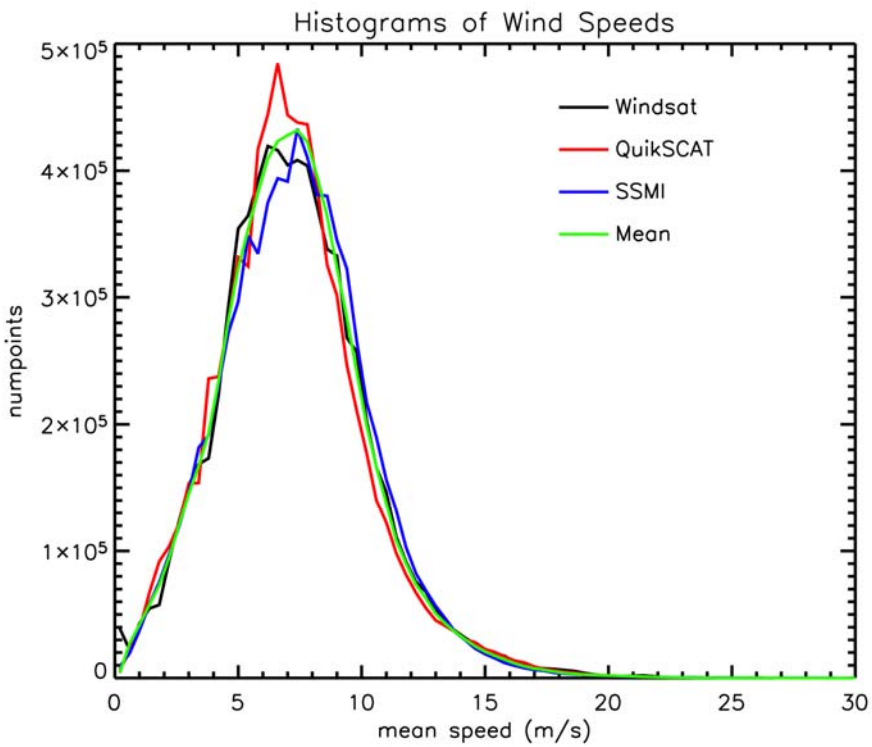
MET3220C & MET6480
Computational Statistics

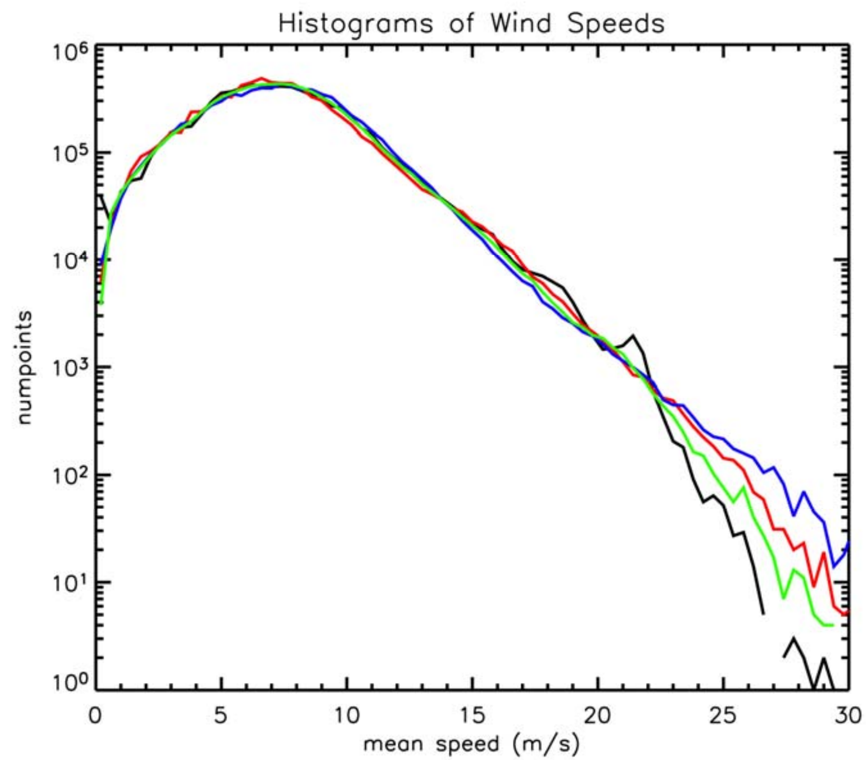
Exploratory Data Analysis
For Paired Data

Scatterplots
Correlation (several types)
Star Plot
Glyph Scatterplots

Key Point: ALWAYS LOOK AT THE DATA!!!!

Wind Speed Histograms Based on Co-located Observations From 3 Satellites





LINEAR

Better for seeing differences in frequently occurring observations

LOG

Better for seeing differences in infrequently occurring observations

Graphics from talk by Mike Freilich and Barry Vanhoff

Standard Deviation and Variance

- Recall that the standard deviation is defined as

$$s_x = \left[\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 \right]^{1/2}$$

- The variance is the square of the standard deviation.

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2$$

- The variance is a particularly useful quantity because it is additive in many applications, and the total variance is often preserved.
- For example, if a variable *f* is dependent on three independent variables *x*, *y*, and *z*, then

$$s_f^2 = s_x^2 + s_y^2 + s_z^2$$

Covariance

- Covariance is a measure of sort of like variance.
- However, covariance (cov) examines how one variable changes in proportion to another.

$$\mathrm{cov}(x, y) = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

- If x' is proportional to y', then

$$\mathrm{cov}(x, y) = s_x s_y$$

- If x' is independent of y', then

$$\mathrm{cov}(x, y) = 0$$

http://campus.fsu.edu/
bourassa@met.fsu.edu

Pearson (Ordinary) Correlation (AKA Linear Correlation)

