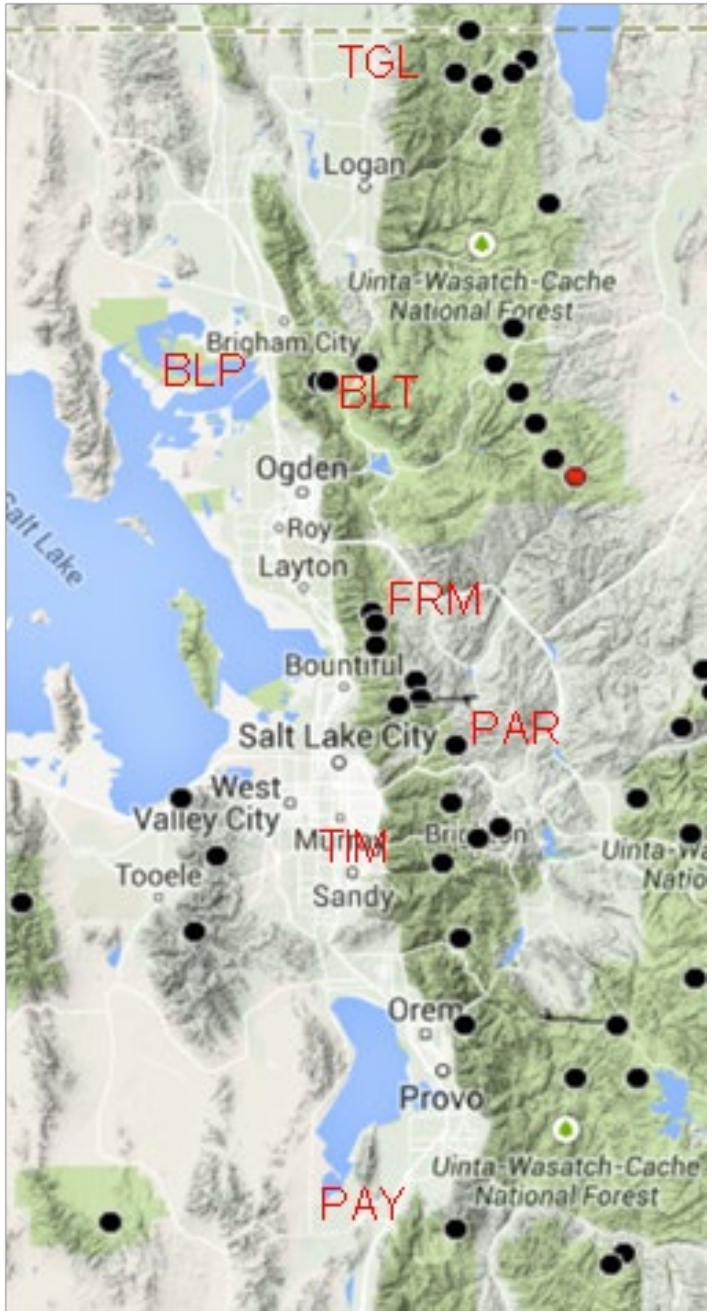
Today

SNOTEL Sites



Accumulated Annual Precipitation

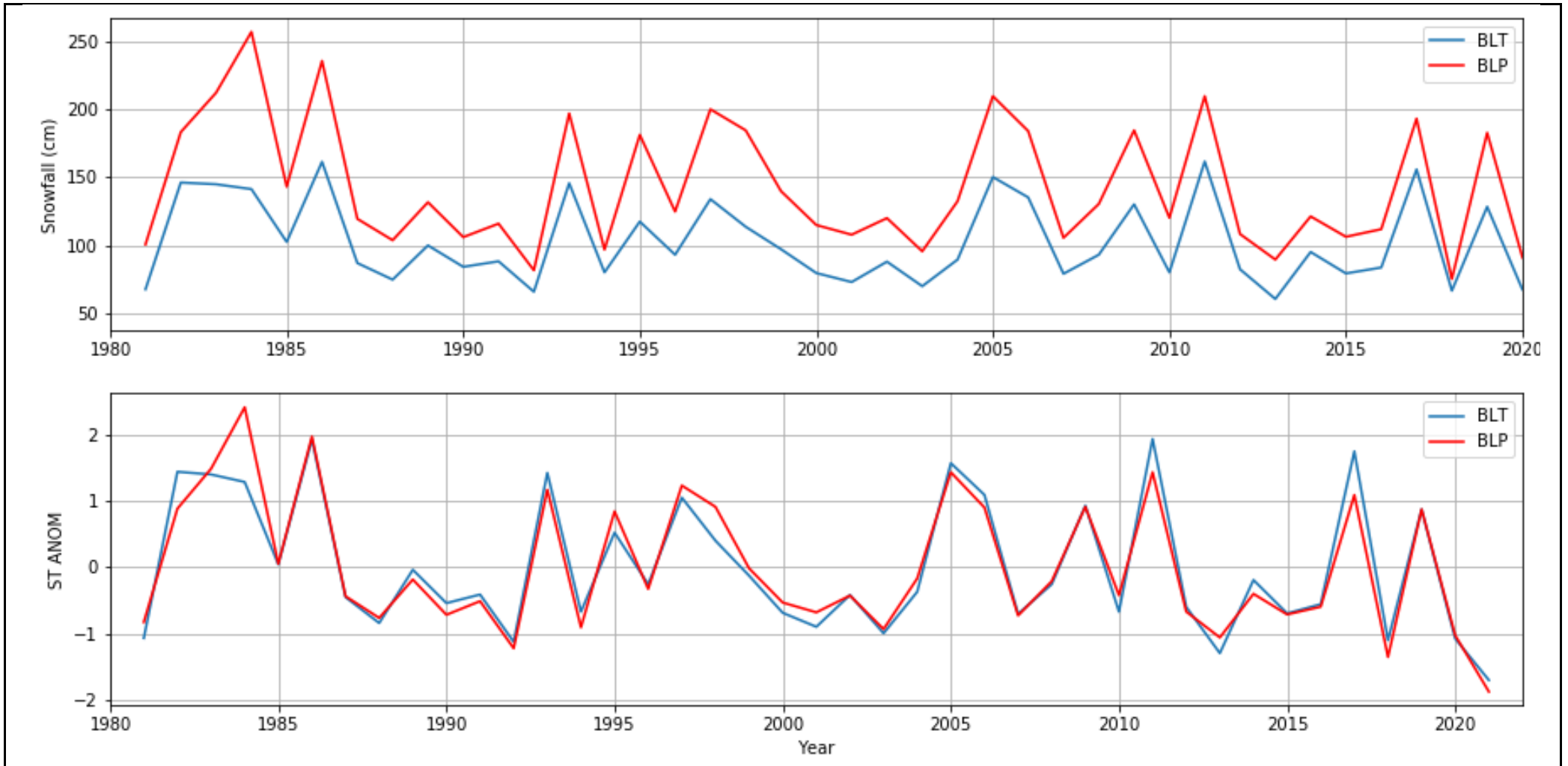


Figure 4.2. Time series (cm) of seasonal total precipitation at Ben Lomond Peak and Trail (top panel) and standardized anomalies (nondimensional) of the time series in the lower panel.

