

4 Exploratory Multivariate Data Analysis

The techniques in the previous two chapters are intended to examine a sample of data from a single variable independent of others. An advantage of many environmental data sets is the large number of simultaneous measurements available. We often want to relate how one or more phenomena are related to others. Besides simply measuring different quantities, we often have access to observations at different locations (both horizontally and vertically). Hence, our sample may have many dimensions: x , y , z , t , and variable, model, etc.. The number of dimensions can easily grow beyond that. For example, if we are dealing with forecasts, then the forecast lead time or perturbations of model parameterizations or initial conditions become other dimensions. Dealing with the dimensionality of environmental data sets in statistical analyses is of general concern (see Murphy 1991; *Mon. Wea. Rev.*, 1590-1601). Obviously, we can slice such data sets up in a number of different ways to simplify the dimensionality of the problem depending on the goals of the study. Exploratory multivariate data analysis encompasses an array of tools to assess relationships between two or more samples.

a. Linear Regression Between Two Variates

The matlab program **chapter_4_1.m** contains all of the code for this subsection. We'll use a data set of daily precipitation collected at high elevation (SNOTEL) sites in the Wasatch Mountains. To keep the analysis manageable, only the time series of precipitation at the 7 stations labeled in Fig. 4.1 will be used and the data are preprocessed in the code to consider only the October-May seasonal totals or totals during January only.

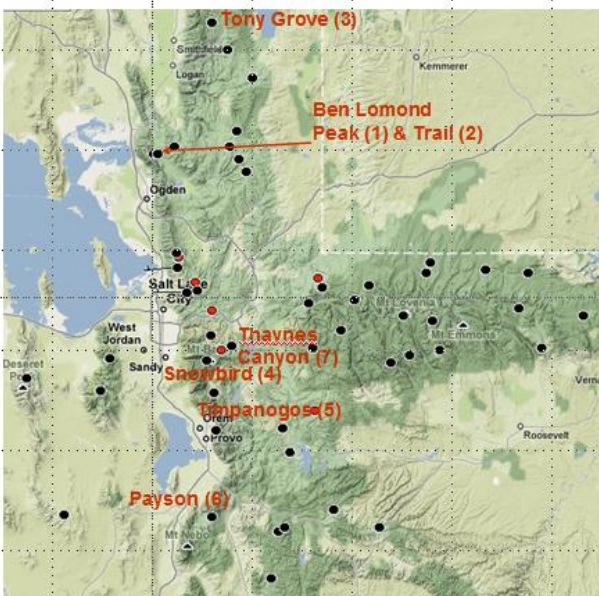


Figure 4.1. Locations of the 7 SNOTEL sites examined

precipitation at Ben Lomond Peak from that at Ben Lomond Trail we should be able to do well, but we're not going to have an easy time estimating Tony Grove precipitation from Payson's during January. You should also recognize that the degrees of freedom in these records is fewer than the 18 years in the sample.

Figure 4.2 shows the time series of seasonal total precipitation at Ben Lomond Peak and Ben Lomond Trail over an eighteen year period. Since the stations are very close to one another, it is not surprising that the year-to-year variations in precipitation at the two sites are very similar. However, since Ben Lomond Trail is at a lower elevation, then its precipitation is distinctly less than that at Ben Lomond Peak. Figure 4.2 also shows similar time series but for January total precipitation at Payson and Tony Grove. Obviously, the relationship between these two records is weaker. The degree of similarity within the two pairs of time series is easier to evaluate after transforming the data into standardized anomalies (bottom panels of Fig. 4.2). You might expect that if we try to estimate the

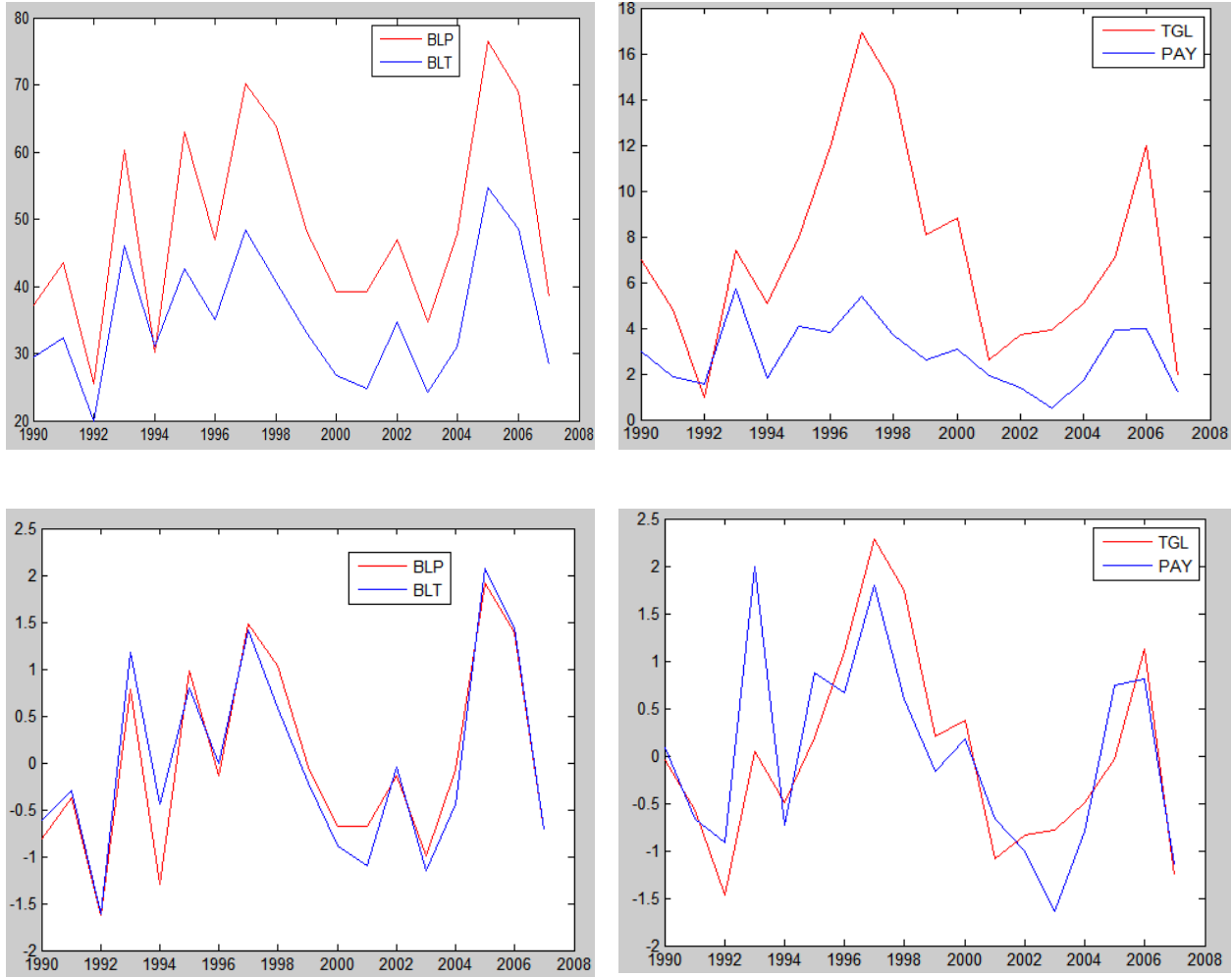


Figure 4.2. Time series (in) of seasonal total precipitation at Ben Lomond Peak and Trail (top left panel) and time series of January total precipitation at Tony Grove Lake and Payson (top right panel). Standardized anomalies of the time series are shown in the lower panels.