- The Pearson correlation assumes that there is (or more accurately could be) a linear relationship between the two variables being considered: x $\propto$ y.
- This correlation coefficient is defined as

$$r_{xy} = \frac{\text{covariance}(x, y)}{s_x s_y}$$

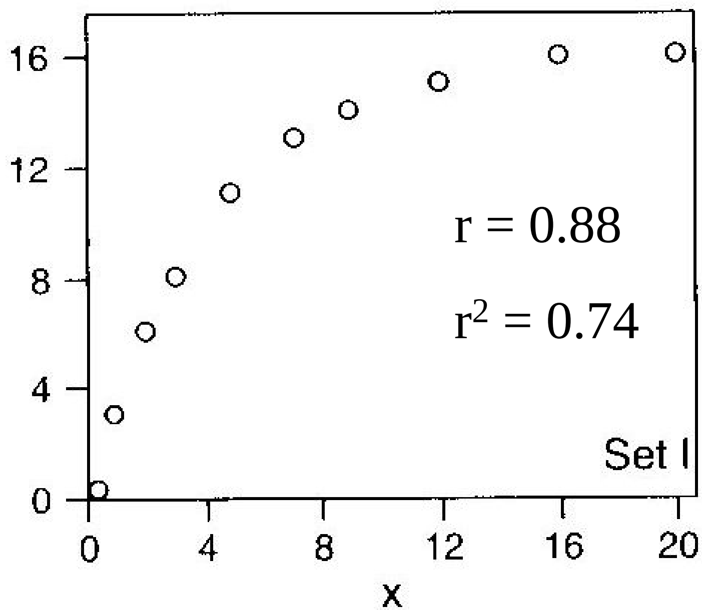
$$r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}$$

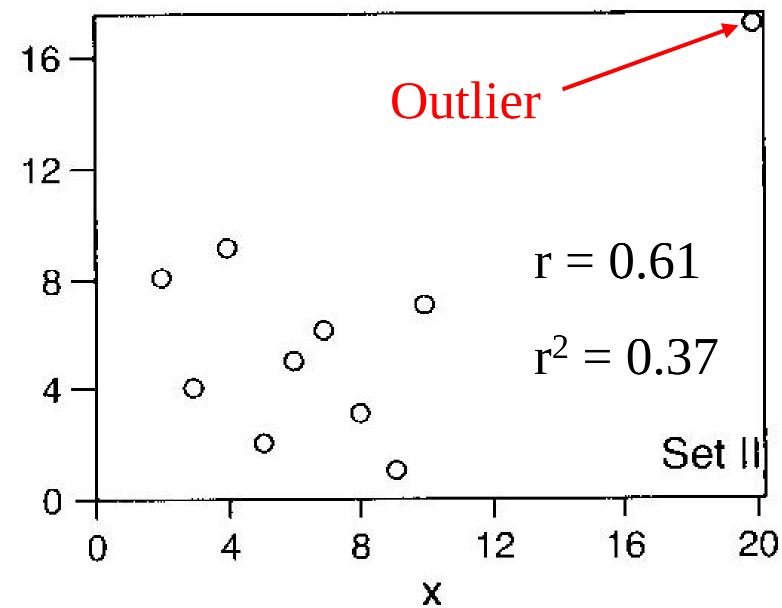
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i' y_i')}{\left[\sum_{i=1}^{n}(x')^2\right]^{1/2}\left[\sum_{i=1}^{n}(y')^2\right]^{1/2}}$$

Properties of the Correlation Coefficient

- $-1 < r_{xy} < 1$
- If $r_{xy} = 1$, then $x' \propto y'$
  - Indicating a positive slope of the best fit line
- If $r_{xy} = -1$, then $x' \propto -y'$
  - Indicating a negative slope of the best fit line
- If $r_{xy} = 0$, then x' is independent of y'
  - The slope of the best fit line is meaningless
- It is often said that r^2 is the fraction of the variance explained by a linear relationship. This is true provided that the uncertainty in both sets of observations is negligible.
  - Another key consideration is that both variables should not be calculated from the same variable or variables.
  - The above problem is called cross correlation, and it results in a much larger correlation than would be other wise determined.
- Correlation does NOT imply cause and effect.

Example Problems

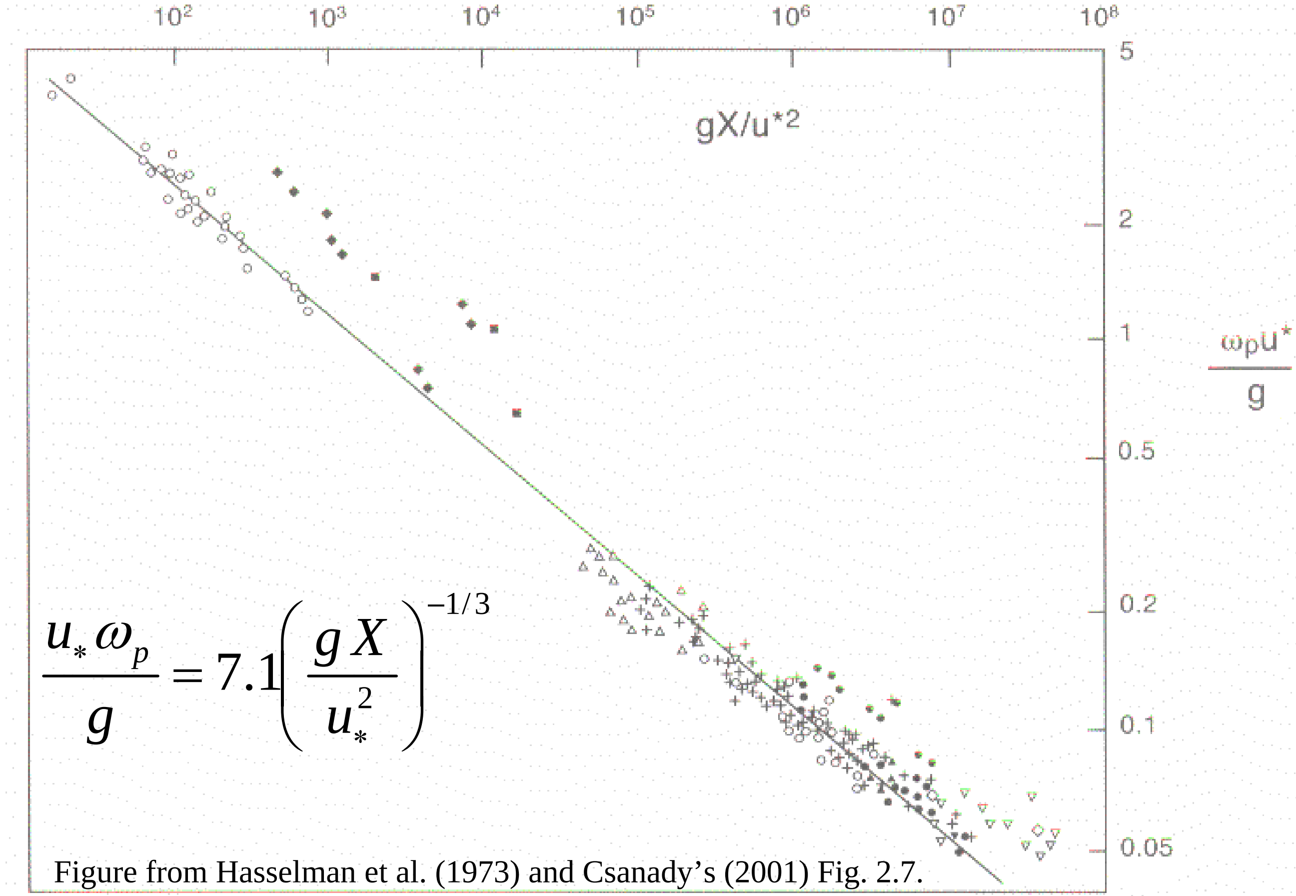




- Pearson linear correlation does not work well in either of these examples.
  - Why are there problems?
- Set I: the relationship is substantially non-linear.
  - An engineering solution might be linear fits over several ranges.
- Set II: The outlier leads to a large covariance, resulting in a questionable value for the correlation.

