METR 5433 - Advanced Statistical Meteorology
Spring 2020 – Homework #2
<u>**Due:**</u> 27 February 2020

(#1) For this problem, you will again use **SfcTAnomalies-1950-Present.nc** along with **ENSO.txt**.

(a) To start, find the *best-fit* between ENSO and global-mean temperatures; i.e., is the best fit the contemporaneous relationship or a lagged relationship? To do so, plot the lag correlation coefficient between ENSO and global-mean temperatures for all lags between -12 months to +12 months. Here, negative lags mean that ENSO *leads* global-mean temperature and positive lags mean that ENSO *lags* global-mean temperature. To calculate lag correlations, remember that you have to truncate the time series for computations. At what lag does the "best" (highest) correlation between ENSO and global-mean temperature exist? Show a plot to justify your choice, including what the correlation coefficient for 90% significance is.

(b) Using (a), calculate the regression coefficient between global-mean temperature (i.e., used a weighted mean, with the weights being cosine latitude) and the ENSO index. The units will be K/std. Make sure to use the right lag and apply it correctly! Then, calculate the *ENSO-residual* global-mean temperature time series; i.e., the global-mean temperature minus the fitted (and lagged, if necessary) ENSO time series: $\mathbf{y}_{\text{residual}} = \mathbf{y} - \mathbf{y}_{\text{ENSO-fit}}$. Make a plot of all three time series as three separate subplots in the same figure: (1) global-mean temperature; (2) the fitted (and lagged, if needed) time series with ENSO; and (3) the residual time series. Stack the plots on top of each other in the order depicted.

(c) Now, determine whether ENSO has impacted recent global climate change. To do so, calculate the linear trends in the raw, ENSO-fitted, and residual global-mean time series. Express the trends in K/decade. In your write-up (2-3 paragraphs), address the following: (i) How you performed the analysis; (ii) The lag (if any) between ENSO and global-mean temperature (and if so, why might this exist?); (iii) What fraction of month-to-month variability in global-mean temperature is linearly related to ENSO?; (iv) Whether you think ENSO has contributed to global climate change.

(#2) Let's investigate some more about red noise time series and red noise processes.

(a) Generate three standardized <u>red noise</u> time series of length $N = 5000$. Time series #1 (T_1) will have a lag-one autocorrelation of $r_1 = 0.1$, T_2 a lag-one autocorrelation of $r_1 = 0.6$, and T_3 a lag-one autocorrelation of $r_1 = 0.9$. Plot each time series in a three-panel plot.

(b) Next, assume the mean (=0) and standard deviation (=1) of the $N = 5000$ time series represent the population values of these statistics. Determine the effect of persistence on the estimate of these statistics based on samples of size $n = 100$. Here is one way to do this: For each time series T_1, T_2, and T_3:

1. Draw a random sample consisting of 100 consecutive elements from the full time series. Calculate the mean and standard deviation of the sample, and save that value.

2. Repeat Step 1 10,000 times.

3. Plot the histogram of the sampling distribution of the means and standard deviations found in Steps #1 and #2, above.

Discuss your results (about 2-3 paragraphs). In the discussion, address the following: (1) Assuming the sample size is fixed, what is the impact of persistence on the distribution of sample means?; (2) How does persistence impact the uncertainty of the sample mean?; (3) What is the impact of persistence on the estimate of the population standard deviation?; (4) How will ignoring persistence in a pair of time series impact the significance of the correlation between those time series?

(#3) Let's build a simple seasonal forecast model for sea surface temperature (SST) anomalies (SSTa) using two predictors: the February-April (FMA)-averaged *North Pacific Meridional Mode* (NPMM) and the FMA-averaged *South Pacific Meridional Mode* (SPMM)[1]. Specifically, you will be testing the forecasting skill of using the two meridional mode indices during FMA to predict the following November - January (NDJ(+1)) average Pacific-basin SSTa field. That is, for example, you would use FMA-average 1950 values of the NPMM and the SPMM for forecasting the NDJ-average 1950-51 SSTa field. The files **NPMM.txt** and **SPMM.txt** contain the monthly-mean values (in standardized units) of the NPMM and SPMM, respectively. You will also use the **SSTs-1950-Present.nc** file for SSTs (Note: Missing values in the file mean land points). Think of this forecasting model as a *multi-linear regression problem*. Also, make all plots for the region 100°E - 60°W, 65°S - 65°N.

(a) First, calculate and plot the correlation coefficients between NDJ(+1)-averaged SSTa and each of the FMA-averaged indices *separately* as a function of space (i.e., two different plots/maps). You will start with FMA 1950 for this calculation and go through NDJ(+1) 2017-18. To calculate SSTa, remove the seasonal cycle from all grid points and use the 1981-2010 base period for climatology. Remember to detrend all data before computing the correlation (explain why in your writeup). Indicate on your plots where the correlation coefficients are significant at the $p < 0.05$ level using a Monte Carlo test.

(b) Now calculate the correlation between the two FMA-averaged meridional mode indices. Are these "independent-enough" predictors for SSTa? Briefly explain your reasoning (HINT: Consider the r^2 values in addition to r).

(c) Regardless of your answer to (b), build a functional form of the general least-squares model: $\mathbf{y} = \mathbf{E}\mathbf{x} + \mathbf{n}$. Sketch / describe what each vector / matrix $\mathbf{y}, \mathbf{E}$, and $\mathbf{x}$ represents (e.g., what will be in the columns of $\mathbf{E}$?).

(d) In order to test the forecast *skill* of the model, you will need to run the model multiple times with the "leave-one-out" approach. The process goes as follows. In the first iteration, you will remove (leave out) values of the indices for 1950 as well as the NDJ 1950/51 SSTa from the matrices constructed in (c). Then, solve for $\hat{\mathbf{x}}_{\mathrm{no}-1950}$ with this reduced set of equations. Next, use $\hat{\mathbf{x}}_{\mathrm{no}-1950}$ and the 1950 values of the two indices to forecast the NDJ(+1) 1950/51 SSTa (i.e., $\hat{\mathbf{y}}_{1950/51} = \mathbf{E}_{1950}\hat{\mathbf{x}}_{\mathrm{no}-1950}$). Then, repeat this process for 1951, 1952, etc. Through running the"leave-one-out" process, you will end up with a [lon x lat x N] array, corresponding to the 1950/51 - 2017/18 (i.e., N maps) of the forecasted NDJ(+1) SSTa. Think of this verification method as setting up a loop to run the least-squares fitting multiple times and retain the forecast values for every year.

After running the "leave-one-out" method, check the skill of the model by calculating the *anomaly corre-*

[1] You can read up about what a meridional mode is by visiting this site. It's OK if you don't wholly understand the modes - you can still do the statistics!

lation coefficient (ACC) for each grid point:

$$ACC = \frac{\sum_{m=1}^{M} f'_m v'_m}{\sqrt{\sum_{m=1}^{M} (f'_m)^2}\sqrt{\sum_{m=1}^{M} (v'_m)^2}} = \frac{\overline{f'v'}}{\sqrt{\overline{f'^2}}\sqrt{\overline{v'^2}}}$$

where f'_m is the forecast anomaly of the variable for time m and v'_m is the observed (or verified) anomaly (i.e., data from the file). Make a plot of the ACC for NDJ(+1) SSTa for your model. Plot *only* positive ACC values. Outline areas where the ACC is significant at the $p < 0.05$ level (use a proper method to test for significance). Discuss the results (1-2 paragraphs). Where does the model have skill? Where doesn't it? Also, compare your map to Figure 12c from You and Furtado [2018]. (Note: It's OK to get different values. :)).