Scatter plots of the values associated with two variables are a convenient way to examine relationships between paired data. I tend to use the basic **plot** command but you can also use a stattool variant with more options, **scatter**. The code also demos the **polytool** command, which will become especially useful later in the course. Clustering, spread, outliers, etc. become apparent in scatter plots. Scatter plots can be done in terms of the raw values, anomalies, or standardized anomalies depending on the application. Since temporal continuity is lost when looking at scatter plots generated from time series of data, you need to be careful to not simply assume that each pair of observations is independent of the others.

Figure 4.3 shows scatter plots of the original and standardized anomalies for the two sets of paired data. The meaning of the lines in each of the panels will become apparent below. Scatter plots are easier to interpret when there is a clear one-to-one association between the two variables, i.e., for a given value of x , the values of y in the sample are similar to one another. If the scatter plot looks like a blob, then that is a clear indication of a lack of one-to-one association. If the pairs of values tend to fall along a line, then it is appropriate to think of the two variables as being linearly related to one another. They may instead exhibit quadratic or

higher order association. The scatter plots on the left side reflect greater linear correspondence between the Ben Lomond stations compared to the Tony Grove/Payson pair.

As part of exploratory data analysis it is common to want to estimate the values of one variable from that of another. I'm going to avoid saying 'predict' one variable from the other for the moment. Let's start by trying to estimate precipitation at Ben Lomond Peak from the values at Ben Lomond Trail. First, we know that there is more precipitation on average at the higher elevation site, so we need to consider the differences between the two means. The simplest linear approach is to assume that for a given value at Ben Lomond Trail \hat{x}_i , our estimate \hat{y}_i (where the subscript i refers to a particular year) at Ben Lomond Peak can be determined as follows:

$$\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x}) \quad (4a.1)$$

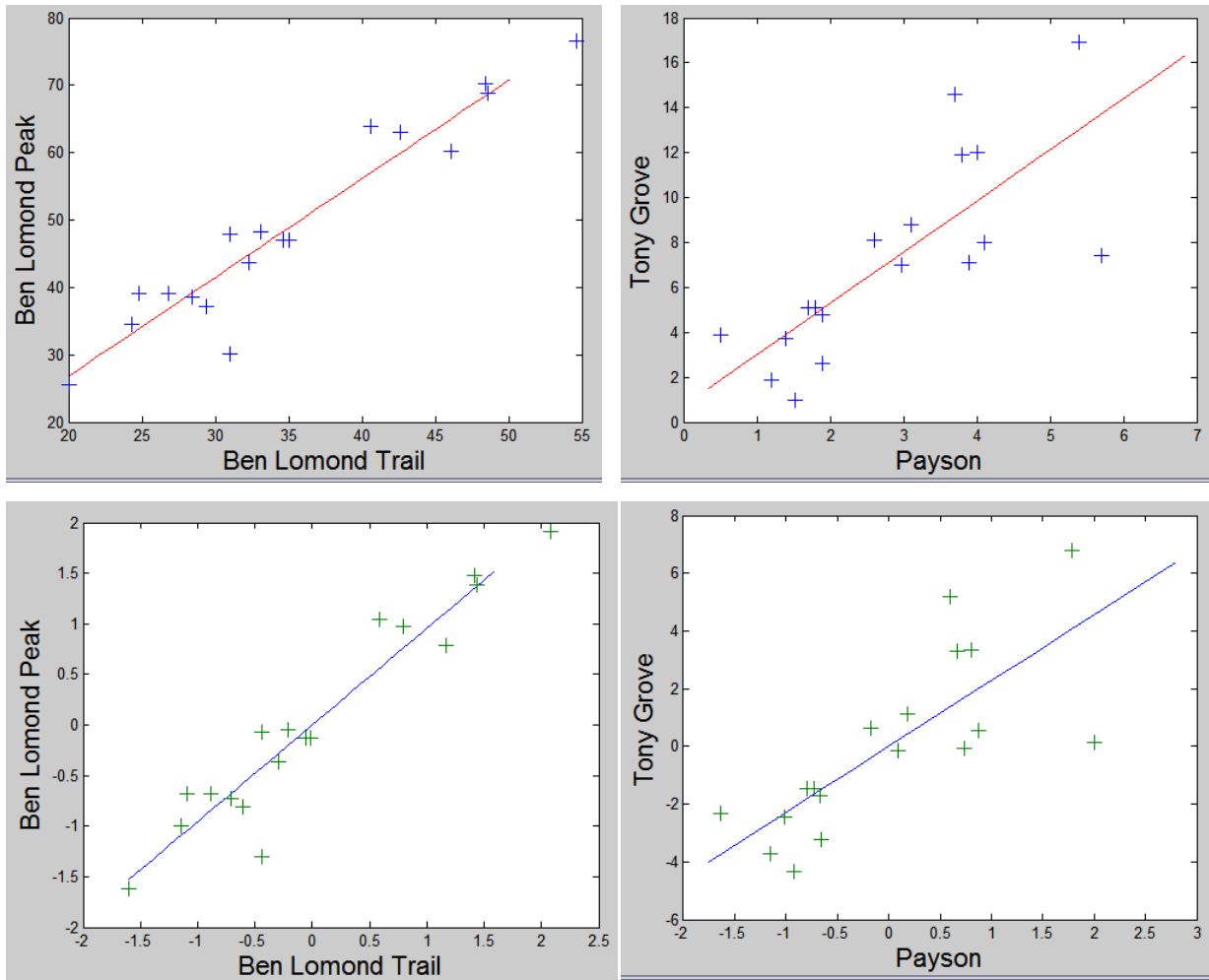


Figure 4.3. Scatter plot of seasonal total precipitation (in) at Ben Lomond Peak and Trail (top left panel) and scatter plot of January total precipitation at Tony Grove Lake and Payson (top right panel). Scatter plots of standardized anomalies are shown in the lower panels.