Scatter plots

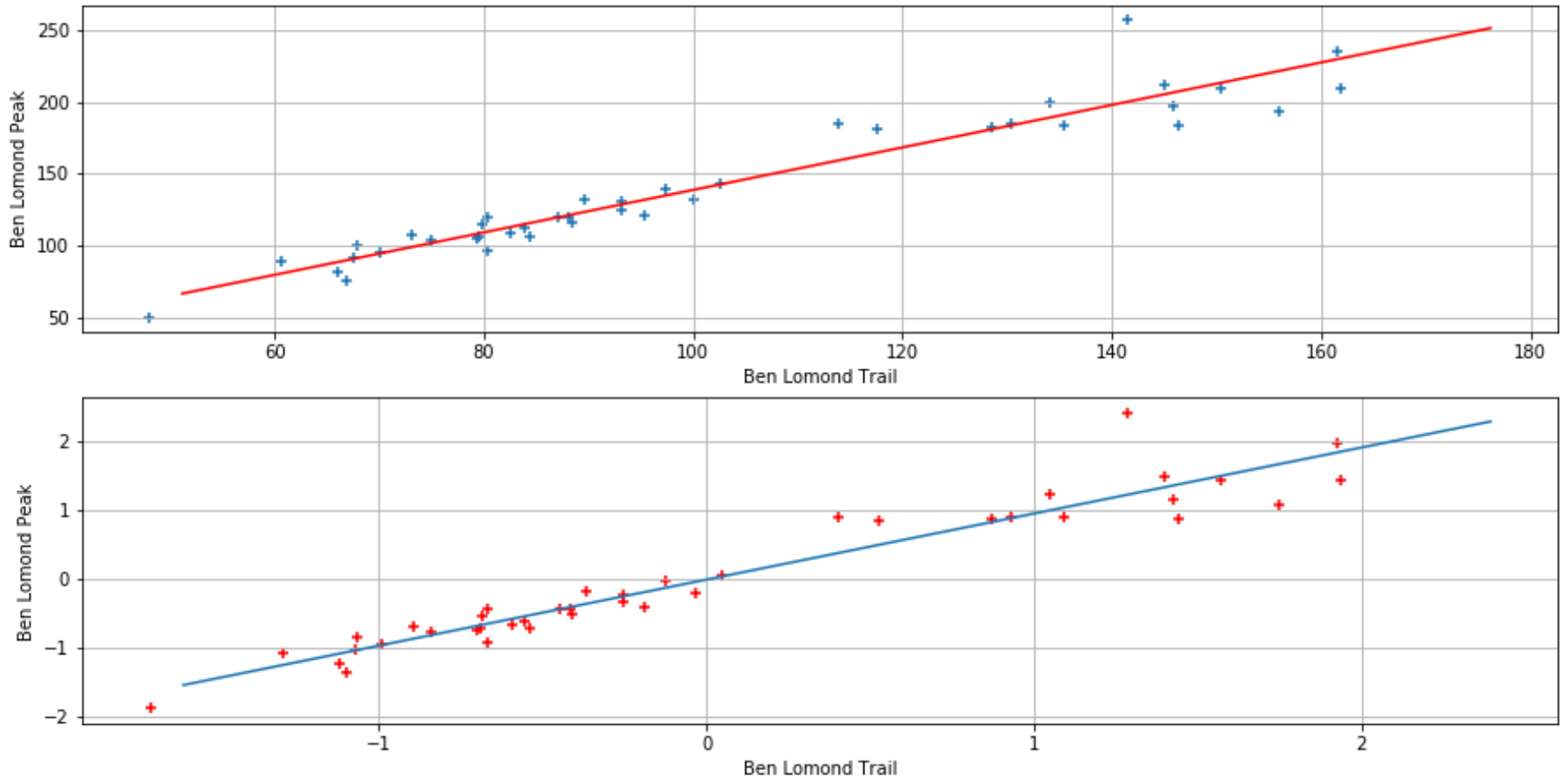


Figure 4.3. Scatter plot of total precipitation (cm) at Ben Lomond Peak vs. Trail (top panel) and of their standardized anomalies (bottom panel).

Estimating Values of One Variable From Another

- X- Ben Lomond Trail
- Y- Ben Lomond Peak
- Want to estimate Peak from Trail
- Use pairs of observations from sample
- Need to determine coefficient b or r
- b- slope of linear estimate
- r- linear correlation

$$\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x})$$

$$\hat{y}^*_i = r\hat{x}^*_i$$

Definitions

- Estimate $\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x})$
- Error of estimate $e_i = y'_i - \hat{y}'_i$
- Want $\sum_{i=1}^{n} e_i^2$ to be a minimum
- Need to find the value of b that minimizes that sum

$$\frac{\partial}{\partial b}\sum_{i=1}^{n} e_i^2 = 0$$

$$b = \overline{x'_i y'_i} \,/\, \overline{(x'_i)^2} = \overline{x'_i y'_i} \,/\, s_x^2$$