Graphics from Wilk's Statistical Methods in the Atmospheric Sciences

Example of Cross Correlation
Non-Dimensional Fetch Relationship



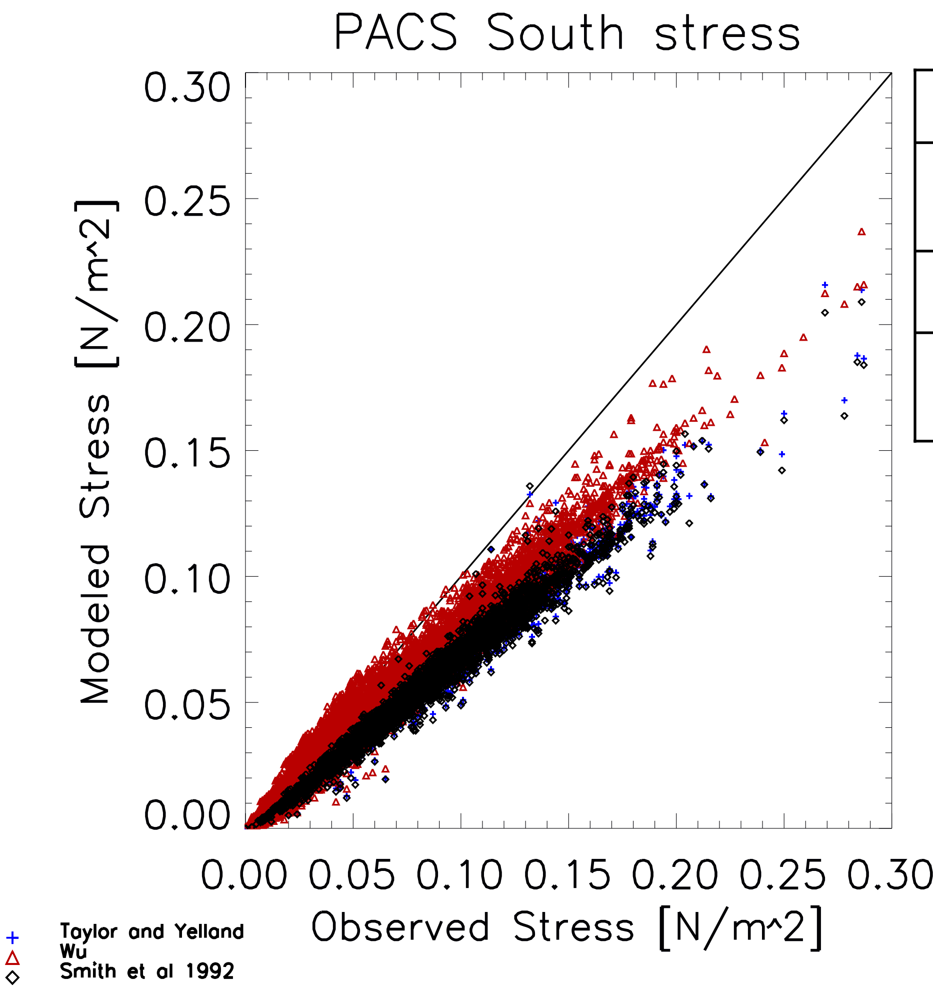
$$\frac{u_* \omega_p}{g} = 7.1 \left(\frac{g X}{u_*^2} \right)^{-1/3}$$

Figure from Hasselman et al. (1973) and Csanady's (2001) Fig. 2.7.

- g gravitational acceleration
- u_* friction velocity
- X fetch
- ω_p peak wave frequency

- Kitiagorodskii has developed a non-dimensional relationship that applies over a wide range of conditions.
- The u_* in both the x and y terms is a serious problem (cross correlation)!

Comparison of PAC3 Stress to Modeled Stress



Total	R	R^2	regression	RMSE
Taylor and Yelland	0.99	0.97	y=0.70x-0.00	0.006
Wu	0.99	0.97	y=0.79x+0.00	0.007
Smith et al.	0.99	0.97	y=0.70x-0.00	0.007

- All three models are very well correlated to the data.
- Which model is better?
  - T&Y model has lower RMS errors, but Wu's model seems to have a much better slope.

Graphics from talk by Yoshi Goto

Computationally Efficient Correlation Step 1: The Covariance

- Computational efficiency can be ignored for small data sets.
- However, for every large data sets it can be very important.
  - Example: one pass through a data set is used to determine the mean
  - A second pass is used to determine the standard deviation.
  - If the data set is read in each time, then the process is miserably slow!
- Consider the covariance squared:

$$\operatorname{cov}(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\operatorname{cov}(x,y)=\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i y_i)-\bar{y}\sum_{i=1}^{n}(x_i)-\bar{x}\sum_{i=1}^{n}(y_i)+\bar{x}\,\bar{y}\sum_{i=1}^{n}(1)\right]$$

$$\operatorname{cov}(x,y)=\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i y_i)-n\bar{x}\,\bar{y}-n\bar{x}\,\bar{y}+n\bar{x}\,\bar{y}\right]$$

$$\operatorname{cov}(x,y)=\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i y_i)-\frac{1}{n}\sum_{i=1}^{n}(x_i)\sum_{i=1}^{n}(y_i)\right]$$

Computationally Efficient Correlation
Step 2: The Standard Deviation

- Consider the standard deviation:

$$s_x = \left[\frac{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}{n-1} \right]^{1/2}$$

$$s_x = \left[\frac{\sum_{i=1}^{n} \left(x_i^2 \right) - n\overline{x}^2}{n-1} \right]^{1/2}$$

$$s_x = \left[\frac{\sum_{i=1}^{n} \left(x_i^2 \right) - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2}{n-1} \right]^{1/2}$$

Computationally Efficient Correlation
Step 3: The Correlation

- Combine covariance and standard deviation to get correlation:

$$r_{xy} = \frac{\text{covariance}(x, y)}{s_x s_y}$$

$$r_{xy} = \frac{\cancel{\frac{1}{n-1}}\left[\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\sum_{i=1}^{n}(x_i)\sum_{i=1}^{n}(y_i)\right]}{\cancel{\frac{1}{n-1}}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]^{1/2}\left[\sum_{i=1}^{n}y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2\right]^{1/2}}$$

- All the terms in the above equation can be calculated in one pass of a data set.
- For example: reading satellite data can be extremely time intensive, and often the data are two massive to store.
- Calculating standard deviations or correlations in one pass is great.