That relationship takes into consideration the differences in the two means. We obviously need to figure out how to determine the coefficient, b , and one approach is to use our sample of data collected over the eighteen year period. First, consider the lines in each of the top panels. They are particular linear estimates using specific values of b . If the Ben Lomond Trail precipitation is 30 inches, then we would estimate Ben Lomond Peak to measure 41.5 inches. If Payson measured roughly an inch, then we would estimate Tony Grove to have 2 inches.

Alternatively, we can use another coefficient, r , to estimate the standardized anomalies at Ben Lomond Peak from the standardized anomalies at Ben Lomond Trail as $\hat{y}_i^* = r\hat{x}_i^*$ where the asterisk indicates a standardized anomaly and r again needs to be determined from the pairs of values in the samples. Linear estimates for particular values of r are shown by the lines in the lower panels. If r is 1, then the standardized anomalies at the two sites would be estimated to be exactly the same. If r is -1, then they would have the same magnitudes but opposite signs of anomalies. If r is 0, then for any x standardized anomaly, the estimate for y would be 0.

How good are those estimates? We can use our sample of data to compute the errors for these specific choices of b (the slope of the line). For example, we have two observations of Ben Lomond Trail precipitation of roughly 31 inches and during those years, Ben Lomond Peak measured 30 inches in one year and 48 inches in the other. Obviously, our linear estimate didn't do particularly well in those cases, but most of the other years had closer estimates to those observed. When Payson's precipitation is high (above 5 inches), we don't have a lot of confidence that we can estimate the precipitation at Tony Grove, since our linear estimate is quite different from the two observed January totals at Tony Grove.

Any particular error in the estimate can be written as $e_i = y_i' - \hat{y}_i$, which is the distance between the line and the specific observation. The best line will be the one which minimizes all the distances e_i , so we want $\sum_{i=1}^n e_i^2$ to be a minimum. For our sample values $y_i' = bx_i' + e_i$, where the primes denote deviations from the respective means. Then if we use the entire sample:

$$\overline{y_i'^2} = b^2 \overline{x_i'^2} + 2b \overline{x_i' e_i} + \overline{e_i^2} \quad (4.a.2)$$

The term on the left is the sample variance of y about the mean and is given as the sum of the variance explained by the linear fit + how the errors and the deviations of x are related over the entire sample + the variance that is not explained by the linear fit, which is what we want to be small. The middle term on the right is assumed to be zero, because e_i is assumed to be random if our sample is large enough.

- Then $s_y^2 = b^2 s_x^2 + \overline{e_i^2} \quad (4.a.3)$

We want to choose b so that the explained variance of the linear fit (the first term on the right) is as big as possible and the last term is as small as possible.

To minimize $\sum_{i=1}^n e_i^2$ means to determine $\frac{\partial}{\partial b} \sum_{i=1}^n e_i^2 = 0$, which by substituting in for e_i yields

$$\frac{\partial}{\partial b} \sum_{i=1}^n e_i^2 = \frac{\partial}{\partial b} \sum_{i=1}^n (y_i' - bx_i')^2 = 2 \sum_{i=1}^n (y_i' - bx_i')(-x_i') = 0$$

or $\sum_{i=1}^n x_i' y_i' = b \sum_{i=1}^n (x_i')^2$. Dividing through by n, using the definition for a mean, and rearranging yields

$$\bullet \quad b = \overline{x_i' y_i'} / \overline{(x_i')^2} = \overline{x_i' y_i'} / s_x^2 \quad (4.a.4)$$

where $\overline{x_i' y_i'}$ is called the covariance and relates how departures from the mean of x and y are related. The covariance has units of the quantity squared, like a variance. Covariances are used in many disciplines: turbulence, planetary-scale dynamics, etc. The covariance is:

- large and positive if there is a general tendency in the sample for large and positive (and/or negative) anomalies of x occurring when large positive (negative) anomalies of y are observed
- large and negative when there is a general tendency for large and positive (and/or negative) anomalies of x to occur at the same time as large negative (positive) anomalies of y when aggregated over the entire sample
- near zero when there is a general tendency for cancellation within the sample, i.e., sometimes large positive anomaly values of x are associated with large positive anomaly values of y and other times large positive anomaly values of x are associated with large negative anomaly values of y.

Returning to 4.a.2, and dividing through by y's sample variance, then we have: $1 = \frac{b^2 s_x^2}{s_y^2} + \frac{\overline{e_i^2}}{s_y^2}$,

which simply says that a fraction of y's variance is due to the variance estimated by our linear regression estimate and the remaining fraction is due to random (or unexplained) errors. Defining r^2 as the squared linear correlation coefficient:

$$\bullet \quad r^2 = b^2 s_x^2 / s_y^2 = (\overline{x_i' y_i'})^2 / (s_x^2 s_y^2) \quad (4.a.5) \text{ or}$$

$$\bullet \quad r = (\overline{x_i' y_i'}) / \sqrt{\overline{x_i'^2} \overline{y_i'^2}} \quad -1 \leq r \leq 1 \quad (4.a.6)$$

In addition, if we standardize the anomalies of x and y by dividing the anomalies by their respective standard deviations:

$$\bullet \quad x_i^* = x_i' / s_x, y_i^* = y_i' / s_y, r = (\overline{x_i^* y_i^*}) \quad (4.a.7)$$

Then, y's sample variance can be described alternatively as: $1 = r^2 + \frac{\overline{e_i^2}}{s_y^2}$ where the squared

correlation coefficient is the fraction of the total variance of y estimated from x. While the covariance is not bounded, $-1 \leq r \leq 1$ and when:

- $r = 1$ - the linear fit estimates all of the variability of the y anomalies and the standardized anomalies of x and y vary identically
- $r = -1$ - the linear fit estimates all of the variability of the y anomalies in the sample but when the standardized x anomaly is positive, then the standardized y anomaly is negative

- $r = 0$ - the linear fit explains none of the variability of the y anomalies in the sample and the standardized anomalies of x and y have no relationship to one another in the sample.

If $r=0$, then the only thing we can say is that the best linear estimation for y is its mean value. If the scatter plot looks like a blob, then the linear correlation coefficient is likely to be close to zero, as there is no linear fit to the data that is going to explain any of the variability of y . As r approaches 1 (or -1), then we gain confidence that we can estimate the behavior of the second variable from the first, and vice versa. The squared correlation coefficient defines the fraction of variance that the two variables have “in common”.