- The value of b that minimizes the total error in the sample

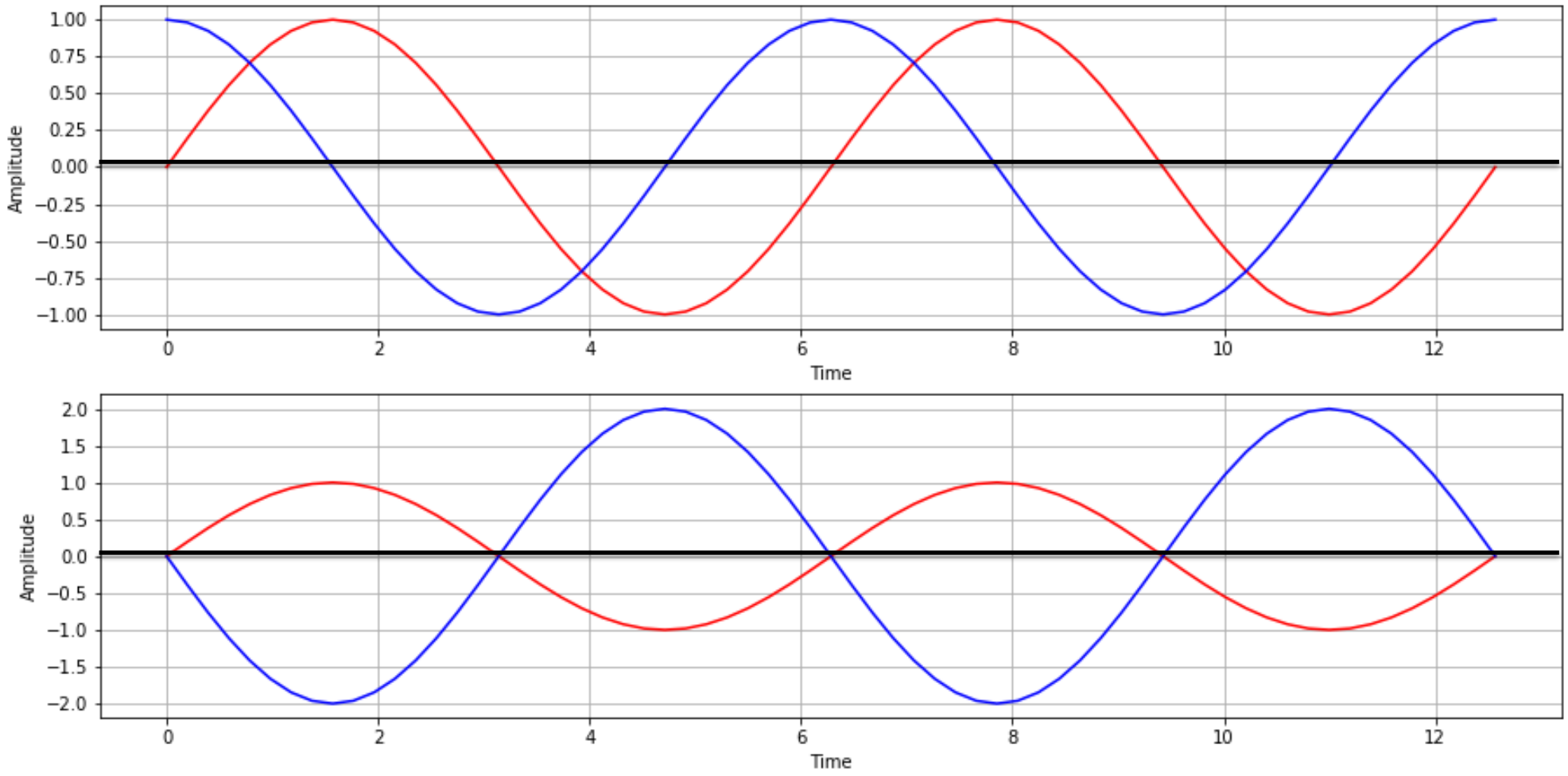


Figure 4.4. The linear correlation between the top two time series is 0 while that between the lower two time series is -1.

Covariance

- Relates how departures of x and y from respective means are related
- Units are the product of the units of the two variables x and y
- Large and positive if sample tendency for:
  - large + anomalies of x occurring when large + anomalies of y AND
  - large - anomalies of x occurring when large - anomalies of y
- Large and negative if sample tendency for:
  - large + anomalies of x occurring when large - anomalies of y AND
  - large - anomalies of x occurring when large + anomalies of y
- Near zero when tendency for cancellation
  - large + anomalies of x occurring when both large – and + anomalies of y AND
  - large - anomalies of x occurring when both large – and + anomalies of y

Linear Correlation

$$r^2 = b^2 s_x{}^2 / s_y{}^2 = (\overline{x_i' y_i'})^2 / (s_x{}^2 s_y{}^2) \qquad r = (\overline{x_i' y_i'}) / \sqrt{\overline{x_i'^2}\,\overline{y_i'^2}}$$

$$x_i{}^* = x_i' / s_x, y_i{}^* = y_i' / s_y, r = (\overline{x_i^* y_i^*})$$

$$1 = r^2 + \frac{\overline{e_i^2}}{s_y{}^2}$$

y's total sample variance = fraction of variance estimated by x + fraction of variance NOT explained by x

- Dimensionless number relates how departures of x and y from respective means are related taking into account variance of x and y
- r = 1. Linear fits estimates ALL of the variability of the y anomalies and x and y vary identically
- r = -1 perfect linear estimation but when x is positive, y is negative and vice versa
- r = 0. linear fit explains none of the variability of the y anomalies in the sample. Best estimate of y is the mean value

Linear Algebra is your friend

$$\vec{X}' = \begin{bmatrix} x'_1 \\ x'_2 \\ \dots \\ x'_n \end{bmatrix} \quad \vec{Y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \dots \\ y'_n \end{bmatrix}$$

$$\overline{x'_i y'_i} = \vec{X}'^T \, \vec{Y}' / n$$

Multivariate Linear Correlations

$$\vec{X}^* = \begin{bmatrix} x^*_{11} & x^*_{12} & \cdots & x^*_{17} \\ x^*_{21} & x^*_{22} & \cdots & x^*_{27} \\ \cdots & \cdots & \cdots & \cdots \\ x^*_{n1} & x^*_{n2} & \cdots & x^*_{n7} \end{bmatrix}$$

