92$.** ***For reference: The time series of yearly surface temperature anomalies of over the southwest United States has $\rho(1)=0.25$. The monthly Cold Tongue index (a measure of ENSO) has $\rho(1)=0.92$.***

a) Plot time series for T_1, T_2, T_3, and T_4. (10 points)

b) For each time series – Draw a random sample consisting of N=100 consecutive elements from the full time series. Note that you must use consecutive values since you are analyzing the persistence. Calculate the mean and standard deviation of the sample, and save your value. Repeat Monte Carlo style until you have a good distribution (i.e. do it 10,000 times). Look at the influence of persistence on your estimate of the mean and the standard deviation based on samples of size N=100. Plot histograms of your estimated sample means and standard deviations for T1, T2, T3, and T4. (10 points)

c) Estimate how the number independent samples (N*) you have for T1, T2, T3, and T4 using Barnes Chapter 2 Equation (88). (15 points)

d) Discuss your results in 1-2 paragraphs. Since you created red-noise time series, the population means are by definition equal to zero ($\mu=0$) and the population standard deviations are by definition equal to one ($\sigma=1$). How does increasing the redness of a time series affect sample estimates of the mean and standard deviation? What are the broader implications for data analysis of time series? (15 points)

2) In this problem you will assess the relationship between two variables: Global mean temperature and ENSO. Use the two provided datasets that run from 1881 through 2017: 1) Global surface temperature anomalies from the GISTEMP dataset, 2) the previous December's ENSO Nino 3.4 Index. Data are in a comma-delimited text file available on the class google drive: homework2_problem2_data.csv. (50 points)

If at a later date you are looking at this assignment and you want to update this analysis to include more years – the original data are here:
GISTEMP data here: https://data.giss.nasa.gov/gistemp/
NINO data here: https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Nino34/

a) First, detrend your data. Next, standardize your data by subtracting the mean and dividing the standard deviation. *Note: Your data should both have a mean of 0 and a standard deviation of 1. After this, your data should both have a mean of 0 and a standard deviation of 1. Check that this is true before you proceed.* Then -- look at your data by making two plots. Plot #1: a time-series plot of standardized detrended global-mean temperature and standardized detrended ENSO index. Plot #2 scatter plot for the standardized detrended values of global-mean temperature (y-axis) versus the standardized detrended ENSO index (x-axis). (10 points)

b) Calculate the correlation and regression coefficients between the standardized detrended global mean temperature and the standardized detrended ENSO index. The regression slope should be in units of standard deviation of global mean temperature per standard deviation of ENSO. Calculate the fraction of the variance in global mean temperature explained by ENSO variability. Is the correlation between these two variables statistically significant (i.e., different than 0) at the 95% confidence level? Assume 1 degree of freedom per year. (15 points)

c) Using composite analysis, calculate the standardized detrended global mean temperature for years when the ENSO 3.4 anomaly exceeds +1 and -1 degree Celsius, respectively. Assess if the composite means are different at the 95% confidence level. *Hint: see example 1.4.2.2 comparison of two sample means problem in the Barnes lecture notes Chapter 1.* (10 points)

d) Discuss your results in 1-2 paragraphs. Discuss the similarities and differences between the results from the composite analysis and the regression/correlation analysis. Explain why they don't provide identical results. (15 points)

3) Select two variables (variable A, variable B) you are using in your research that you think might be related. Use what we have been learning in ATOC7500 to analyze them and assess if one can be used to predict the other using linear regression techniques. In other words – can you use variable X to predict variable Y using linear regression? (50 points)

a) **Plot your raw data as a line plot (i.e., as a function of the sampling direction, commonly time) and as a scatter plot. Explain why you decided which variable would be X and which would be Y. Explain what you see visually and if this gives you hope or not that you might be able to predict Y using X. (5 points)**
b) **Standardize your data: subtract the mean and divide by the standard deviation. In other words, ensure that both variables have a mean of 0 and a standard deviation of 1. Also de-trend if needed. Plot your standardized data as an x-y plot and as a scatter plot. (5 points)**
c) **Calculate the autocorrelation and estimate the number of independent samples for both of your variables. (10 points)**
d) **Complete a regression between two variables. Assess the statistical significance taking into account the number of independent samples. Place confidence intervals on your regression coefficient. Provide any other statistical results that are helping you interpret your data. Provide a plot or two showing your results. (10 points)**
e) **Apply Granger Causality. Can X help you predict Y more than Y itself? (10 points)**
f) **Discuss what have you learned. Can you use X to predict Y using linear regression? Are your results statistically significant? If Yes – can you explain the underlying physical reasons for a potential correlation between the two variables? If No – why? What are the next steps if you wanted to take this analysis further? Please explain in a paragraph or two. (10 points)**

4) Future homework assignments will continue to require that you analyze your own data. Identify a dataset you wish to use and describe it here briefly. The dataset should have at least two dimensions (e.g., time&space, size distribution&mass). You will apply EOF/PCA analysis to this dataset. Please discuss with the professor if you do not have a dataset in mind. (0 points)