Examine carefully the supplied program, **chapter_4_1.m**. The coefficients b and r are computed using several different approaches. The first approach uses (2.c.7) and requires that the sums of the product $x_i y_i$ be computed as well as the sum of squares and sums of the two variables. i.e., $\text{cov} = \overline{xy} - \overline{x}\overline{y}$. This is a useful approach when processing large data sets. The second and third approaches use linear algebra. Define the column vector \vec{X}' for the x anomalies (BLT or PAY) and the column vector \vec{Y}' for the y anomalies (BLP or TGL), then

$$\vec{X}' = \begin{bmatrix} x'_1 \\ x'_2 \\ \dots \\ x'_n \end{bmatrix} \text{ and } \vec{Y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \dots \\ y'_n \end{bmatrix} \text{ and the covariance } \overline{x'_i y'_i} = \vec{X}'^T \vec{Y}' / n \quad (3.a.7)$$

where the superscript T denotes the transpose of the column vector (i.e., the column vector is switched to a row vector). The resulting matrix multiplication of the $1 \times n$ row vector times the $n \times 1$ column vector yields a scalar number, which divided by the total number of elements, is the average of the vector product. Similar matrix multiplications can be done to obtain the sample variances.

Using the example programs, you should verify that the linear fits are as shown in the above figures and that the linear correlation between the seasonal precipitation anomalies at Ben Lomond Peak and Trail is 0.95 while the linear correlation between the January precipitation anomalies at Payson and Tony Grove is 0.38. Hence, 90% of the variance of seasonal total precipitation at Ben Lomond Peak can be explained by the variability of seasonal total precipitation at Ben Lomond Trail. In contrast, only 14% of the variance of Tony Grove seasonal precipitation totals can be estimated from that of Payson's.

The Pearson correlation coefficient is another name for the linear correlation coefficient defined here. You can compute the Pearson correlation coefficient in matlab by using the function **corr(x,y)**, i.e., it is not necessary to subtract out the mean value or any other intermediate steps. The linear correlation coefficient is not a robust and reliable statistical measure, because the covariance and variance terms are quite sensitive to outliers. The Spearman rank correlation coefficient (**corr(x,y,'type','Spearman')**) is a more robust measure and it is determined by sorting the data for the two variables in order from least to greatest and then computing the covariance as a function of rank, i.e., the correlation would be high if the highest (and lowest) values occur at the same time in both records. The Spearman approach is particularly

appropriate for analyzing variables with skewed distributions, e.g., precipitation and wind speed. Verify in the case of the pairs of precipitation records that the value of the Spearman correlation is not substantially different from that of the Pearson correlation.

There are a number of limitations of linear correlation coefficients that must be recognized:

- First, there is a widespread tendency to use correlation coefficients of 0.5-0.6 to be indicators of “useful” association. However, 75%-64% of the total variance is unexplained by a linear relationship if the correlation is in that range.
- Second, linear correlations can be made large by leaving in signals that may be irrelevant to the analysis. For example, if we correlate over many years two temperature records from opposite sides of the earth, the linear correlation will be large if we do not remove the annual cycle. Perhaps we may be interested in knowing that the annual cycle in Great Britain is similar to that in North Dakota, but usually we are more interested in examining departures from the seasonal cycle.
- Third, large linear correlations between two variates may occur simply at random, especially if we try to correlate one variate with many, many others. This situation arises frequently when we relate interannual or intraseasonal anomalies in one part of the globe to those over the entire globe. Tests are available to weed out some of these situations. We will formalize later what steps should be taken when an unexpected strong association crops up vs. one that we have hypothesized to exist.
- Fourth, relationships in the data that are inherently nonlinear will not be handled well.
- Fifth, when two time series are in quadrature with one another (e.g., one time series corresponds to a cosine and another corresponds to a sine) as shown in the figure to the left below, then the linear correlation is 0 (verify that using matlab and/or analytically). You should be able to recognize that as the relative phase of two sinusoidal time series progresses from 0 to 90 to 180, then the linear correlation changes from 1 to 0 to -1. Since the environment is filled with propagating features, the limitations of the use of linear correlations for such phenomena should be readily apparent.
- Linear correlation provides no information on the relative amplitudes of two time series. For example, the linear correlation between the two time series on the right below is 1.0, yet the amplitude of one of the time series is 3 times that of the other. The normalization by the standard deviation of each variable removes the relative amplitude information.

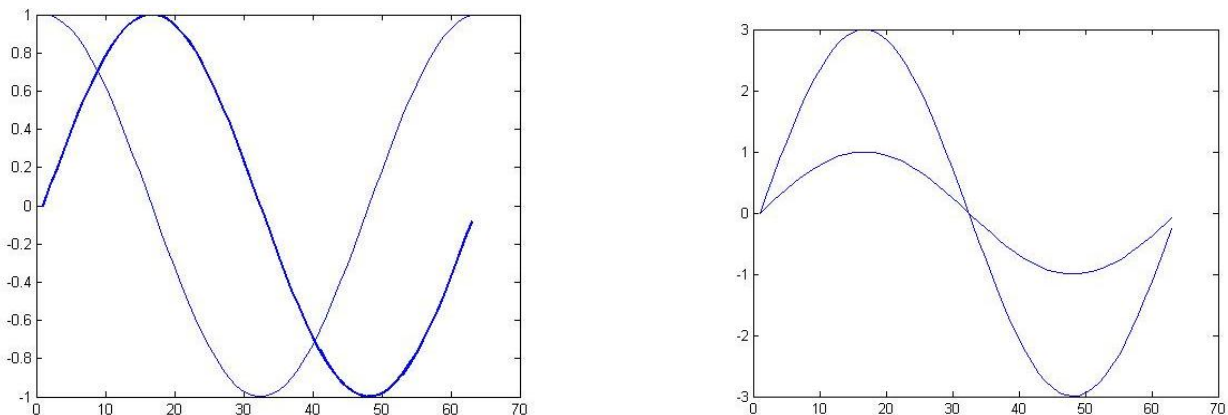
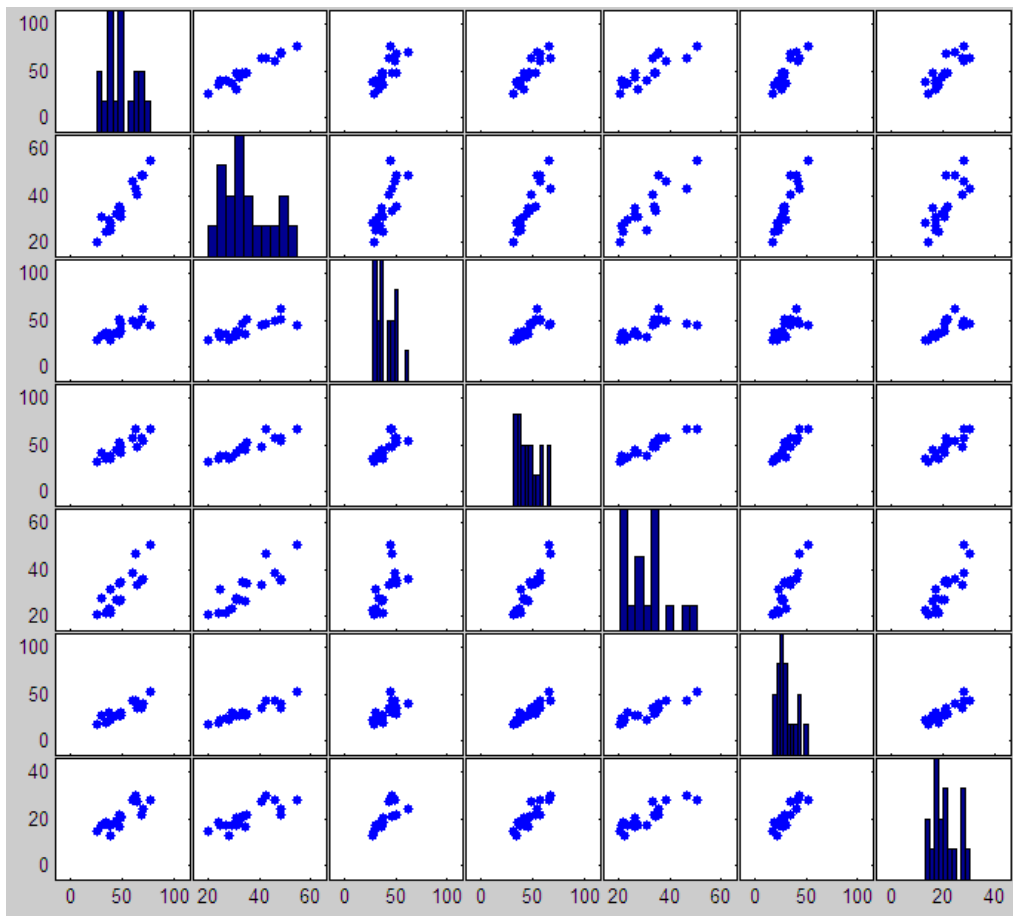