- 7 stations and n=40 years
- Standardized anomalies

Hovmuller Diagram

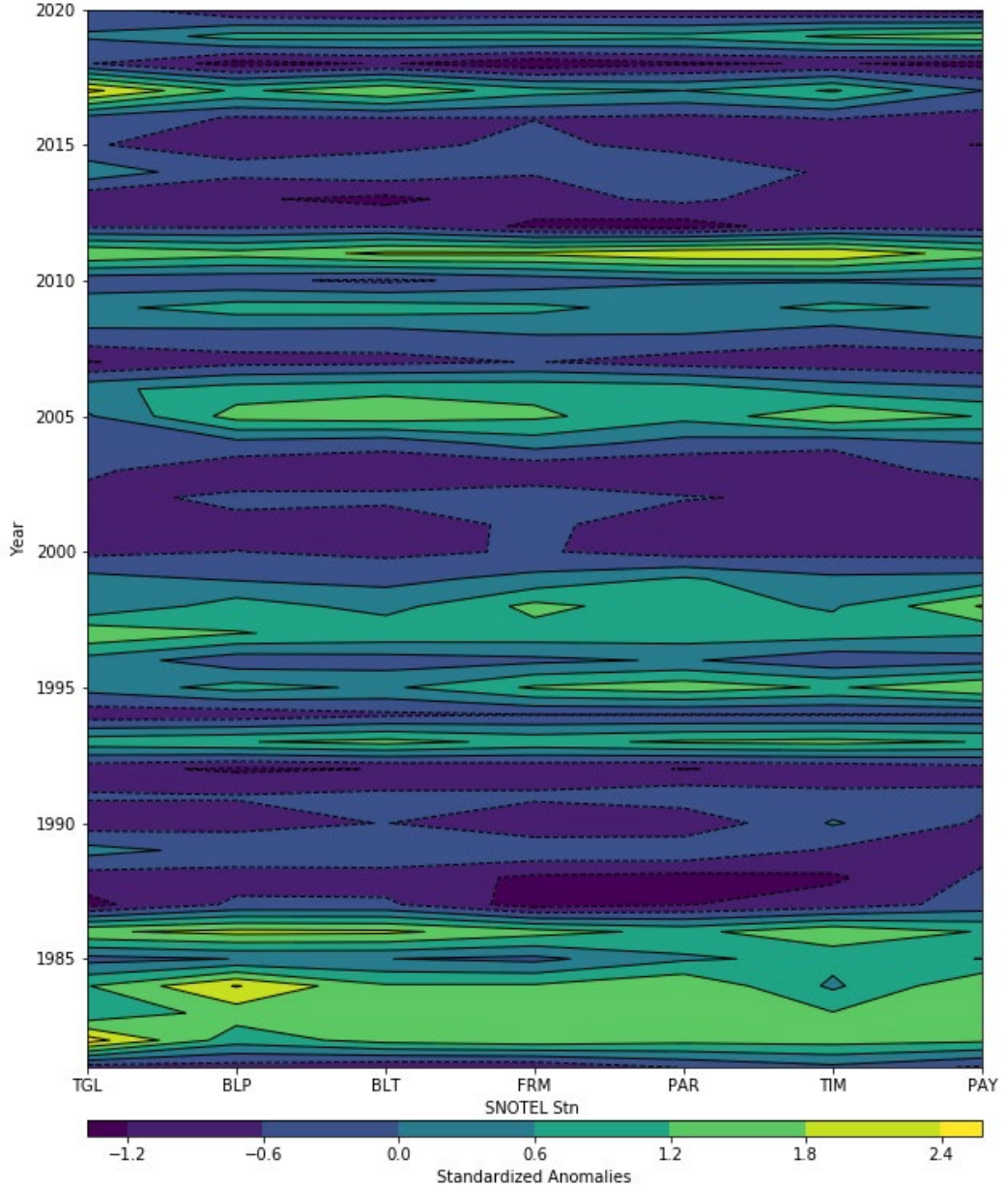
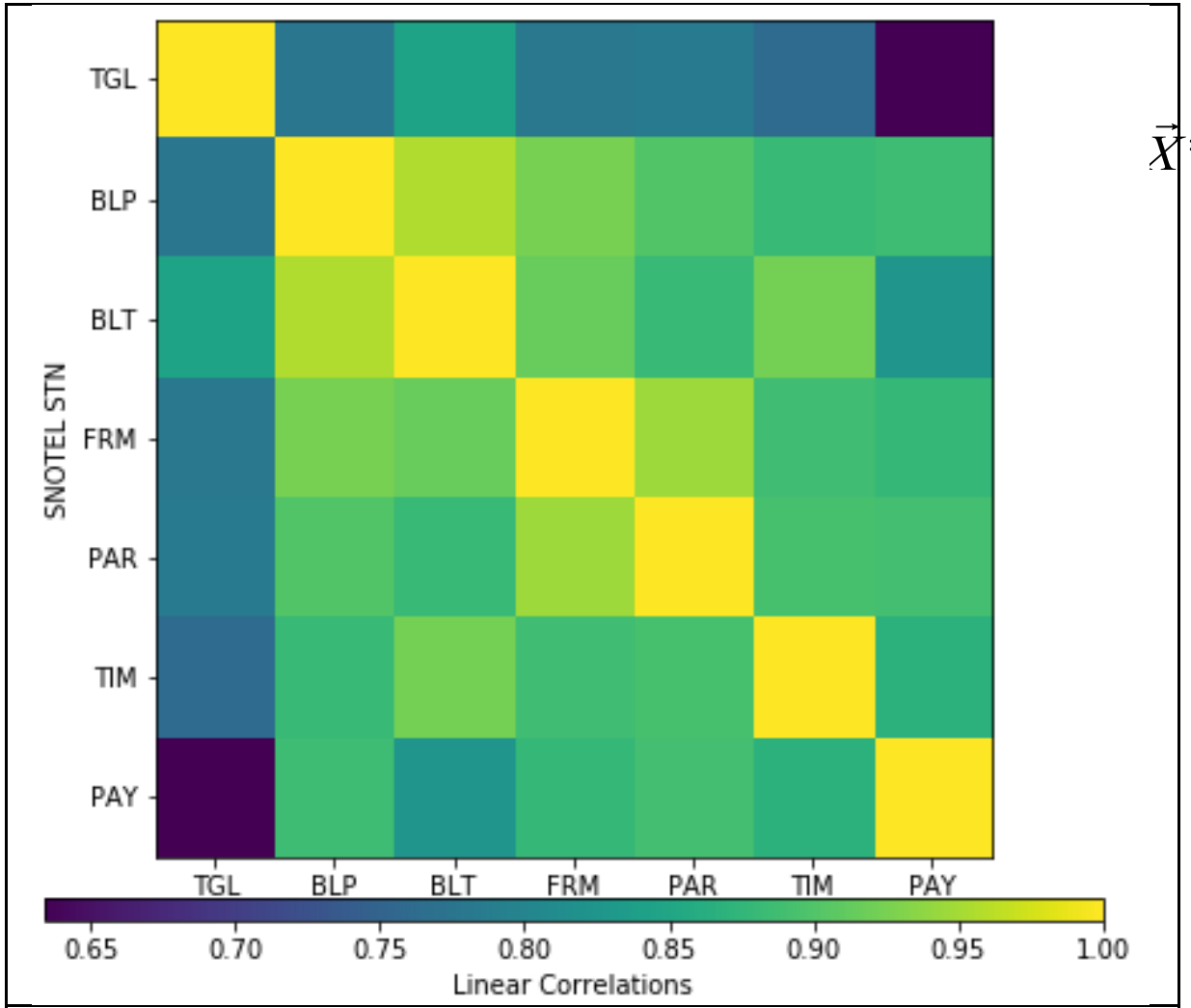


