Data:

- Gridded $[5^\circ \times 5^\circ]$ monthly surface temperature anomalies for all months of January 1958 to December 2008.
- Time series of ENSO index for all months of January 1958 to December 2008.

Note: The complete code to produce these results is contained in the IDL file kcw_hw2.pro.

Problem 1: Calculate and plot the global mean temperature anomaly.

Figure 1: Time series of monthly global mean temperature anomaly for all months of Jan. 1958 to Dec. 2008.

Figure 1 shows my result for this time series. Now for some details on how I did the calculation.

The global temperature data file is arranged somewhat inconveniently, i.e., with all the temperatures (all lats and lons) in a single row for each month. To make this and all other calculations much easier, the first thing I did was reformat the global temperature grid from its native $[nLon \cdot nLat, nTime]$ format to a more intuitive 3-D form, i.e., like this:

$[nLon, nLat, nTime]$  

To calculate the global mean temperature for each month, I first calculated the mean temperature at each latitude for that month, i.e., take the simple mean of the temperatures of all longitudes at a given latitude:

$$T_{lat} = \frac{1}{nLon} \sum_{i=1}^{nLon} T_{lon,i}$$
Note that this does not require any special weighting because all values are at the same latitude. However, I had to disregard any ‘NaN’ values when computing this mean.

Once we have the mean at each latitude, take the cosine weighted mean of those, giving the mean global temperature for this month:

\[
T_{\text{month}} = \frac{\sum_{j=1}^{n_{\text{Lat}}} \cos(lat_j)T_{\text{lat},j}}{\sum_{j=1}^{n_{\text{Lat}}} \cos(lat_j)}
\]

However, when I did this weighted mean, I also accounted for the possibility that there could have been entire latitudes that had no data. Those latitudes should not be included either in the \(T\) for that latitude nor should they be included in the weighting. Simply setting the temperatures at those missing latitudes to zero, then computing the cosine-weighted mean, would be incorrect.

From Figure 1, it is clear that there is quite a bit of month-to-month variation, but there is also clearly a positive trend.

**Problem 2: Investigate the relationship between global mean temperature and ENSO.**

I first restricted both the global mean temperature time series (from problem 1) and the ENSO index time series to only the months of Dec, Jan, and Feb. After this restriction, I standardized the ENSO time series:

\[
z = \frac{\text{ENSO} - \bar{\text{ENSO}}}{\sigma_{\text{ENSO}}} \tag{1}
\]

By definition, this standardized ENSO time series now has a mean of zero and a standard deviation (and variance) of one. This gives us a more useful measure of the behavior of ENSO during only the winter months. For example, we have now essentially defined what a positive (1σ) ENSO event is for any given winter. I did not standardize the temperature time series though because we want our regression coefficient to be in units of temperature (degrees C) per unit standard deviation of ENSO index.

Figure 2 shows the resulting time series. Just eyeballing these two time series suggests that there is some relationship between temperature and ENSO. The regression quantifies this relationship. Figure 3 shows the scatter plot of the two time series along with the “best fit” line and fit parameters. For completeness, here is what the regression does:

Let \(y\) = temperature, and \(z\) = standardized ENSO.

Linear regression:

\[
\hat{y} = a_0 + a_1 z
\]

Regression coefficient, covariance in \((z,y)\) per unit variance of \(z\):

\[
a_1 = \frac{z'y'}{z'^2} \tag{3}
\]

Correlation coefficient, similar to covariance but normalized by the variance of both \(z\) and \(y\):

\[
r = a_1 \cdot \frac{\sqrt{z'^2}}{\sqrt{y'^2}} \tag{4}
\]

But since the ENSO index is standardized, the regression and correlation coefficients simplify to:

\[
a_1 = \frac{z'y'}{y'^2}
\]

\[
r = \frac{a_1}{y'^2}
\]
Figure 2: Time series of (black) global mean temperature anomaly and (red) standardized ENSO index for winter months of Jan. 1958 to Dec. 2008. The ENSO index is plotted against the right axis.

The units of $a_1$ are degrees C per unit standard deviation of ENSO index. The fraction of the variance in temperature that is “explained” by ENSO is simply $r^2$.

This “best fit” was accomplished by simply doing this (in IDL):

```idl
fit = poly_fit(z, y, 1)
a0 = fit[0]
a1 = fit[1]
r = a1/STDDEV(y)
yhat = z*a1 + a0
```

Despite the eyeballed impression of a good correlation between ENSO and global mean temperature that we get from the time series in Figure 2, the correlation between the two is not very large. It is only about $r \approx 0.21$ which means that only about 4.5% of the variance in global temperature is “explained” by ENSO. (Although, as we’ll see in the next homework, this correlation is a bit different if we lag the temperature relative to ENSO).