Figure 4.4. The linear correlation between the two time series on the left is 0 while that between the two time series on the right is 1.

b. Multivariate Linear Correlations

As an extension to exploratory tools for pairs of data, it is straightforward to simultaneously examine the association between many simultaneous observations. I'll use as an example the seasonal totals of precipitation from the 7 SNOTEL sites. If all the observations are loaded into a single matrix as shown in the demo code **chapter_4_2.m**, then *plotmatrix* can be used to view scatter diagrams between all pairs of simultaneous observations as shown in Fig. 4.5. The diagonal subplots are histograms while the scatter subplots below the diagonal are simply inverted from those above the diagonal. There are obviously some strong associations (e.g., the strongest linear relationship appears to be the one already examined in the previous section between Ben Lomond Peak (1st row and column) and Ben Lomond Trail (2nd column and 2nd row). One of the weaker ones is between Tony Grove Lake (3rd column and row) and Payson (6th column and row).



As shown in the sample code, we have computed the average and sample standard deviation for each of the 7 stations over all eighteen years and thereby computed the standardized anomalies for each station as a function of time. Then, we can define the $n \times 7$ two-dimensional array of standardized anomalies as \bar{X}^* where n is the total number of years and 7 is the number of stations. In other

Figure 4.5. Histograms of the 7 seasonal precipitation records (diagonal) and scatter plots between each pair of records.
words,

$$\bar{X}^* = \begin{bmatrix} x^*_{11} & x^*_{12} & \dots & x^*_{17} \\ x^*_{21} & x^*_{22} & \dots & x^*_{27} \\ \dots & \dots & \dots & \dots \\ x^*_{n1} & x^*_{n2} & \dots & x^*_{n7} \end{bmatrix} \quad (4.b.1)$$

A Hovmuller diagram (time vs. location) is simply a plot of the matrix defined in 4.b.1. For example, all the stations had large standardized precipitation anomalies during the 2005 season but the positive standardized anomaly at Tony Grove (column 3) was smaller than that at all the other locations during that year. Tony Grove had its largest precipitation anomaly during 1997.

Then, we can compute the linear correlation coefficients between every pair of stations (pairs of columns) from

$$\bar{R} = \bar{X}^*{}^T \bar{X}^* / n, \quad (4.b.2)$$

where \bar{R} is a 7 x 7 matrix. This computation is trivial in matlab: **corr(x)** or **corrcoef(x)**, i.e., those

functions handle computing the standardized anomalies and the other intermediate steps. However, the sample code also shows how to compute the linear correlation matrix using matrix algebra. The resulting matrix for this example is shown graphically in Fig. 4.6. Also, look at the values after running the sample code. The correlation between each time series and itself is 1 (the diagonals). In addition, the correlation matrix is symmetric, i.e., the columns and rows are interchangeable for a given index value.

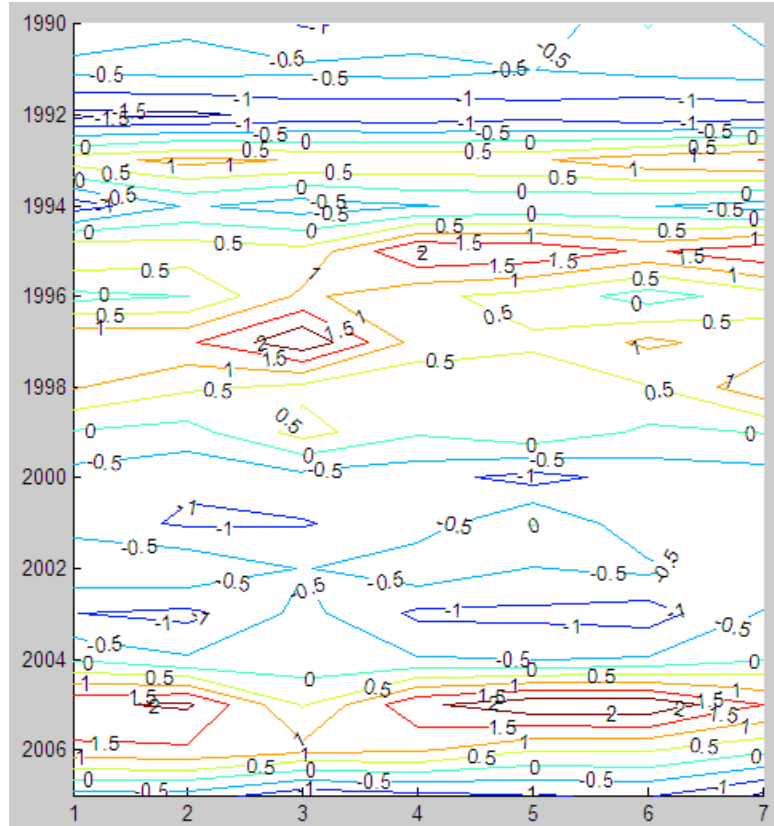


Figure 4.6. Hovmuller diagram (time increasing down the page and location advancing across the page) of standarized seasonal precipitation anomalies.