FORTRAN Tidbit: Procedures AKA Subprograms

- There are two types of procedures: functions and subroutines.
- In general, procedures have zero or more arguments
  - E.g., subprogram1(x1, x2, x3, x4)
  - The variables x1, x2, x3, and x4 are arguments
  - Each procedure must end with the END command
  - Each procedure will cause the program to stop if the program reaches any END command
  - If procedure is not suppose to cause the program to stop, then the program must reach a RETURN command prior to the END.
- Subroutines change the value of one or more of the arguments.
  - Executed with a CALL command. E.g., CALL MEAN(x, ave)
- Functions do not alter any arguments, but return a value.
  - Example: y = mean(x)
- There are several ways to declare procedures.
  - Procedures must be declared in any program that uses them.

FORTRAN90 Example Function

- Consider a subroutine to calculate standard deviation.

```
FUNCTION STANDEV( x, n )
! n   the number of values in array x
! x   array of values for which the standard deviation will be determined
INTEGER :: i_data, n
REAL :: standev, sum_x, sum_x_sqd
REAL, dimension( n ) :: x
sum_x = 0.0
sum_x_sqd = 0.0
DO i_data = 1, n
  sum_x = sum_x + x(i_data)
  sum_x_sqd = sum_x_sqd + x(i_data) ** 2
ENDDO
standev = SQRT( ( sum_x_sqd – ( sum_x ** 2 ) / REAL(n) ) / REAL(n–1) )
RETURN
END FUNCTION STANDEV
```

Alternative Correlation Methods

- Spearman Rank Correlation
  - More robust than Pearson correlation
  - Computes a Pearson correlation using the ranks of the data, rather than the actual data values.
    - Reduces the influence of outliers
    - Still hampered by noisy data
- Can be simplified to

$$r_{rank} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

  - Where D_i is the difference in ranks between the i^{th} pair of ranks.

Autocorrelation

- Autocorrelation is used to investigate how information at one time is related to information at other time.
- It is useful for examinations of:
  - Persistence
  - Repeating cycles
- Autocorrelation is a correlation of a data set with itself, but with one of the series lagged
- For example: With a series of daily temperatures $\{T_0, T_1, T_2, \ldots, T_{31}\}$ could be correlated with a series lagged by one day $\{T_{-1}, T_0, T_1, \ldots, T_{30}\}$
  - Correlations for days of two or more days could also be calculated.
- The autocorrelation for the k^{th} lag could be written as

$$r_k = \frac{\sum_{i=1}^{n-k}\left[\left(x_i - \overline{x}_0\right)\left(x_{i+k} - \overline{x}_k\right)\right]}{\left[\sum_{i=1}^{n-k}\left(x_i - \overline{x}_0\right)^2 \sum_{i=k+1}^{n}\left(x_i - \overline{x}_k\right)^2\right]^{0.5}}$$

Where the k^{th} mean is based on values from x_{i+k} to x_n

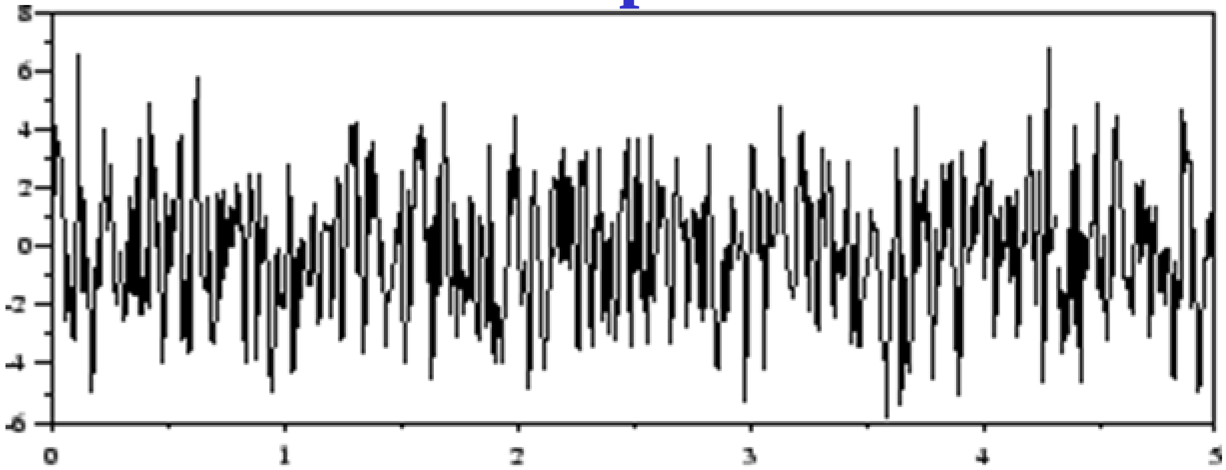
Alternative Autocorrelation

- The previous version of autocorrelation is useful when the overlapping portions of the data set are too small.
  - However, it is odd to work with means from different periods
- If there is a large number of overlapping points, then an alternative version can be applied.
  - Examines only the overlapping period.
  - The means are identical
- For example: With a series of daily temperatures $\{T_0, T_1, T_2, \ldots, T_{31}\}$ could be correlated with a series lagged by one day using the values $\{T_0, T_1, T_2, \ldots, T_{30}\}$
- The autocorrelation for the k^{th} lag could be written as

