Tues, 4-02-19

9. Correlation

* Lightning talk
* Feedback on A8

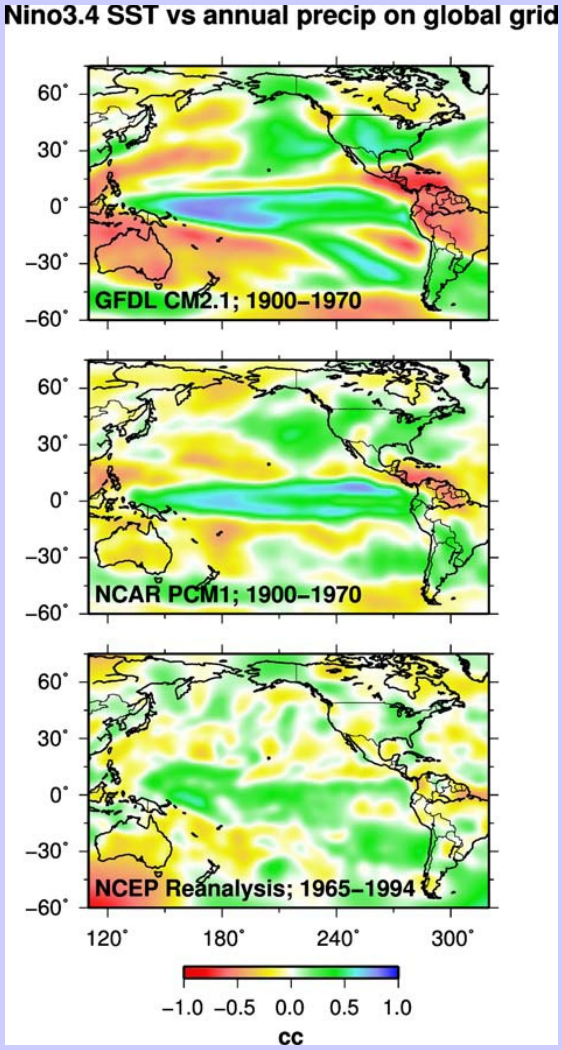
1. **Applications and applications**
2. **Statistical significance**
3. **Temporal stability**

Read notes_9.pdf

A8 Feedback

1. Download A8x.pdf from D2L
2. Automatic points, for running assignment and having uploaded by due time, is already marked in parentheses at top of first page
3. Each assignment has maximum possible 10 points; if you make no deductions, score is 10/10
4. A8x is color coded for points; purple=1; yellow=0.5; blue=0.5
5. Open your copy of the same assignment pdf you uploaded
6. In Acrobat Reader, using “Add text box,” mark in right margin for deductions only, with deduction and segment reference : (eg., -0.5 A); round to tenths in deductions (e.g., no -0.25)
7. At top of your pdf, mark grade like this : 9.5/10
8. If necessary, put any comments at top near the grade
9. Upload your self-graded pdf to folder A8_graded in D2L