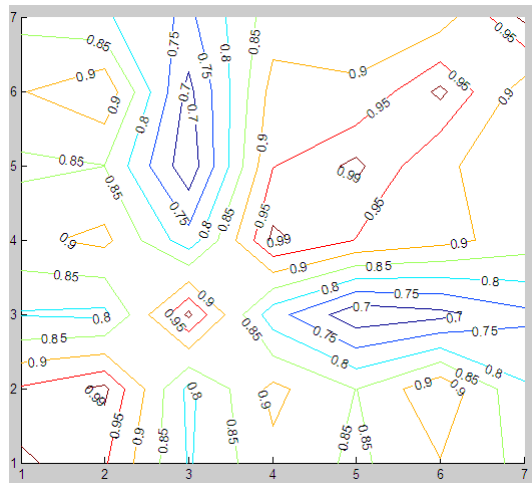


Figure 4.67. Correlations between every possible pair of seasonal precipitation totals.

The matlab contouring routine doesn't do a particularly good job of indicating that the values on the diagonal are 1. The lowest correlations (less than 0.7) are between Tony Grove Lake (column or row 3) and Timpanogos and Payson (column or row 5 or 6). This result shouldn't be too surprising, since they are the ones separated by the largest distance and there are some differences in the temporal evolution of the precipitation anomalies over time evident in the Hovmuller diagram of Fig. 4.6. We could use the Spearman correlation (**corr(x,'type','Spearman')**) to reduce the sensitivity of the correlation matrix to outliers.

The above exploration of the data centers on the question: how do seasonal precipitation departures from the eighteen year temporal mean at one location compare to those at another location when considered over all 18 years?

Linear correlations of this sort are commonplace in the environmental field. A time series of one variable is often correlated with time series of variables at every location on a grid result. That results in a temporal anomaly correlation map. As shown in Fig. 4.8 from Horel and Wallace (1981, *Mon. Wea. Rev.*, 813-829), time series of 700 mb height at grid points poleward of 20°N are related to various indices. “Teleconnection” maps are where the time series at each gridpoint of a

variable is related to the time series of that same variable at every point and then this procedure is repeated for every possible gridpoint. Then, the largest correlation values for each gridpoint for locations beyond a specified range are tabulated and displayed on a single figure.

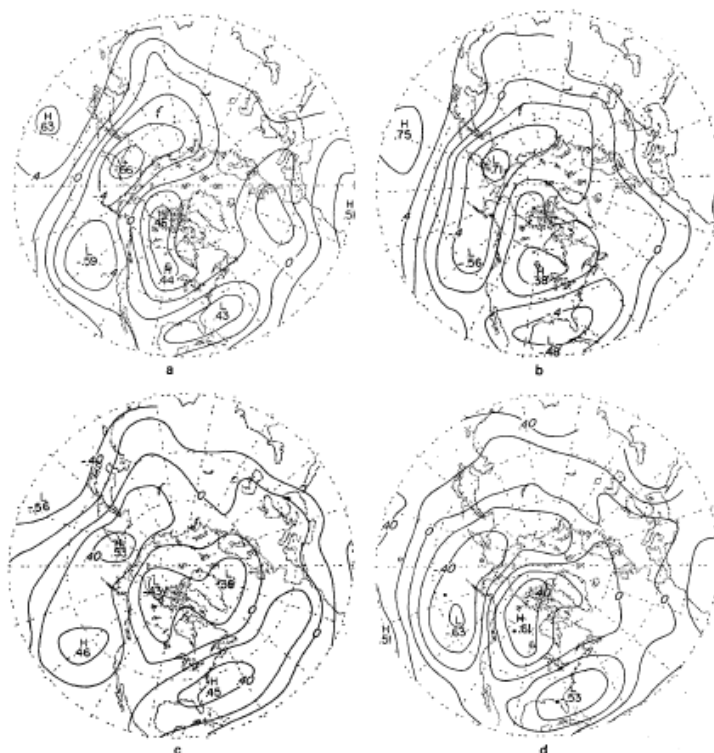


FIG. 9. Correlation coefficients between 700 mb geopotential height at gridpoints poleward of 20°N and (a) the Sea Surface Temperature Index, (b) December–February rainfall at Fanning, (c) our Southern Oscillation Index, and (d) the tropical 200 mb Index. Contour interval 0.2. The locations of the centers of action of the Pacific/North American and West Pacific patterns are denoted, respectively, by dots and open circles in Fig. 9d.

Figure 4.8. Examples of correlation maps.

Anomaly correlation maps with many different climate indices can be computed from the CDC web site:

<http://www.cdc.noaa.gov/Correlation/>. For example, Fig. 4.9 shows the correlation between the January Multivariate ENSO Index (one of the better El Niño indices (<http://www.cdc.noaa.gov/people/klaus.wolter/MEI/index.html>)) with 500 mb January monthly height anomalies. Positive correlations imply that when the SST in the equatorial Pacific is above (below) normal then 500 mb heights are above (below) normal. The tendency during El Niño winters for enhanced troughing (lower than normal 500 mb heights) in the Gulf of Alaska and over the southern U.S. combined with above normal heights in western Canada is evident.

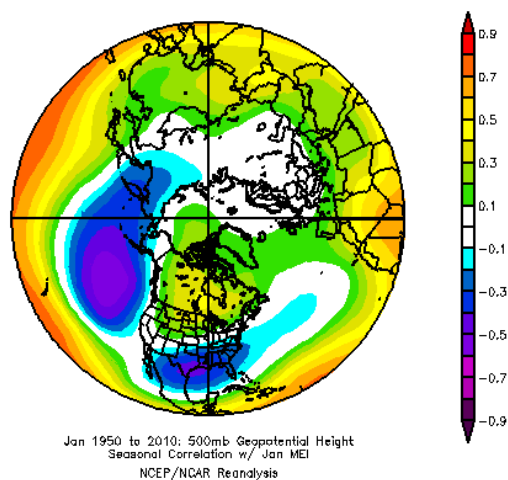


Figure 4.9. Correlation between the MEI index during January with 500 mb height anomalies in the Northern Hemisphere.

The above analysis has focused on how the

year-to-year variations in precipitation (rows) at locations (columns) relate to similar variations at other locations. Alternatively, we could transpose the original matrix and view the data as maps (rows) at specific times (columns):

$$\hat{X} = \begin{bmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \dots & \hat{x}_{1,18} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \dots & \hat{x}_{2,18} \\ \dots & \dots & \dots & \dots \\ \hat{x}_{7,1} & \hat{x}_{7,2} & \dots & \hat{x}_{7,18} \end{bmatrix} \quad (4.b.3)$$

We can then compute the spatial average over the 7 locations and the variability about that spatial average for a specific year.