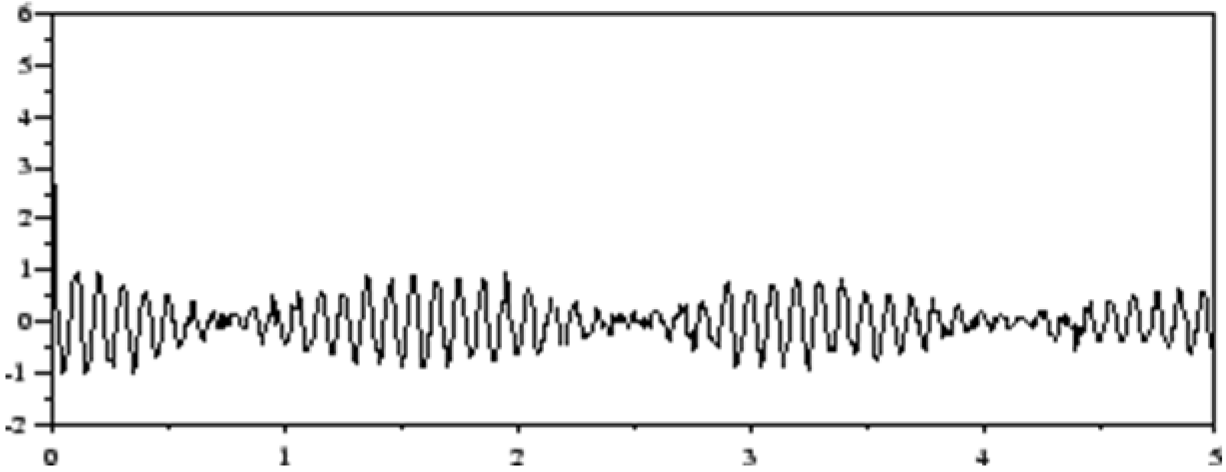
$$r_k = \frac{\sum_{i=1}^{n-k}\left[\left(x_i - \overline{x}\right)\left(x_{i+k} - \overline{x}\right)\right]}{\sum_{i=1}^{n-k}\left(x_i - \overline{x}\right)^2}$$

Where all means are based on values from x_1 to x_n

Example Autocorrelation



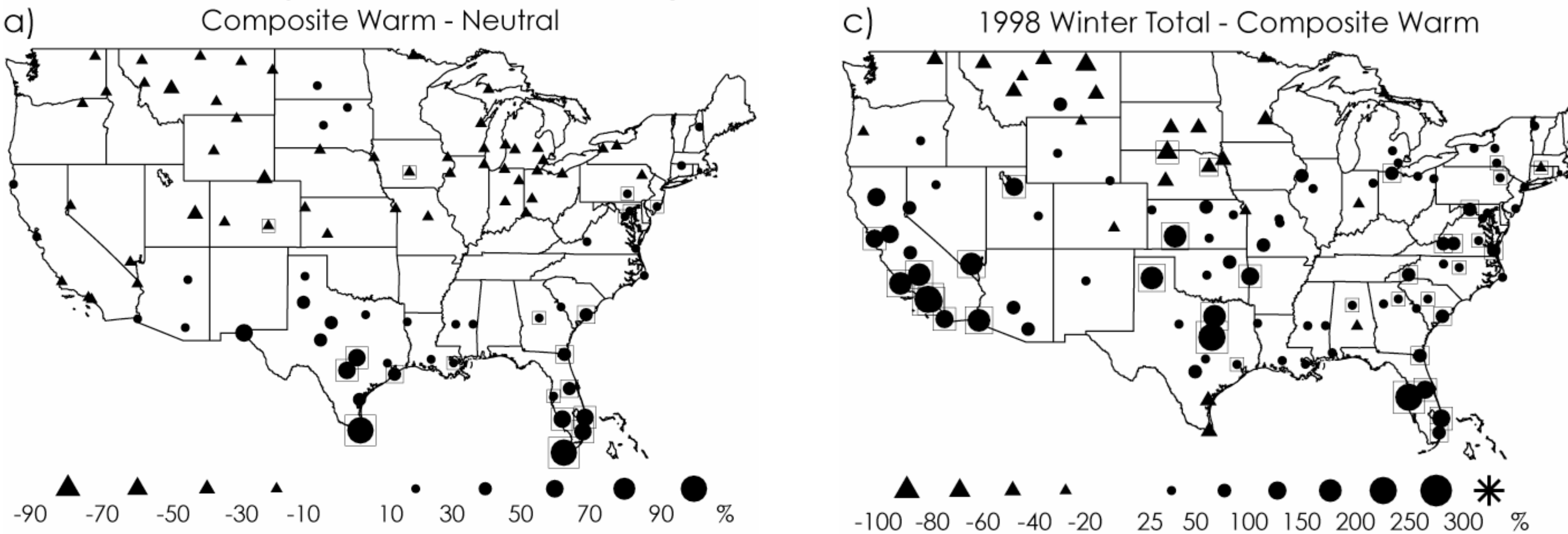
Time Series

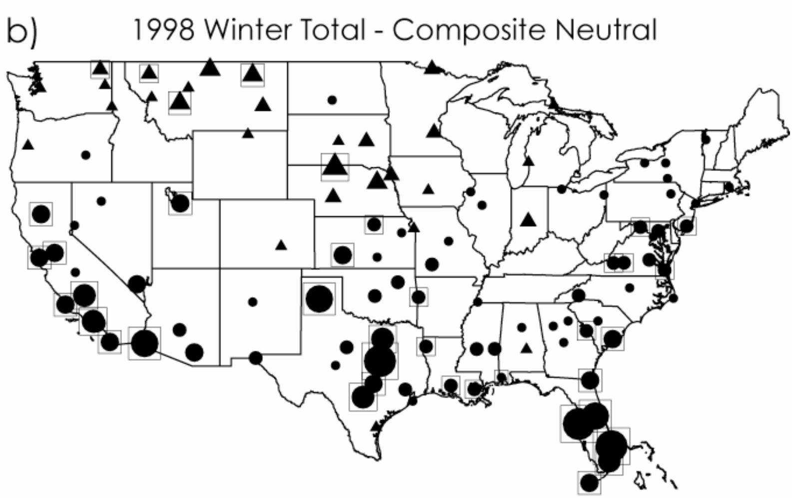


Autocorrelation with different time lags

Figure from [www.neurotraces.com/ scilab/scilab2/node39.html](www.neurotraces.com/scilab/scilab2/node39.html)