The regression coefficient ($a_1$) tells us that we would expect about a 0.05°C change in global temperature per unit standard deviation change in ENSO index. It was useful to standardize ENSO first for a couple reasons: (1) it makes the calculations easier, (2) it gives a more useful result. The correlation coefficient does not depend on the amplitude of ENSO (regardless of whether or not ENSO is standardized), but the regression coefficient does depend on that amplitude. A “raw” (not standardized) ENSO index isn’t really very meaningful (what are the units, for example?).
Figure 3: Scatterplot of global mean temperature anomaly versus standardized ENSO index for winter months of Jan. 1958 to Dec. 2008. The line shows the “best-fit” linear regression between the two. The numbers in the bottom-right list the number of samples (N) in the time series, the regression coefficient (a1), the correlation coefficient (r) and the square of the correlation coefficient ($r^2$).

Whereas the standardized ENSO index gives us a better feel for what a positive or negative value means. For example, $z = 1$ means that ENSO is one standard deviation above its long-term winter mean.

**Assessing Significance**

To assess the significance of the correlation between temperature and ENSO index, I used the Student’s t-statistic. We get our sample value for the t-statistic via:

$$t = \frac{r(N - 2)^{\frac{1}{2}}}{(1 - r^2)^{\frac{1}{2}}}$$  \hspace{1cm} (5)

where $r = .209$ is the sample correlation we just computed and $N$ is the number of independent samples. $N - 2$ is the degrees of freedom.

The Null Hypothesis we are testing here is that the true correlation ($\rho$) between ENSO and temperature is zero. The t-distribution is the probability distribution of $\rho$ about this zero mean (the t-distribution is almost identical to the normal distribution in this case due to the large number of degrees of freedom). There is some finite probability that our sample correlation ($r$) would be different from zero even though the true correlation is zero. So, basically, we are using this t-distribution to quantify just how confident we can be that a non-zero sample correlation is not due to chance. We will strictly define rejection of the null hypothesis as “the correlation is not due to chance with 95% confidence”. So this means that our value for $t$ in the above equation must exceed the **two-sided** value for 95% confidence from the
t-distribution. I’ll call this the confidence value. I’m using the two-sided confidence value because I have no a priori reason to expect either a positive or negative correlation. So we will consider our r to be significant if it lies beyond 95% of the t-distribution on either side. In short, this amounts to using the 2.5% value (0.025) for our confidence value.

Now then, how many degrees of freedom (ν = N − 2) do we have? If we think that all the months in our time series are independent, then ν = 153 − 2 = 151. In that case our confidence value is:

\[ t_{0.025} \approx 1.976 \]

It is more likely that all months are not independent. It would be more appropriate to consider only that each winter is independent, so ν = 51 − 2 = 49. In that case our confidence value is:

\[ t_{0.025} \approx 2.009 \]

The change of ν didn’t affect our confidence value much. It’s about 2 in both cases, i.e., like a 2σ of a normal distribution. However, this change of ν does affect the t-stat we calculate using Equation (5) above.

- ν = 151, our t-stat is \( \approx 2.621 \)
- ν = 49, our t-stat is \( \approx 1.493 \)

So we could reject the Null Hypothesis if we treated all months as independent, but we could not reject it if we treated only all winters as independent. (As a side note: when I did this calculation about 8 years ago, I found a much higher correlation, and I also found that I could reject the Null Hypothesis regardless of whether I treated all months or just all winters as independent. I guess the last 8 years of temperature and ENSO and/or “refinements” to the data record changed something.)

### Composite Analysis

Composite analysis is an alternate (or perhaps supplemental) way to assess the relationship between ENSO and global temperature. The time series and scatterplots of Figures 2 and 3 show a tendency for winter temperature to be warmer during positive ENSO phases and colder during negative ENSO phases. Hence, a reasonable thing to do is to grab sub-sets of the temperature data from each phase. Specifically, I will define a warm/cold phase as being when our standardized ENSO index > 1 and < −1. So we have our two composite categories:

1. Global mean temperature anomalies in winter months when ENSO > 1.
2. Global mean temperature anomalies in winter months when ENSO < −1.

Are the statistics of temperature in these two categories significantly different from the total population (i.e., all winter months)? Are they different from each other?