We can compute the linear correlation coefficients between every pair of maps from

- $\vec{S} = \hat{X} *^T \hat{X} * / 7, (4.b.4)$

Linear correlations between pairs of anomaly maps are commonly used to verify model forecast fields vs. analysis grids. Usually, the long-term daily mean is removed at each grid point and then the departures from the spatial mean are computed for the forecast and analysis grids. Such spatial anomaly correlations have been computed for forecast grids by the operational centers for years as shown in Fig. 4.8 (in this case for the 5-day 500 mb height forecast grids in the Northern and Southern Hemisphere). If the spatial anomaly correlation was equal to one, then the forecast and the analysis would exhibit the same spatial anomaly patterns. If the correlation is 0, then the model forecast and analysis fields are completely unrelated in a linear sense.

These spatial anomaly correlations between two fields are computed as follows. Let the analysis grids at m locations (rows) and n times (columns) be \hat{X}' and the forecast grids for one specific model at m locations and n times be \hat{Y}' .

Then we can compute the spatial anomaly correlations between every matched pair of forecast and verifying analysis maps and generate a figure like that above from

- $\vec{S} = \hat{X}'^T \hat{Y}' / m, (4.b.5)$

Besides the information on the relative accuracy of the various models shown in Fig. 4.10, the magnitude of the

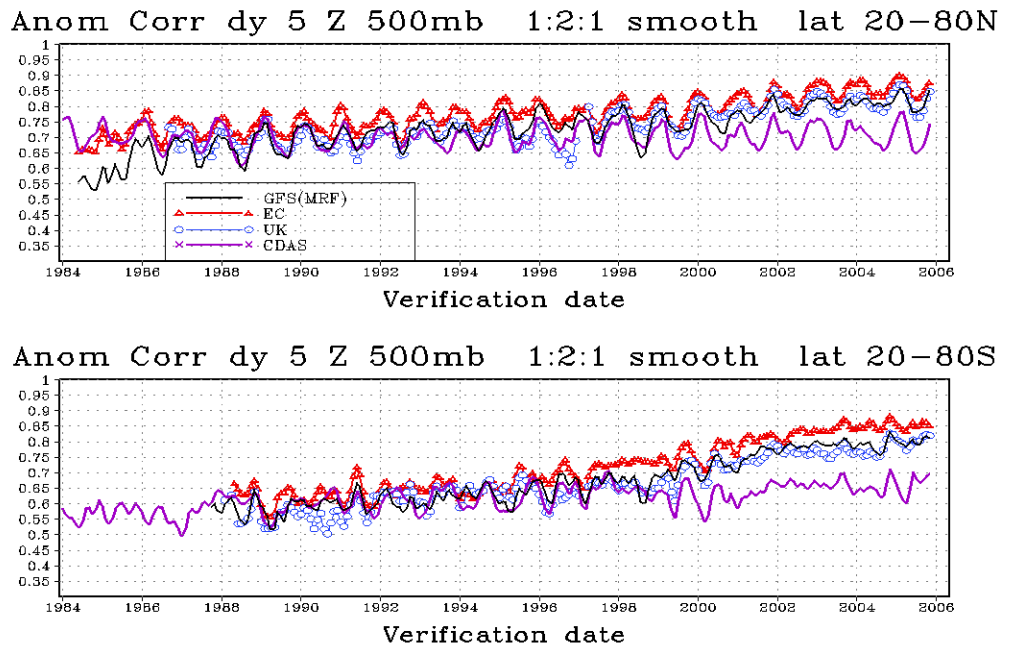


Figure 4.10. Anomaly correlations between model forecasts and analyses computed each day.

anomaly correlations indicates greater accuracy in the Northern Hemisphere than Southern Hemisphere as well as an improving trend over time. All of the caveats regarding linear correlation apply to the spatial anomaly correlations. Hence, for this type of forecast verification, we are unable to assess if the forecasts have large errors in amplitude. In addition, a relatively good forecast with a slight phasing error (i.e., ridges and troughs captured properly but shifted in longitude) will be counted as a relatively poor forecast. For many other examples of the uses of spatial anomaly correlations and other accuracy measures, browse around <http://www.emc.ncep.noaa.gov/gmb/STATS/STATS.html>.

c. Compositing

Compositing (or superposed epoch analysis) is frequently used to assess the common environmental features associated with a sample of events. For example, the occurrences of some relatively rare event are identified (e.g., local floods or warm sea surface temperature in the equatorial Pacific). The goal is to identify the average conditions within some large data set before, during, and after those rare events. The availability of the NCEP/NCAR reanalysis grids and the CDC web software available at <http://www.cdc.noaa.gov/Composites/Day/> and <http://www.cdc.noaa.gov/cgi-bin/data/composites/printpage.pl> has helped to spawn a cottage industry of compositing applications.

The steps in the compositing process can be summarized as follows (Hartmann 2005):

- select the basis for compositing and define the categories on which the compositing will be defined. It is preferable to have some physical reasoning for the categories or else the results may have limited usefulness.
- compute the means and statistics for each category
- organize and display the results
- validate the results (the methods for which we will discuss later) either in terms of: significance tests; breaking the data record into parts and showing that the results are reproducible in smaller samples; examining the relationship on an independent data set; show consistency in space and time; or verify consistency with a well-founded theory.

Relating environmental phenomena to ENSO variability is of interest in many fields. The multivariate ENSO index (Fig. 4.11) is one of the better indicators of ENSO variability (<http://www.cdc.noaa.gov/people/klaus.wolter/MEI/index.html>).

From the page <http://www.cdc.noaa.gov/people/klaus.wolter/MEI/rank.html> it is possible to identify when the biggest El Nino and La Nina events have occurred. For this example, I'll limit it to just the top 6 years during Jan-Feb during the available period of record: 1958, 1973, 1983, , 1992, 1998, and 2010. Using <http://www.cdc.noaa.gov/cgi-bin/data/composites/printpage.pl> it is very straightforward to develop the composite 500 mb height anomaly map shown in Fig. 4.12 for those 6 January's. While the basic information obtained from this simple composite is similar to that obtained from the linear correlation shown in Fig. 4.9 between the MEI and 500 mb height anomalies (i.e., below normal heights in the Gulf of Alaska and over the southern United States), the composite analysis provides information on the amplitude of the anomalies as well.