Correlation coefficient
example of application



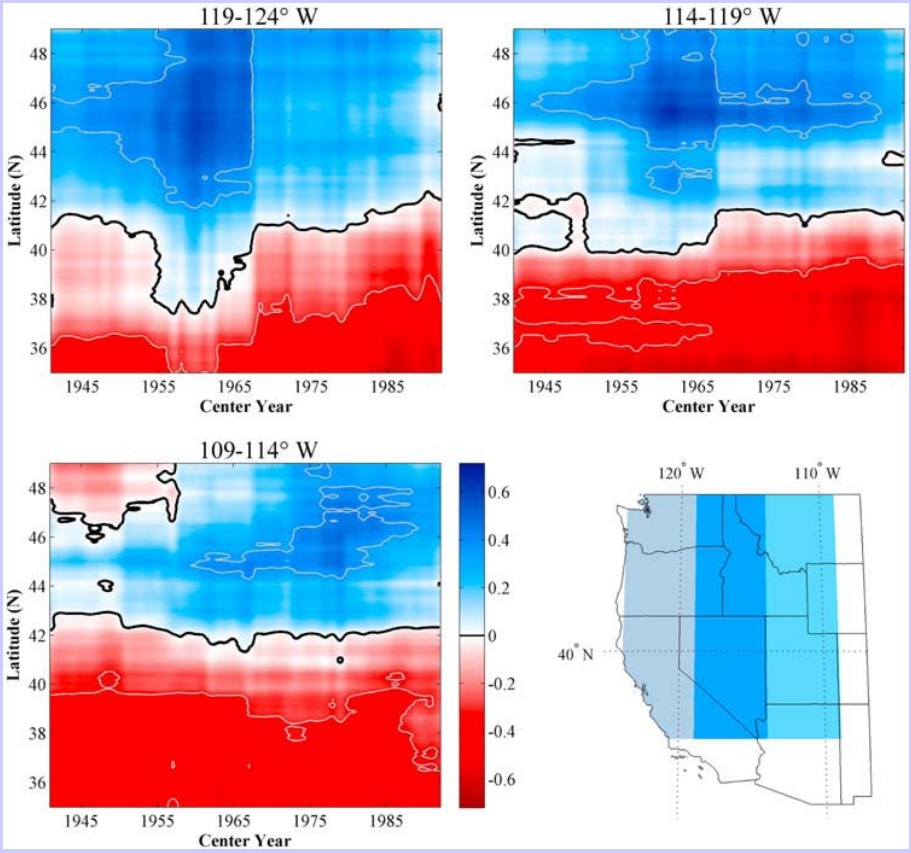
Mapped correlations

Nino3.4 SST with annual modeled precipitation

Spatially dependent response to some stimulus

Cayan et al., 2008, Climate change scenarios for the California region, Climatic Change 87 (Suppl 1):S21–S42

Correlation coefficient
example of application



Sliding-window correlations

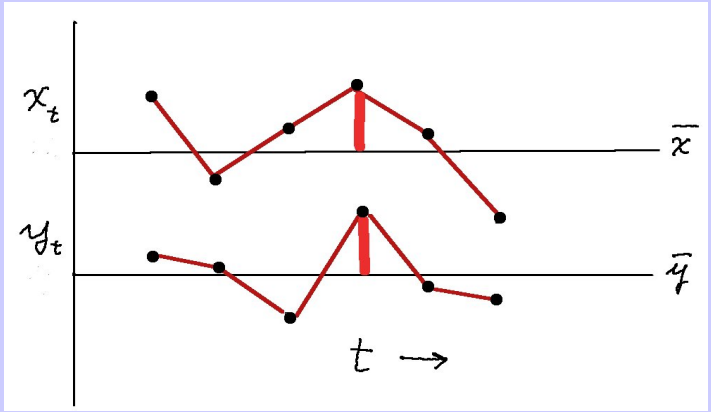
Southern Oscillation Index with Oct-Mar precipitation as function of latitude and time

Temporal shift in dipole transition zone

Wise, E. 2010. Spatiotemporal variability of the precipitation dipoleb transition zone in the western United States. GRL 37, L07706, doi:10.1029/2009GL042193, 2010

Correlation coefficient: equation

1. Time series, length N, and departures from mean



2. "Covariance": average product of departures

$$\mathrm{cov}(x,y) = \frac{1}{N-1}\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y})$$

3. "Correlation": covariance scaled by standard deviations

$$r = \frac{\mathrm{cov}(x,y)}{s_x s_y}, \quad \rightarrow \quad -1 \le r \le 1$$

