Problem Set 5

J. S. Wright
jswright@tsinghua.edu.cn

5.1 The provided data file includes monthly mean surface air temperature anomalies for January 1950 through December 2016 based on the most recent version of the GISTEMP surface temperature analysis.

(a) Use one of the python linear regression models (e.g., scipy.stats.linregress, statsmodels.OLS, sklearn.linear_model.LinearRegression) to calculate the linear temperature trend between January 1950 and December 2016 for every grid cell. Note that some grid cells have masked data, especially early in the record, and that a few locations have no valid data at all. Choose an approach to deal with these masked data and briefly summarize this approach in your final submission.

(b) Plot a map that shows the spatial distribution of the 1950–2016 trend in surface air temperature anomaly in units of K per decade. Make sure you choose appropriate contour levels and color scale. In particular, when plotting a variable that can be either positive or negative, it is useful to make the zero line clearly and easily identifiable. You may find the Color Brewer website helpful — matplotlib and seaborn allow you to access many of these color schemes by name.

(c) Remove the linear trend from the time series at each grid cell, so that you have a "detrended" data set. Regress these detrended data against the Multivariate ENSO Index (MEI) from problem set 3.

(d) Plot the spatial distribution of detrended surface air temperature anomalies associated with an MEI index of 1. Mark the locations where this regression is significant at the 95% level using Student's t test. Briefly describe some of the key features of this spatial distribution.

(e) **Extra credit:** Calculate the trends from part (a) using the Theil–Sen estimator, an alternative to linear regression that is more robust to outliers and less sensitive to the choice of endpoints. Implementations of this operator are provided in both scipy (scipy.stats.theilslopes) and sklearn (sklearn.linear_model.TheilSenRegressor). Compare the resulting distribution of trends to the distribution you plotted in part (b) — are there any differences worth noting?

5.2 Using the provided data files, regress precipitation anomalies for northern hemisphere winter (DJF) in each grid cell against the Niño3.4 SST anomaly.

(a) Plot a map of the precipitation anomalies your regression model predicts for the El Niño of DJF 1997–98.

(b) Plot a map of precipitation anomalies your regression model predicts for the La Niña of DJF 1998–99.

(c) Use masked arrays to mask relationships that are not significant at the 95% confidence level (see the slides and script from week 5), and replot your results for (a) and (b).

(d) How do your predicted anomalies compare with the actual anomalies observed during DJF 1997–98 and DJF 1998–99? How do they compare with NOAA's illustrations for warm and cold ENSO events?

5.3 A dynamical index for the intensity of the East Asia Winter Monsoon (EAWM) may be calculated from the isobaric distribution of potential vorticity at 500 hPa (Huang et al., 2016):

$$I_{\mathrm{PV}} = \langle PV_{500\mathrm{hPa}} \rangle_{\mathrm{EA}} - \langle PV_{500\mathrm{hPa}} \rangle_{\mathrm{ET}} \tag{1}$$

where $\langle\rangle_{\mathrm{EA}}$ is the area average over East Asia (25–55°N, 100–150°E) and $\langle\rangle_{\mathrm{ET}}$ is the average over the global extratropics (25–55°N).

(a) Using the provided data files, calculate the EAWM index using equation 1.

(b) Construct a time series of the wintertime (DJF) mean EAWM index for 1958 through 2014 (i.e., December 1958 through February 2015). Plot this time series, as well as a map showing the climatological DJF distribution of PV at 500 hPa and the boundaries of the EA box used to calculate the index.

(c) Normalize the DJF mean time series you calculated in part (b) to have a mean of zero and a standard deviation of 1. After this normalization, an index of 1 should represent a winter for which the mean EAWM index is one standard deviation above the mean. Plot this normalized time series.

(d) Using the normalized time series, calculate the linear trend in the EAWM index and evaluate its statistical significance. Plot the trend line together with the normalized time series and show the value of the coefficient of determination (R^2) on the plot.