Let’s start with the statistics of the total population. Using all 153 winter months, I get the following sample mean (\( \bar{T} \)) and sample standard deviation (\( s_T \)) for mean global temperature anomalies:

\[
\begin{align*}
\bar{T} &= 0.103°C \\
{s_T} &= 0.253°C
\end{align*}
\]

I’ll use the Student’s t-statistic to place some confidence intervals around the sample mean. The t-statistic in this case is of the form:

\[
t = \frac{\bar{T} - \mu}{s_T} \cdot \sqrt{N - 1}
\]

(6)

where \( \mu \) is the “true” population mean, and \( N \) is the number of independent samples. Assuming that we have only independent winters, \( N = 51 \). For 95% confidence intervals, our t-stat should lie in the range:

\[
t_{0.025} < \frac{\bar{T} - \mu}{s_T} \cdot \sqrt{N - 1} < t_{0.025}
\]

(7)

So, “invert” this to find the confidence intervals for our \( \mu \):

\[
\left( \bar{T} - t_{0.025} \cdot \frac{s_T}{\sqrt{N - 1}} \right) < \mu < \left( \bar{T} + t_{0.025} \cdot \frac{s_T}{\sqrt{N - 1}} \right)
\]

(8)
or more simply:
\[ \mu = \bar{T} \pm t_{0.025} \cdot \frac{s_T}{\sqrt{N-1}} = \bar{T} \pm \delta T \] (9)

Using \( N - 1 = 50 \), our t-stat is \( t_{0.025} \approx 2.01 \), our \( \delta T = 0.072 \). So our confidence intervals on the population mean are:

\[ 0.031^\circ C < \mu < 0.175^\circ C \]

Now, what do we get from our two composite categories? I computed the mean, standard deviations and confidence intervals of temperature for the two composite categories. The table below summarizes the results:

<table>
<thead>
<tr>
<th>Sub-set</th>
<th>N</th>
<th>( T )</th>
<th>( s_T )</th>
<th>( \delta T )</th>
<th>( \bar{T} - \delta T )</th>
<th>( \bar{T} + \delta T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>51</td>
<td>0.104</td>
<td>0.25</td>
<td>0.072</td>
<td>0.031</td>
<td>0.175</td>
</tr>
<tr>
<td>ENSO &gt; +1</td>
<td>23</td>
<td>0.227</td>
<td>0.209</td>
<td>0.093</td>
<td>0.135</td>
<td>0.320</td>
</tr>
<tr>
<td>ENSO &lt; -1</td>
<td>19</td>
<td>0.098</td>
<td>0.285</td>
<td>0.141</td>
<td>-0.043</td>
<td>0.239</td>
</tr>
<tr>
<td>ENSO &gt; +1</td>
<td>13</td>
<td>0.227</td>
<td>0.209</td>
<td>0.132</td>
<td>0.095</td>
<td>0.356</td>
</tr>
<tr>
<td>ENSO &lt; -1</td>
<td>10</td>
<td>0.098</td>
<td>0.285</td>
<td>0.215</td>
<td>-0.117</td>
<td>0.313</td>
</tr>
</tbody>
</table>

The first line gives the results for all winter months, with only 51 independent winters. The next set of two lines gives the results for the composites, assuming all months were independent. So, for example, I found 23 months when ENSO > 1 and 19 months when ENSO < -1. However, these months spanned only 13 and 10 different winters, respectively. The last set of two lines gives the results for composites assuming only independent winters.

The mean temperature for the ENSO > 1 subset is 0.227\(^\circ C\) which is beyond the range of our estimate for the population mean, i.e., it is > 0.175\(^\circ C\). The mean temperature for the ENSO < -1 subset is 0.098\(^\circ C\) which is within the range of our estimate of the population mean. So far, our results suggest ENSO’s effect on temperature is significantly more pronounced during positive ENSO phases. Note, however, that there is overlap in the confidence ranges (the \( \bar{T} \pm \delta T \) columns of the table). That is, the > 1 ranges overlap the < -1 ranges, and both overlap with the population ranges.

Finally, I tested the Null Hypothesis that the > 1 and < -1 sub-sets have the same mean (i.e., they are from the same distribution). The t-statistic in this case has the form:

\[ t = \frac{\bar{T}_1 - \bar{T}_2}{s \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \] (10)

where

\[ s = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \] (11)

The \( T, N, s \) are the mean, number of samples and sample standard deviations of each composite category, i.e., the \( N \) and \( s_T \) values from the table above. The quantity \( \nu = N_1 + N_2 - 2 \) is the number of degrees of freedom. Assuming all months are independent, I get \( \nu = 40 \) and \( t \approx 1.65 \). Assuming only winters are independent, I get \( \nu = 21 \) and \( t \approx 1.20 \). Both of these t-stat values are much less than the two-tailed Student’s t-statistic for 95% confidence. That value is approximately 2. So, we cannot reject the Null Hypothesis.