Glyph Plots
El Nino Winter Precipitation Anomalies

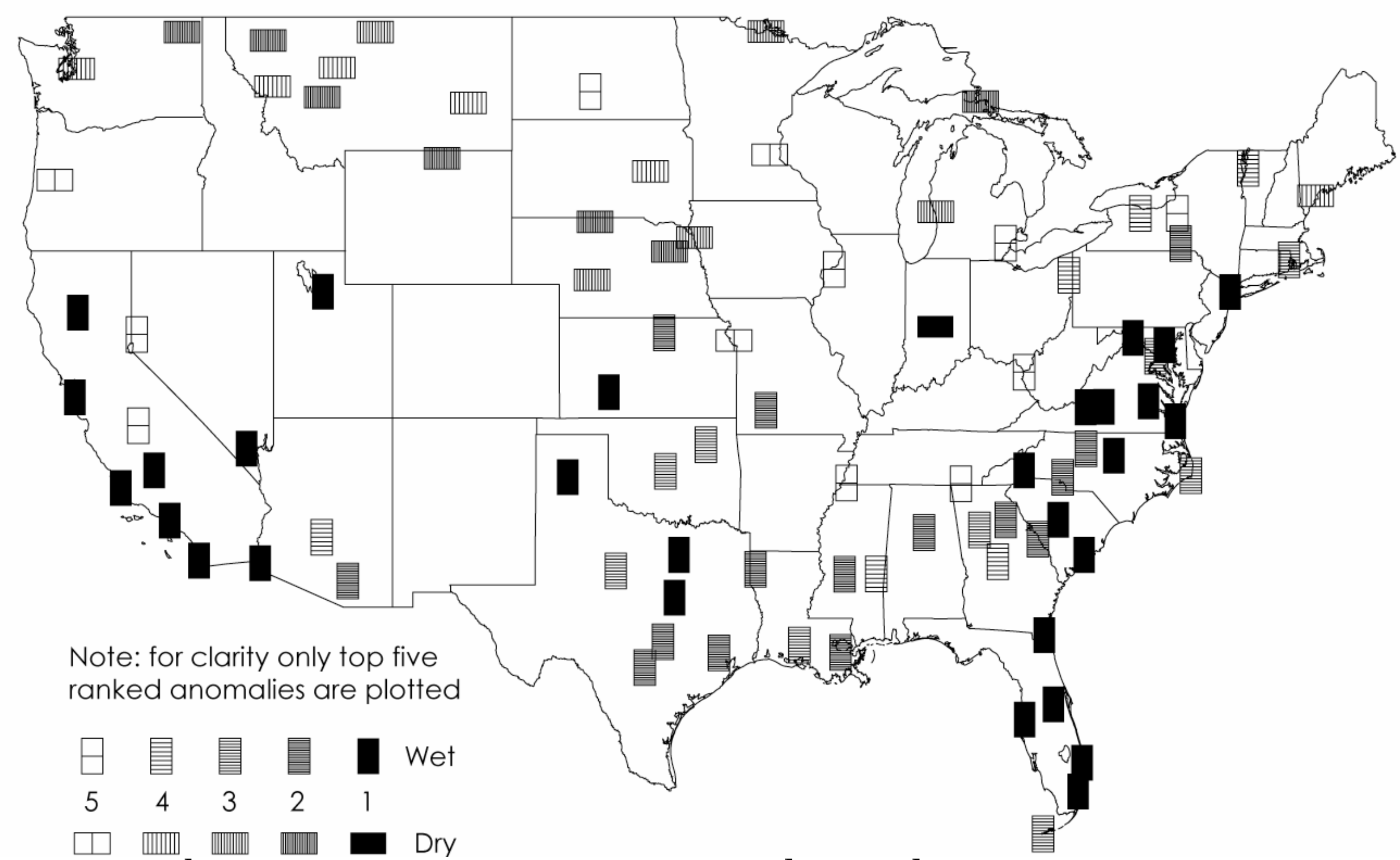




- Data are from the Global Historical Climate Network (GHCN).
- Boxes indicate 1998 station data that falls within the upper or lower 5% of values based on the (b) neutral or (c) warm phase data.

Figure from Smith, Legler, Remigio and O'Brien, JCLIM, 1999

Glyph Plots
1998 Winter Precipitation Anomaly Ranked vs. Nine Historical Warm Phases



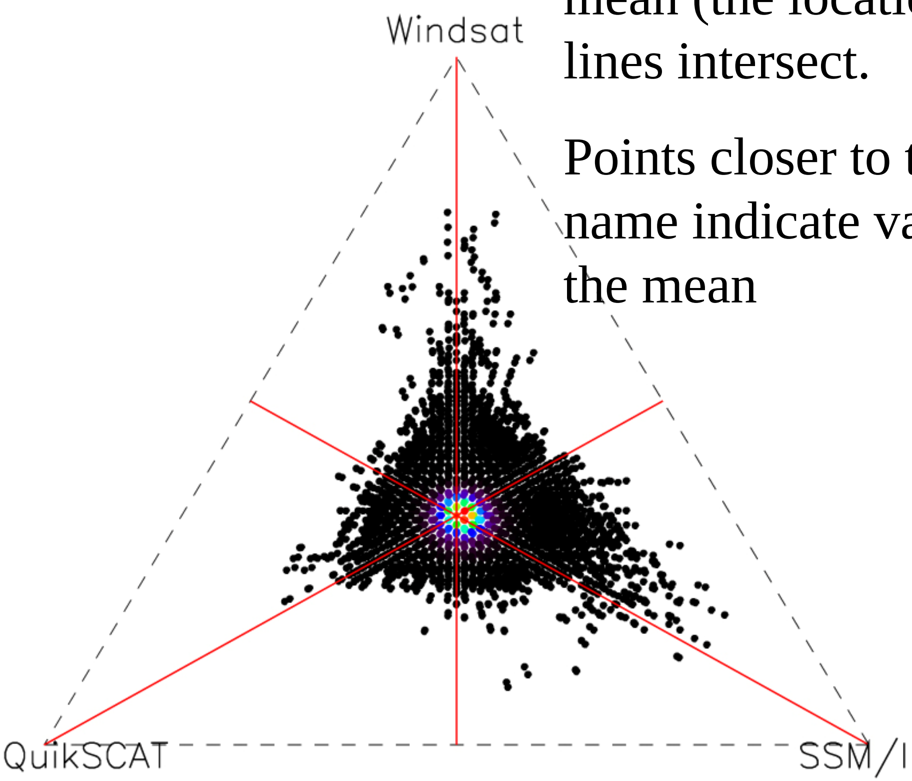
- Warm phase GHCN winter precipitation data and 1999 Winter.
- Boxes indicate rank of the 1999 data.