Figure 4.5. Hovmuller diagram (time decreasing down the page and location advancing south across the page) of standardized precipitation anomalies.

Multivariate Linear Correlations



$$\vec{X}^* = \begin{bmatrix} x^*_{11} & x^*_{12} & \cdots & x^*_{17} \\ x^*_{21} & x^*_{22} & \cdots & x^*_{27} \\ \cdots & \cdots & \cdots & \cdots \\ x^*_{n1} & x^*_{n2} & \cdots & x^*_{n7} \end{bmatrix}$$

$$\vec{R} = \vec{X}^{*T}\ \vec{X}^* / n$$

Figure 4.6. Correlation between the 7 pairs of SNOTEL precipitation time series computed over the 40 year sample.

Cross Validation

- Assess confidence of linear regression by applying the linear regression to an independent sample.
- Data divided in sample to “train” the regression while the rest is kept for verification of the regression.
- How the data are split between the “dependent/training” and “independent/testing” samples can be tricky,

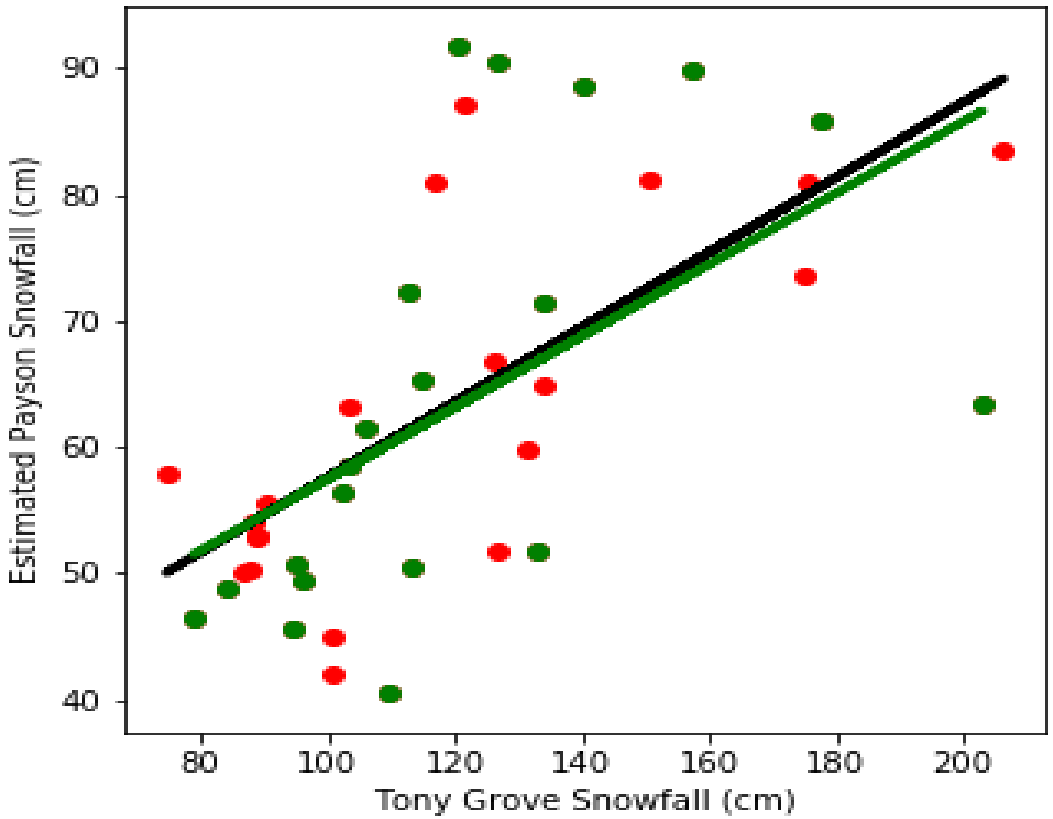


Fig 4.12. Linear fit of Payson snowfall estimated by Tony Grove snowfall using all years (the training years- red dots- and testing years-green dots) is shown by the black line. The green line is based on the linear fit using only the training years (red dots).