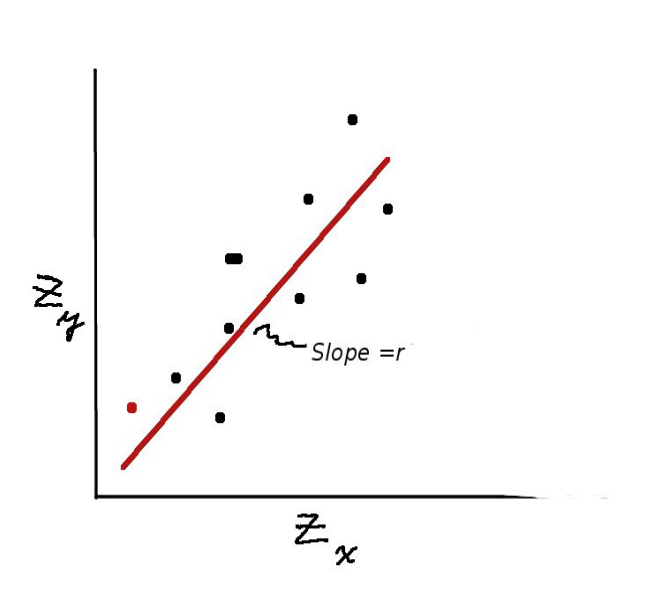
Correlation coefficient – z-score notation

1. Time series as "z-scores"

$$z_{x,t} = \frac{x_t - \bar{x}}{s_x}, \qquad z_{y,t} = \frac{y_t - \bar{y}}{s_y}$$

2. Correlation coefficient

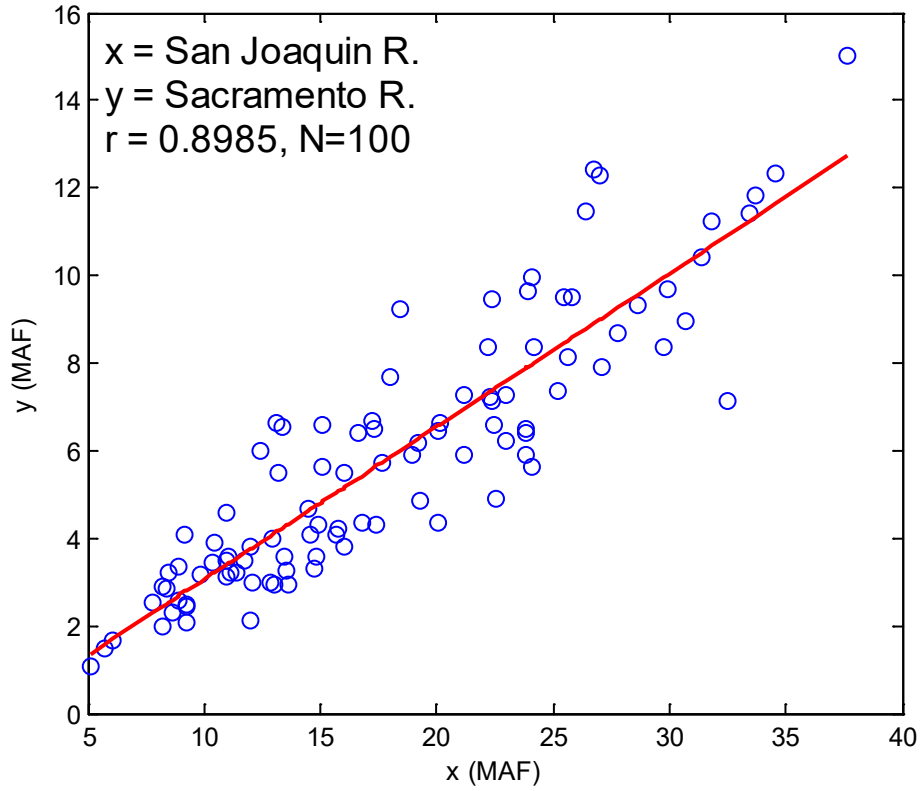
$$r = \frac{1}{N-1}\sum_{t=1}^{N} z_{x,t} z_{y,t}$$