In the end, this problem has shown us that there is certainly some sort of positive correlation between ENSO and global mean temperature. There seems to be a greater temperature response to ENSO during the warm phase (i.e., when the standardized ENSO index > 1). However, the relationship does not appear to be very significant, nor does it explain much of the variance in temperature.
Problem 3: Investigate the relationship between surface temperature and ENSO at all spatial grid points.

Figure 4: Map showing the number of useable winter months at each grid point.

Here we basically apply the same methodology as Problem 2, i.e., regress a time series of temperature against a time series of ENSO index. However, in this problem we are grabbing a time series of temperature at each grid point, calculating regression and correlation with ENSO, then summarizing the results by plotting these coefficients on a map.

As in Problem 2, I first restricted both ENSO and global mean temperature to the winter months. I then simply looped over the lat, lon grid to extract a time series at each point which I then regressed against ENSO. This results in a grid of $r$ and $a_1$ values. For example:

\[
a_{1,i,j} = \frac{zT_{i,j}}{s_{i,j}}
\]

\[
r_{i,j} = \frac{a_{1,i,j}}{s_{i,j}}
\]

where $z$ is the standardized ENSO, $T_{i,j}$ is temperature at each grid point, and $s_{i,j}$ is the standard deviation of temperature at each grid point. One detail here: not all the grid points will have the same number of useable months in their time series. To account for that, I simply disregarded any 'NaN' values. That is, at each grid point, I removed months that had 'NaN' in the $T$. I removed these months from both $T$ and ENSO, then standardized ENSO using only the "good" months. So, theoretically, I might use a different $z$ in the calculation at each grid point. I'm not entirely convinced that this was the best way to do it: do I use the same $z$ for all grid points, then just remove values from it where the $T$ has 'NaN'? Or do I calculate a different $z$ for each grid point? I couldn't decide, so I did the latter.

While computing my grid of $r$ and $a_1$, I also kept track of how many "good" samples were used in the calculation at each grid point. Figure 4 shows how many useable months were at each grid point. Most of the domain has at least 100 useable samples. The polar regions are obvious exceptions to this. I didn't do the regression calculation at a grid point unless there were at least 10 samples. Initially, I planned to account for the variability of $N$ so that I could use the appropriate number for $N$ when assessing significance. However, in the end, this didn't affect things very much, so I used the same value for $N$ at all points. I used $N = 51$, i.e., all winters (not all months) were independent.
Figure 5 shows the resulting maps of regression and correlation coefficients. The results in the Pacific are reassuring: positive ENSO phase (i.e., an El Nino) is clearly related to significantly warmer temperatures in the eastern Pacific and colder temperatures in the west Pacific. So the strongest ENSO signal is clearly in the Pacific. No surprise there since that’s the region where the ENSO index measurements are done. There doesn’t appear to be much of a significant relationship to ENSO in the interior of the continents (with the interesting exception of southern Africa), but there are some interesting “teleconnection” features in the ocean basins and along the coasts.

Wouldn’t it be fun to now re-do this analysis using precipitation instead of temperature?

Problem 4: Estimate the incidence of extreme cold temperature events.

Figure 6 illustrates the concept of this problem. The three thin lines in the figure are just the standard normal distribution $f(z)$:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

(12)

the black is f(z) for a mean of zero, the red and blue are shifted right and left by 0.2, respectively. The thick dashed line is the cumulative distribution function:

$$F(z) = \int_{-\infty}^{z} f(t)dt$$

(13)

Assuming the winter temperatures in Lubbock are normally distributed, the black $f(z)$ curve in Figure 6 is essentially the population distribution of Lubbock temperatures. In this context, the probability of having an extreme cold (-2$\sigma$) event is simply $F(-2) = 0.02275$.

Now, the question is: how would a correlation between temperature and ENSO alter this probability? Specifically, how much more probable is an extreme cold event during cold versus warm ENSO phases? Well, if both temperature and ENSO are standardized (i.e., mean is zero, variance is one), then the linear regression between the two would give us a correlation coefficient of:

$$r = a_1 = ENSO' \cdot T'$$

and thus our “best-fit” line would be of the form:

$$\hat{T} = r \cdot ENSO' = (ENSO' \cdot T') ENSO'$$

So, a change of ENSO by some amount $\delta$ will correspond to a temperature change of $r\delta$. When ENSO is exactly $\pm 1\sigma$ from its long term mean, that would correspond to a standardized temperature change of $\pm r$. Hence, the mean of our standardized temperature distribution would shift by $\pm r$, as illustrated in Figure 6.