Figure from Smith, Legler, Remigio and O'Brien, JCLIM, 1999

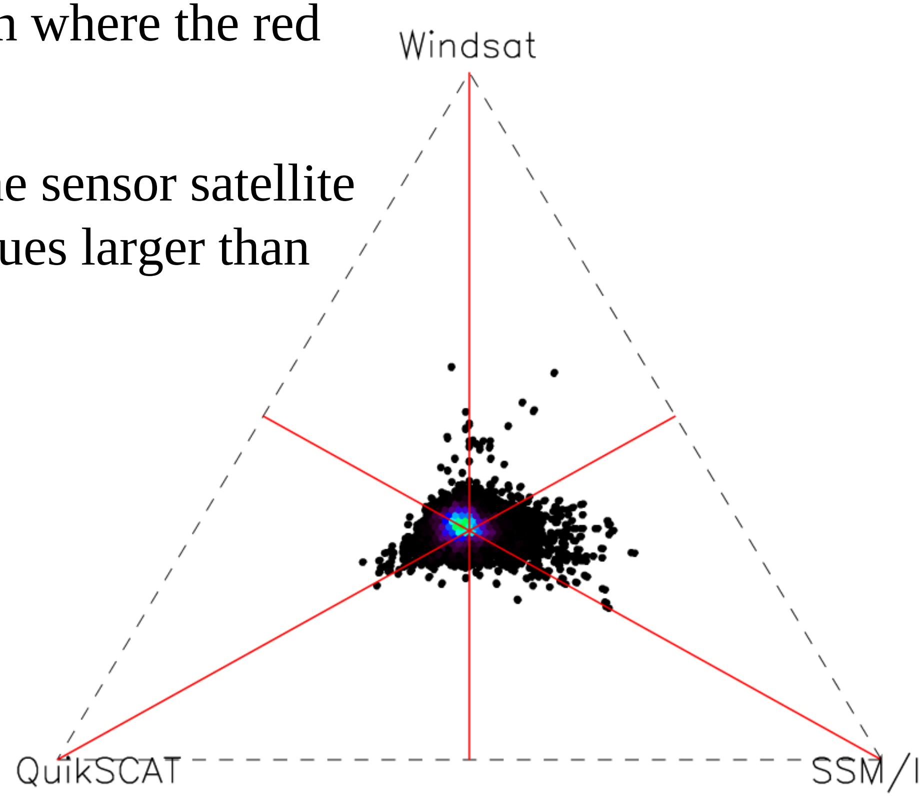
Example Star Plot: Differences from the mean of three co-located satellite observations

The plots show departure form the mean (the location where the red lines intersect.

Points closer to the sensor satellite name indicate values larger than the mean



6.4 – 7.2 m/s

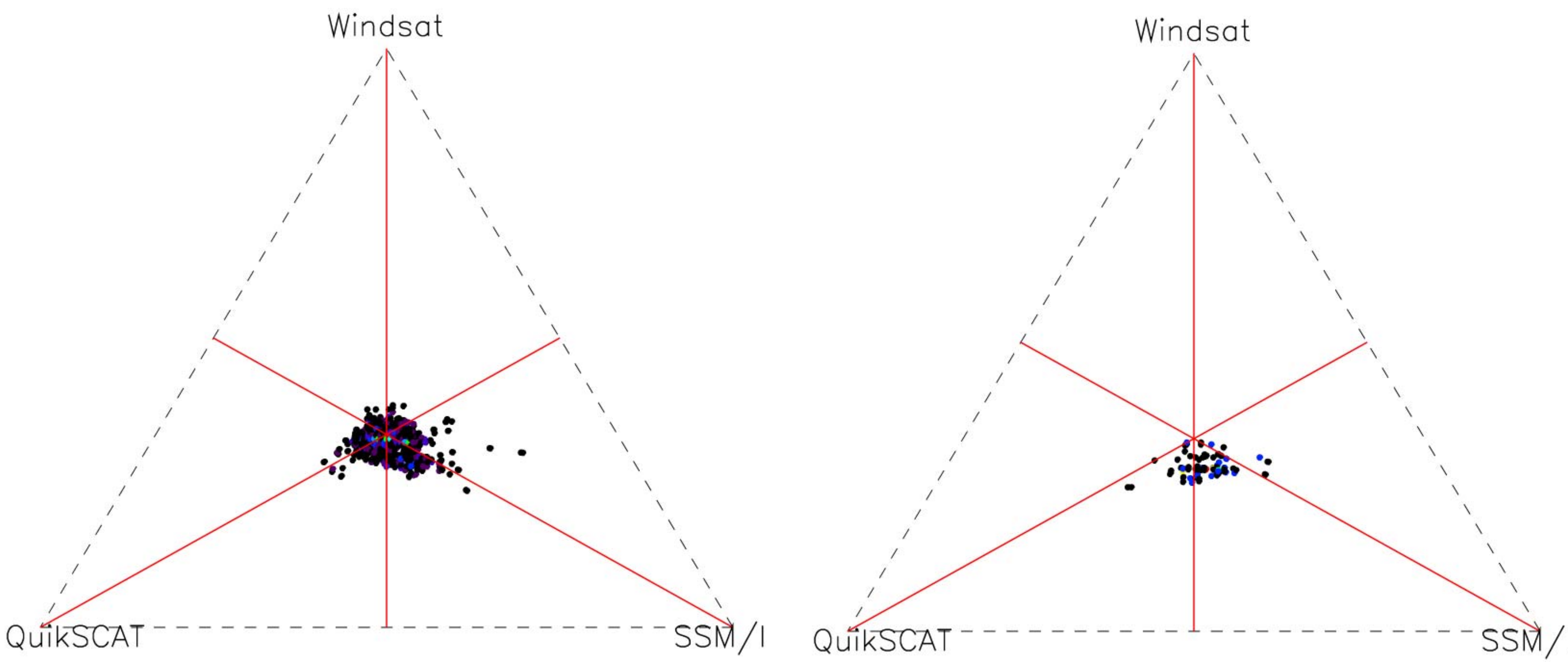


15.2 – 16.0 m/s

Graphics from talk by Mike Freilich and Barry Vanhoff

http://campus.fsu.edu/
bourassa@met.fsu.edu

Example Star Plot: Differences from the mean of three co-located satellite observations