Principal Component Analysis

- Year-to-year variations in precipitation at the Wasatch stations are generally similar
- Want one index of Wasatch mountain precipitation over time
  - Average the records at the 7 sites?
  - Use Principal component (PC) analysis to identify structures in Wasatch precipitation?
- Purpose of PC analysis is to reduce the dimensionality of data sets through linear recombination of the variables
  - Exploratory analysis tool that can provide insight into spatial and temporal variations within a data sample
  - PC analysis is one of the most widely used multivariate statistical analyses in the environmental sciences, has a long history in other fields, and figures prominently in machine learning applications
- There are key caveats about the approach:
  - it depends on linear regression which suffers from many limitations described earlier
  - some of the implicit mathematical assumptions of the technique constrain the results into predictable and often artificial modes.

Principal Component Analysis

- PC analysis is not the only approach available to find modes.
  - Canonical correlation: identify pairs of patterns in two data sets
  - Principal oscillation analysis (POP): fit data to a linear low-order model
  - Discriminant analysis: separate data into groups defined in advance
  - Cluster analysis: separate data into groups based on similarity within sample of data
  - Artificial neural networks: adaptive system that evolves as information input during learning phase
  - Harmonic and spectral techniques.

- Before using PCs or any other of these approaches, you need to spend time deciding which is the most appropriate to use for your application.

- There are still some critical preprocessing steps to consider:
- What are the critical temporal and spatial scales for the phenomena of interest?
- Should trends or seasonal or diurnal cycles be removed?
- Are the phenomena sampled frequently enough to observe without aliasing?
- Should the data be transformed to reduce skewness, i.e., increase normality?
- Have outliers been eliminated?

1st principal component

- PC1 time series explains the maximum ***sum*** of shared variance among all 7 time series
- There are a total of 7 units of variance in the data set since each time series has been standardized to unit variance.
- The goal is to have one (or a few) PCs explain such a large fraction of the total variance in the data set that we can ignore the rest
- PC1 time series: linear combination of the 7 standardized anomaly time series with the year indicated by index i:

$$p_{i,1} = (e_{1,1}x'_{i,1} + e_{2,1}x'_{i,2} + e_{3,1}x'_{i,3} + e_{4,1}x'_{i,4} + e_{5,1}x'_{i,5} + e_{6,1}x'_{i,6} + e_{7,1}x'_{i,7}) / \lambda_1$$

- In matrix notation all 7 PC time series are:

$$\vec{P} = \vec{X}'\vec{E}\Lambda^{-1}$$

- PC analysis follows from the characteristics of symmetric matrices such as R, correlation matrix
- Any symmetric matrix can be decomposed into eigenvalues and eigenvectors as follows:

$$\vec{R}\vec{E} = \vec{E}\vec{\Lambda}$$

- The loadings or weights (E's) applied to the original time series are the elements of a eigenvector array (m x m) E that can be computed from the correlation array R where Λ is a diagonal eigenvalue array (m x m).

PC output

- Eigenvalues (the λ's) :
  - Amount of normalized variance explained by principal components
  - Percent variance explained by the first principal component is λ_1*100/m where m is the number of original time series
- Principal components (the P's):
  - If the original rows are elements of a time series, then the principal components are time series
  - These time series are linearly independent of one another (which means they are linearly uncorrelated with each other)
- Eigenvectors (the E's):
  - if the original columns were locations, then the eigenvectors are recombinations of locations, or think of them as maps
  - Each of these "maps" are linearly independent of each other (or spatially uncorrelated)

Eigenvalues: λ's

PC	Value	%
1	6.2	88.6
2	0.4	5.4
3	0.1	2.1
4	0.1	1.7
5	0.1	1.3
6	0.0	0.6
7	0.0	0.2