r is invariant of any linear scaling of the time series

Statistical significance of a sample r

Annual River Flows



$$r = b\frac{s_x}{s_y}$$

b = slope of regression line

s_x, s_y = standard deviations of x, y

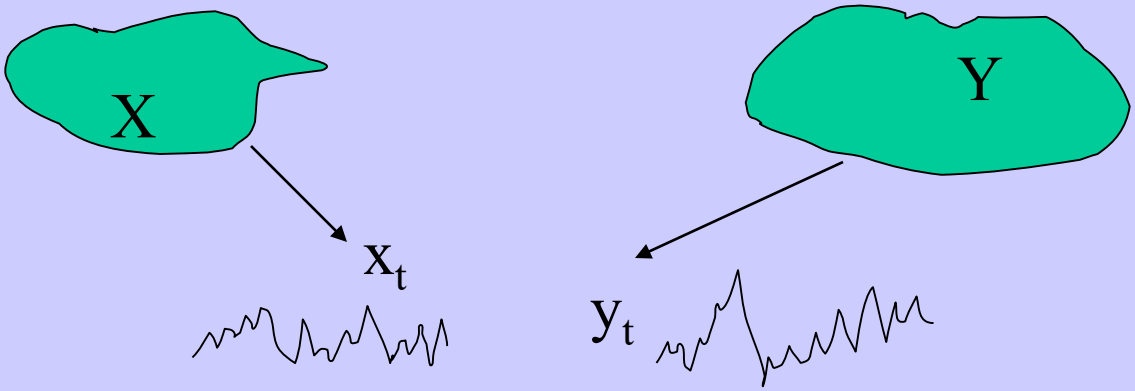
T-test of slope of regression is equivalent to test of significance of correlation

Assume:

1. x,y of sample size N drawn from populations following a bivariate normal distribution
2. Samples are randomly drawn
3. The population correlation is zero

$$T = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$ follows a t-distribution with $N-2$ degrees of freedom

Large Sample Size Approximation
Statistical significance: Testing H0 that ρ=0



Assume:

- **Populations normally distributed**
- **Populations uncorrelated**
- **Pairs of observations drawn at random**
- **Sample size “large”**

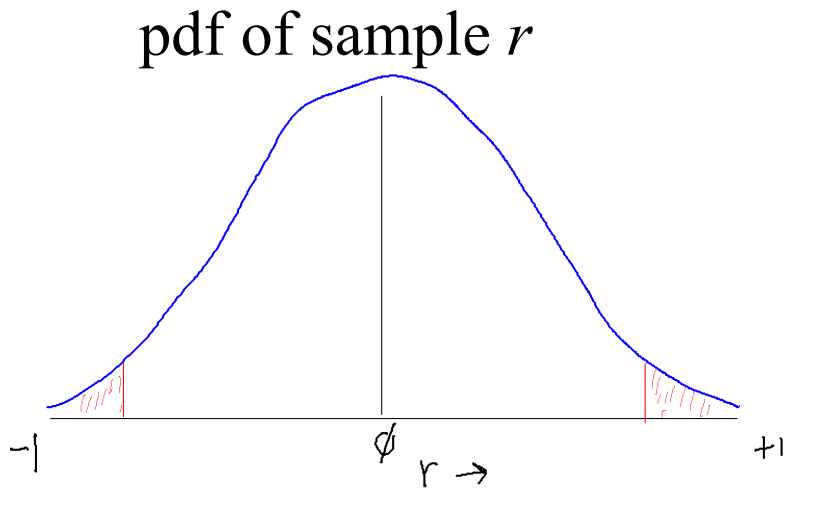
Testing H0 that ρ=0 (continued)