(e) **Extra credit.** Many alternative indices for measuring the intensity of the EAWM have been proposed. One of these is the index proposed by Li and Yang (2010) based on DJF upper tropospheric zonal wind:

$$I_U = \langle U_{200\mathrm{hPa}} \rangle_{\mathrm{ML}} - \frac{1}{2}\left[\langle U_{200\mathrm{hPa}} \rangle_{\mathrm{HL}} + \langle U_{200\mathrm{hPa}} \rangle_{\mathrm{LL}}\right]$$

where $\langle\rangle_{\mathrm{ML}}$ is a mid-latitude area average over (30–35°N, 90–160°E), $\langle\rangle_{\mathrm{HL}}$ is a high latitude area average over (50–60°N, 70–170°E), and $\langle\rangle_{\mathrm{LL}}$ is a low latitude area average over (5°S–10°N, 90–160°E). Another is the index proposed by Wang and Chen (2014) based on DJF sea level pressure:

$$I_{\mathrm{SLP}} = \langle \mathrm{SLP} \rangle_{\mathrm{SH}} - \frac{1}{2}\left[\langle \mathrm{SLP} \rangle_{\mathrm{NP}} + \langle \mathrm{SLP} \rangle_{\mathrm{MC}}\right].$$

where $\langle\rangle_{\mathrm{SH}}$ is an area average over the Siberian high region (40–60°N, 70–120°E), $\langle\rangle_{\mathrm{NP}}$ is an area average over the mid-latitude North Pacific (30–50°N, 140–170°E), and $\langle\rangle_{\mathrm{MC}}$ is a low latitude area average over the Maritime Continent (20°S–10°N, 110–160°E).

Using the provided data files, calculate these two alternative indices for EAWM intensity. Put all three indices in a dataframe and calculate the correlations among them. Use one of the example applications of the seaborn PairGrid environment as a template to make a joint plot showing relationships among the three indices. Plot maps showing the climatological DJF distribution of the analyzed quantity (zonal wind or sea level pressure) and the boxes used for calculating the index.

5.4 Next we examine the relationship between the EAWM and climate conditions over East Asia.

(a) Calculate the DJF surface air temperature anomaly ($T_{\text{DJF}} - \overline{T_{\text{DJF}}}$) at every grid cell. Regress these surface air temperature anomalies onto the normalized EAWM index you calculated in Problem Set 5. This should be done by calculating linear regressions with a slope and intercept at every grid cell. The surface air temperature anomaly is the dependent variable (y) and the normalized EAWM index is the independent variable (x). See the example for details.

(b) Plot a map of DJF surface air temperature anomalies regressed onto a normalized DJF EAWM index of 1 (this map can be regional or global, but should at least include China). Don't forget the intercept! Use a two-sided Student's t test to evaluate the significance of the regression at each grid cell. Mark the points with significant relationships on the map.

(c) Surface air temperature and the EAWM may both be significantly impacted by global warming, which could confound the statistical relationship between the EAWM and surface air temperature anomalies. Remove the linear trends from (detrend) the normalized EAWM index and the surface air temperature anomaly at every grid cell. Regress these detrended surface air temperature anomalies onto a normalized EAWM index of 1, and plot the map. What differences do you find relative to part (b)?

(d) Regress the distributions of detrended DJF-mean sea level pressure anomalies and 10-m near surface wind anomalies onto the detrended normalized EAWM index. Plot a map of these quantities regressed onto a normalized EAWM index of 1. For plotting the winds, you might try the streamplot or quiver functions provided by matplotlib.

(e) **Extra credit:** The decorrelation timescale for the correlation (or regression) between two variables $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ can be estimated by

$$\tau = 1 + 2\sum_{k=1}^{n-1} \rho_{\mathcal{X}}(k)\rho_{\mathcal{Y}}(k)$$

where $\rho_{\mathcal{X}}(k)$ is the autocorrelation of $\mathcal{X}(t)$ at lag k and $\rho_{\mathcal{Y}}(k)$ is the autocorrelation of $\mathcal{Y}(t)$ at lag k. Estimate the decorrelation timescale for the relationship between area-mean DJF temperature and the normalized EAWM index. Can we assume that each year is statistically independent when evaluating the significance of the correlation calculated in part (a)?

(f) **Extra credit:** The decorrelation timescale is an important factor in estimating the effective sample size n', which can then be used to adjust the parameters used in statistical testing (replacing the sample size n, which assumes every sample is independent). For a regression or linear correlation,

$$n' = \frac{n}{1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_{\mathcal{X}}(k)\rho_{\mathcal{Y}}(k)}$$

where n is the size of the sample, $\rho_{\mathcal{X}}(k)$ is the autocorrelation of $\mathcal{X}(t)$ at lag k and $\rho_{\mathcal{Y}}(k)$ is the autocorrelation of $\mathcal{Y}(t)$ at lag k. Write a function to calculate the effective sample size, and use it to re-evaluate the significance of the regressions you calculated in part (c). You may need to do some additional research on the role of n in the two-sided Student's t test. See also `scipy.stats.t`.