PC1 and PC2

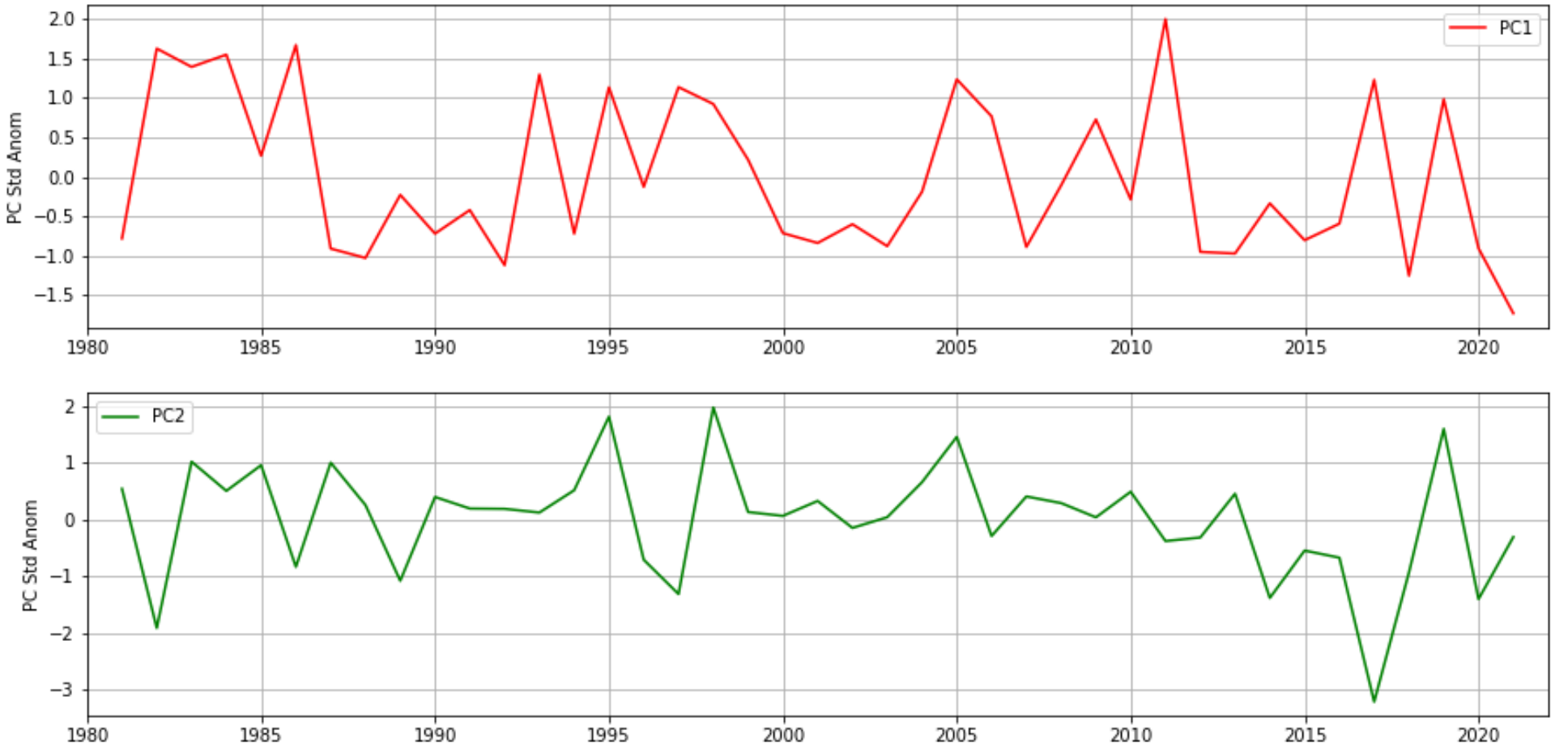


Figure 4.13. Time series of PC1 and PC2 derived from the 7 SNOTEL precipitation time series.

Eigenvectors

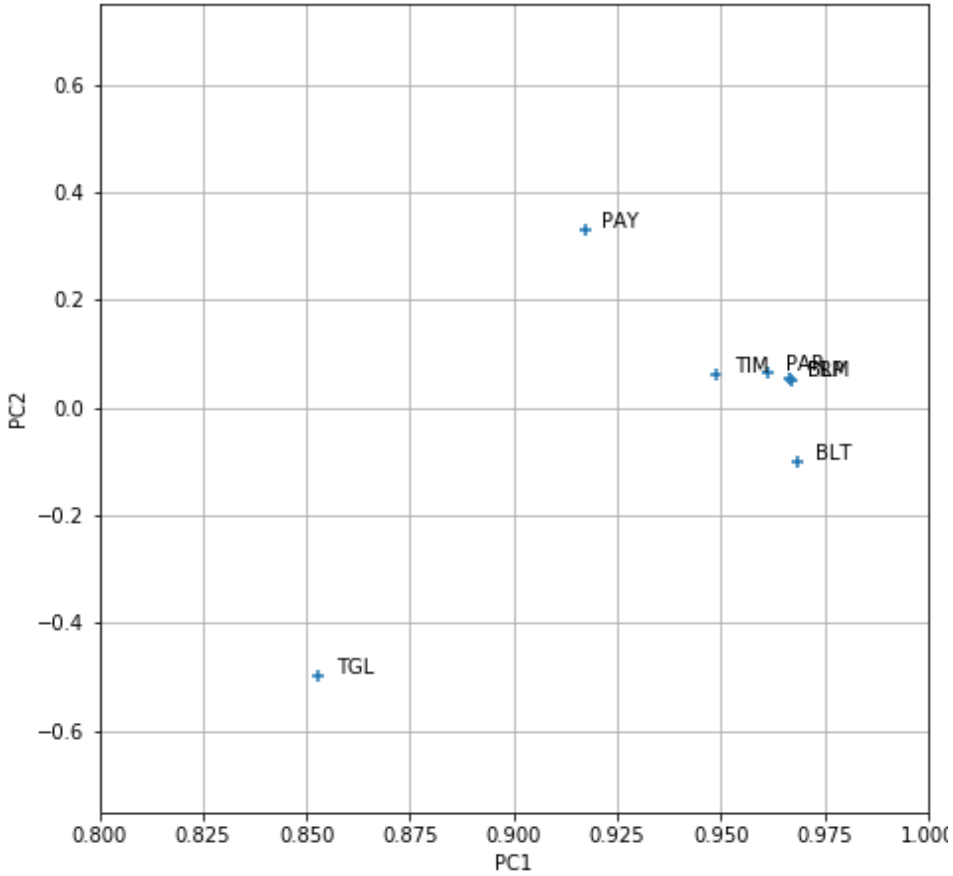


Figure 4.14. Correlations between SNOTEL time series and PC1 and PC2 time series.