Naturally, if the correlation between temperature and ENSO is positive, then extreme cold events are more probable when $\delta < 0$ and less probable when $\delta > 0$. This will scale with $r$. That is, the greater the correlation with ENSO, the more probable cold events will be when $\delta < 0$.

To estimate the ratio $\eta$ of the probability of 2$\sigma$ extreme cold days during cold ($\delta = -1$) versus warm ($\delta = +1$) ENSO phases, we are essentially just computing the ratio of the cumulative distribution at $-2 + r$ versus $-2 - r$, i.e.,

$$\eta = \frac{Pr\{cold\}}{Pr\{warm\}} = \frac{F(-2 + r)}{F(-2 - r)}$$

So, all we have to do is solve for “$r$” given a value for $\eta$ that is 2, or 3, or 4. This is difficult to do analytically because the equation for $F(z)$ is difficult to solve analytically. Hence, we have to try something else. Here are some options:

1. Use trial and error with a table of values for $F(z)$.
2. Use a numerical method like Newton’s Method.
3. Use a graphical method.
All of these options are fine. I was hoping someone would try using Newton’s Method. For example, on page 91, Wilks gives a close approximation to the cumulative distribution $F(z)$ which has the form:

$$F(z) \approx \frac{1}{2} \left[ 1 \pm \sqrt{1 - e^{\frac{-2z^2}{\pi}}} \right]$$

So, one could use this form of $F(z)$ in Newton’s method to solve for the root, i.e., solve for $r$ such that:

$$F(-2 + r) - \eta \cdot F(-2 - r) = 0$$

Instead of using Newton’s Method, I solved this graphically. Figure 7 shows the result. To get this plot, I used IDL’s built-in function called GAUSS_PDF which gives the exact value of $F(z)$ for any $z$. So, all I did here was to compute and plot $F(-2 + r)$ and $\eta \cdot F(-2 - r)$ as a function of $r$. Plot them and see where they intersect. The values of $r$ where they intersect are the $r$ values we are looking for. I get the following:

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.146</td>
</tr>
<tr>
<td>3</td>
<td>0.231</td>
</tr>
<tr>
<td>4</td>
<td>0.292</td>
</tr>
</tbody>
</table>

Interesting things to note here:

- If $r = 0$, the probability of an extreme cold event is obviously the same in both warm and cold phases of ENSO.
- It doesn’t take too much of an increase in $r$ to make cold days 2, 3, or 4 times more probable in the cold versus warm phase of ENSO.

We could only solve the problem this way if we define opposing phases of ENSO as situations in which the ENSO index “is 1 standard deviation from its long-term mean” because that gives a specific value for $F(z)$ in our computed ratio. If we defined opposing phases as situations in which the ENSO index “exceeds its long-term mean”, then we have no specific value for $z$. 
Figure 5: Map of (left) regression and (right) correlation coefficients between surface temperature and standardized ENSO index for winter months of Jan. 1958 to Dec. 2008. Contour interval is 0.2 with zero contour omitted. Positive values are in red. Negative values are in blue. The units of the regression coefficient are °C per unit standard deviation of ENSO index. The gray-shaded areas indicate where the correlation exceeds the two-tailed 95% confidence level assuming 51 independent winters (49 degrees of freedom) at each grid point.
Figure 6: The standard normal distribution. The probability distribution functions $f(z)$ are plotted against the left axis. The cumulative probability $F(z)$ is plotted against the right axis. The thin black line is $f(z)$ for a mean of zero. The thin red and blue lines are $f(z+0.2)$ and $f(z-0.2)$ (i.e., a mean of 0.2 and -0.2 respectively). The thick, black dashed line is $F(z)$. The green line at $z=-2$ indicates the “2-sigma” extreme cold event based on the zero mean distribution.

Figure 7: Probability of a $-2\sigma$ event for 1$\sigma$ cold and warm phases of ENSO as a function of the correlation coefficient. The blue line corresponds to cold ENSO, i.e., $F(-2 + r)$. The black line is the warm ENSO, i.e., $F(-2 - r)$. The three red lines are $F(-2 - r)$ multiplied by 2, 3, and 4. The points of intersection with the blue line indicate the value of the correlation coefficient required to make a $-2\sigma$ cold event 2, 3, and 4 times more probable during the cold ENSO phase relative to the warm ENSO phase. Note that a cold event is equally probable in both warm and cold ENSO phase if $r = 0$. 

Neg/Pos Ratio = 2:1, 3:1, 4:1