21.6 – 22.4 m/s

25.6 – 26.4 m/s

Graphics from talk by Mike Freilich and Barry Vanhoff

http://campus.fsu.edu/
bourassa@met.fsu.edu

Correlation Maps

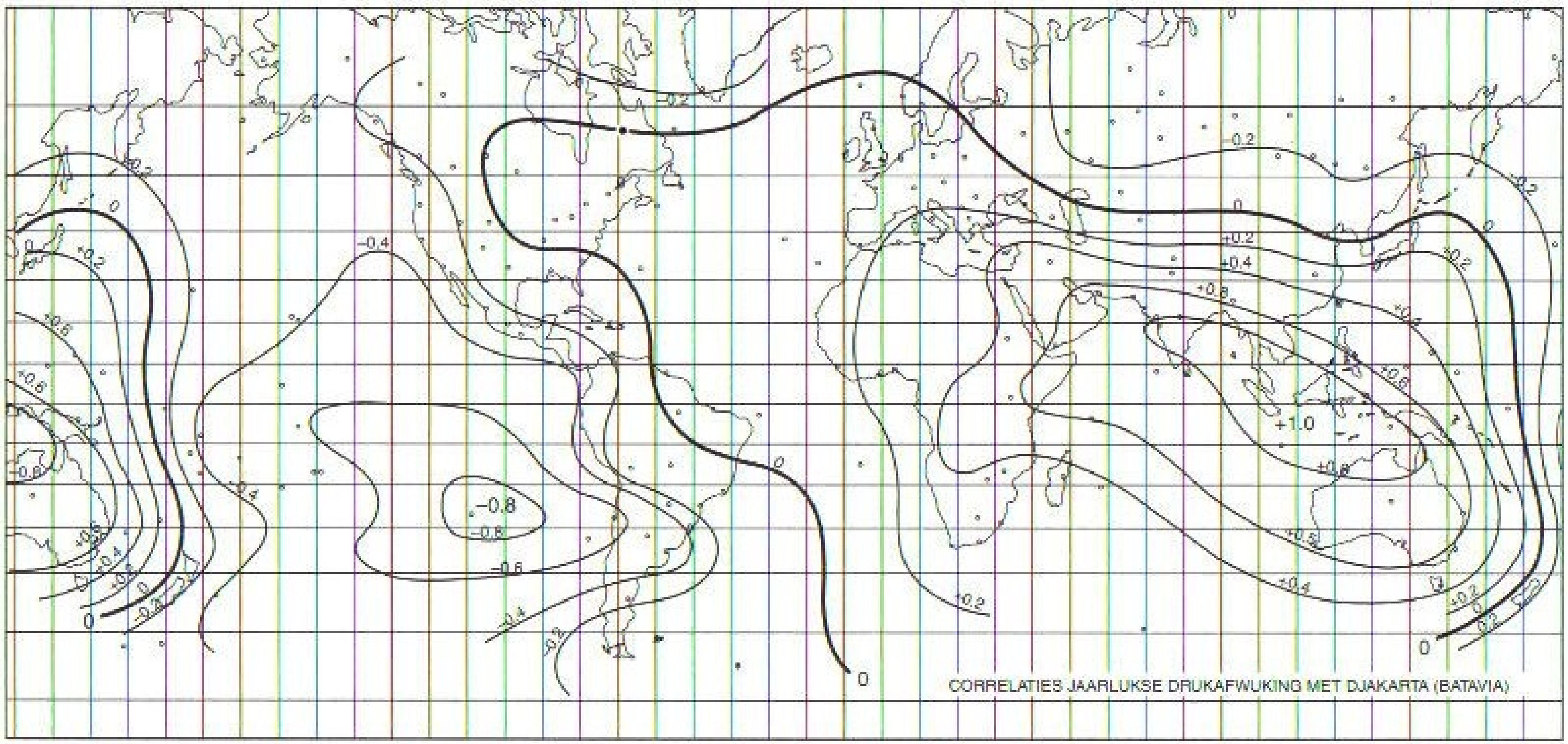
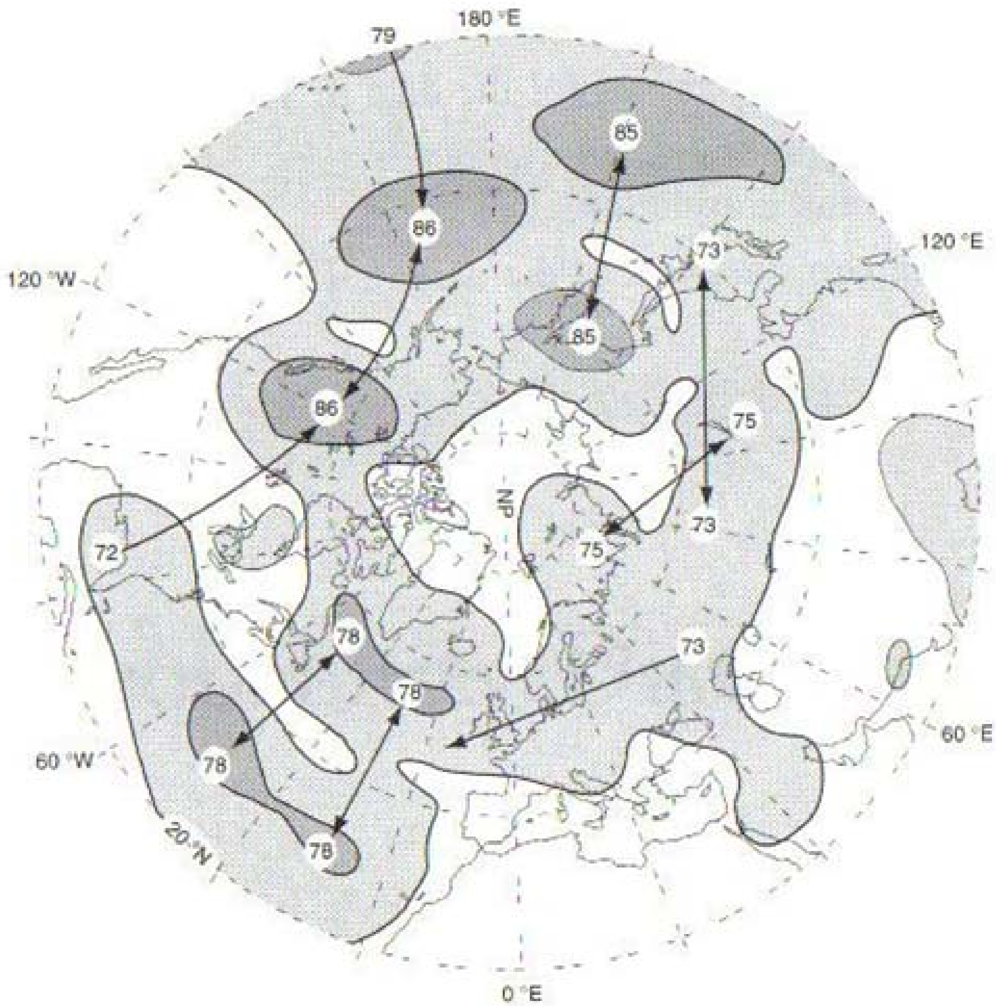


FIGURE 3.27 One-point correlation map of annual surface pressures at locations around the globe with those at Djakarta, Indonesia. The strong negative correlation of −0.8 at Easter Island is related to the El Niño-Southern Oscillation phenomenon. From Bjerknes (1969).

Application of Correlation Maps



- Remote influences (teleconnections) can be identified.

Figure originally from Wallace and Blackmon, 1983