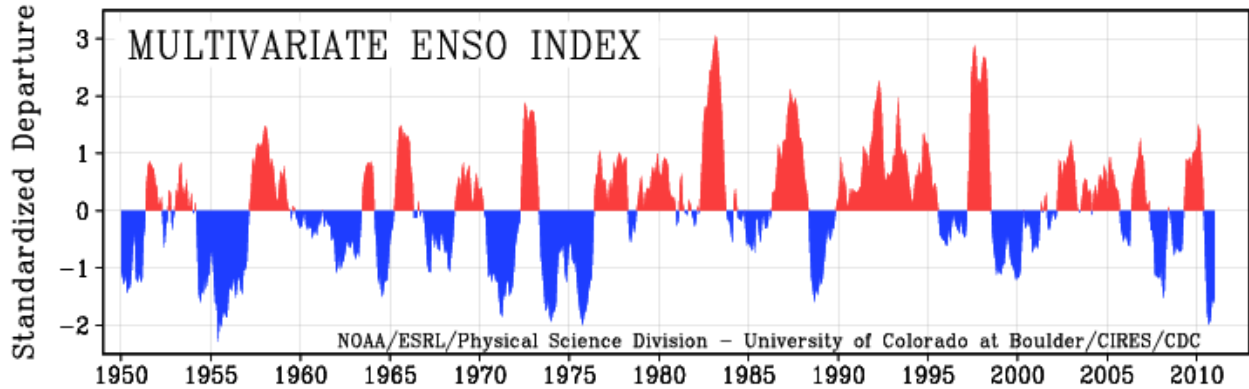


Figure 4.11. Multivariate ENSO index

One of the principal strengths of composite analysis relative to linear correlation analysis is that no assumption about the linearity of the system is made in the composite analysis. As will be discussed later, the primary limitation on composite analysis is the extent to which the sample mean can be judged to differ from the population mean. That will depend on the sample size and how much variability is present within the members of the sample.

As summarized by Hartmann (2005), compositing studies need to be carefully evaluated:

- was there a reason before the analysis started to expect the relationship found in the study? We will discuss the advantage of *a priori* expectations in greater detail later.
- what is the basis for choosing the compositing categories? How arbitrary was the selection or is it based on physical reasoning?
- was there an opportunity for subjective judgment or bias to enter the composite analysis?
- do the composite results make sense logically and physically? Are there simpler explanations possible?

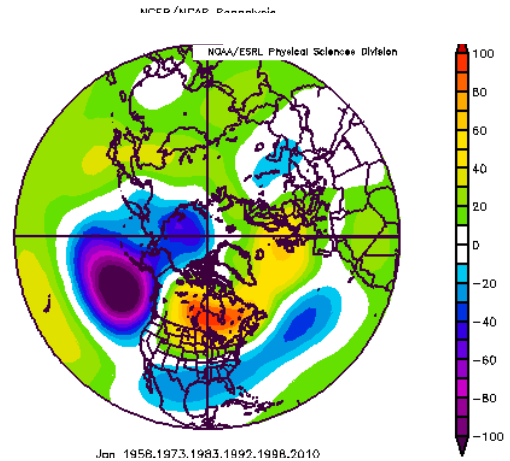


Figure 4.12. Composite 500 mb height anomaly for the 6 January's with the highest MEI index.

d. Check Your Understanding

1. Go to the NRCS web page (<http://www.wcc.nrcs.usda.gov/snow/>) and learn more about how the precipitation data set used in this chapter was created. The data we're using are from the precipitation sensor. What are some of the differences between how the precipitation is measured compared to snow water equivalent?
2. Estimate the number of degrees of freedom in the seasonal total precipitation records at Ben Lomond Peak and Trail.
3. Using the information obtained from the chapter_4_1.m code, write out the regression equation between Ben Lomond Trail and Ben Lomond Peak (i.e., 4.a.1). Imagine that a drought hits the Wasatch Front and the seasonal total precipitation at Ben Lomond Trail was only 10

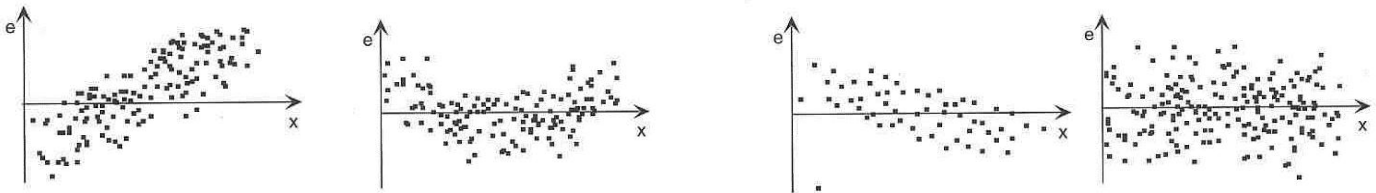
inches. What would be the estimated precipitation at Ben Lomond Peak? Write out the regression equation for precipitation at Ben Lomond Trail estimated from that at Ben Lomond Peak. It's even a worse drought year and the precipitation at Ben Lomond Peak is only 10 inches. What would be the precipitation estimated at Ben Lomond Trail?

4. Experiment with the polytool command as used in the demo code chapter_4_1.m (but ignore the red dashed lines for now). What does robust regression mean as opposed to least squares regression? How does the regression equation change for Ben Lomond Peak precipitation estimated from Ben Lomond Trail precipitation if the regression estimate is changed from least squares to robust?

5. Consider the two vectors $x = (-1, -5, 4, 3, -2, 2, 2)$ and $y = (5, -1, 0, 0, -1, -1, 0)$. Both are temperature records, so the relevant units are °C. Create a scatter plot of x vs. y . Compute the following and be sure to include the units where appropriate : a) median of x and y ; b) mean of x and y ; c) sample standard deviation of x and y ; d) sample variance of x and y ; e) standardized anomaly vectors of x and y ; f) covariance between x and y ; g) linear regression coefficient b used to estimate y from x ; h) regression equation to estimate y from x ; i) linear correlation coefficient between x and y ; j) percent of variance in common between x and y ; k) percent of y 's total variance that is not explained by x .

6. Consider the set of 4 scatter plots between variable 'x' and variable 'e'.

(1) (2) (3) (4)



Select the scatter plots (1-4) that best describe

the following. There may be more than one scatter plot that is appropriate for each answer.

- a) linear correlation that may be affected by a single large outlier.
- b) positive linear correlation.
- c) nonlinear relationship between x and e .
- d) near zero linear correlation;
- e) negative linear correlation;
- f) large fraction of the variability of e can be explained by the variability of x .

7. Explain what the matlab function **mldivide** does.

8. The temperature at one location is completely defined by a daily sinusoid with a peak at 3PM. The daily range of temperature at this location is 10°C. The temperature at another location is completely defined by a daily sinusoid with a peak at 2 PM and a daily range of 1°C. a) Write analytic expressions for the temperature T in each case where time t is defined in hours. b) Compute the mean and standard deviation for each of the two time series over an entire year analytically. c) Determine analytically the linear correlation between the two time series over an entire year. d) Defining a time array of length 24x365 hours, verify your answers in (b) and (c) using matlab. e) The temperature at a third location is completely defined by a yearly sinusoid with a peak on Julian day 225 and a yearly range of 10°C (it has no diurnal cycle). Repeat the appropriate parts of (a)-(d).

9. Compute the composite 500 mb height anomaly for La Nina's during January for the top 5, top 10, top 15 January-February months. How sensitive are your results to the number of months used in the composite? Composite similarly and discuss another atmospheric field of interest to you for the January-March period.