If the assumptions are true, sample correlations are normally distributed with

$$\text{Mean} = 0$$

$$\text{Variance} = 1/(N-2)$$

Distribution of sample r when ρ=0



$$r \sim \mathrm{N}\left(0, \frac{1}{N-2}\right)$$

$$\mathrm{std}(r) = 1/\sqrt{N-2}$$

95% CI: $$r \pm \frac{1.96}{\sqrt{N-2}} \approx r \pm \frac{2}{\sqrt{N}}$$

$\rho =$ population correlation coefficient, X vs Y

$r =$ sample correlation coefficient, x vs y

$$N = 100 \rightarrow r_{.95} = 2/\sqrt{100} = 0.20$$

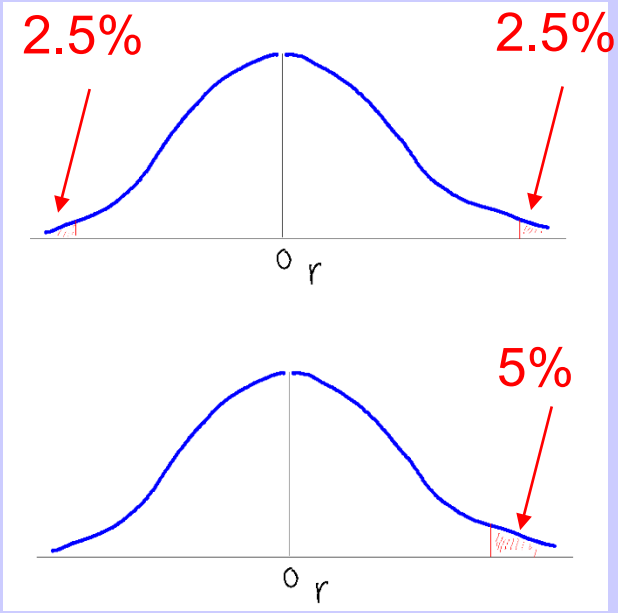
$$N = 10000 \rightarrow r_{.95} = 2/\sqrt{10000} = 0.02$$

Assume

1) X and Y distributed bivariate normal

2) N random samples of the distribution

2-sided vs 1-sided test for significance of a sample *r* when $\rho=0$



$$2\text{ sided} \rightarrow \text{H0: } \rho = 0, \quad \text{H1: } \rho \neq 0$$

$$1\text{ sided} \rightarrow \text{H0: } \rho \leq 0, \quad \text{H1: } \rho > 0$$

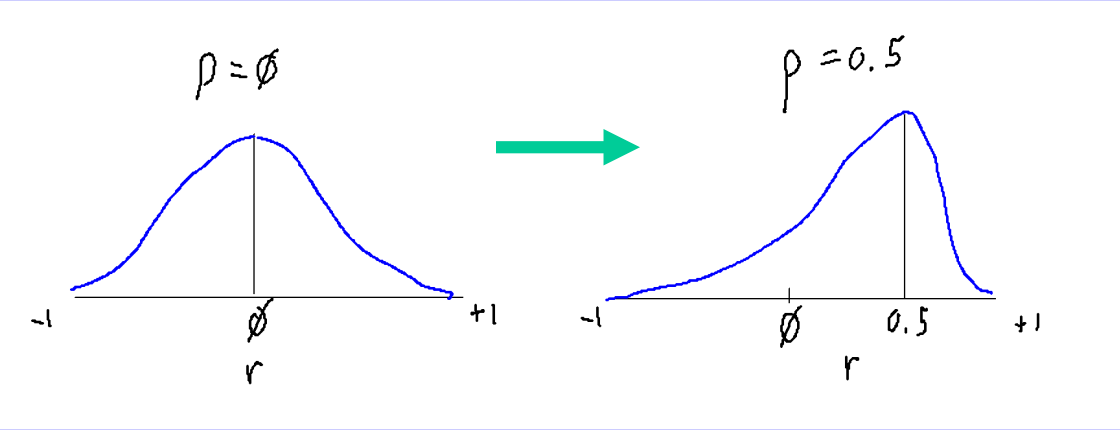
Snowpack depth "affects" tree-growth (2-sided)
Deep snowpack "favors" growth & vice versa (1-sided)

For symmetric distributions (e.g., normal):

The same sample *r* would be judged "more significant"
by one-tailed test than by rwo-tailed test

What if cannot assume ρ=0? Testing significance of difference between two non-zero sample correlations

- Population correlations are not zero
- Distributions of sample correlations are not normal



For example:

- Correlation between climate and tree-ring index is r_1 in last century and r_2 in this century, with $r_2 < r_1$
- Might hypothesize that relationship between climate and tree growth is *weakening*: $\rho_2 < \rho_1$
- Cannot use tests that rely on normal distribution of sample r

… but can use Fisher's Z transform

r : a sample correlation

ρ: population correlation

N : sample size

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$Z \sim \mathrm{N}\left(\mu_z, \sigma_z^2\right), \text{ where}$$

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \text{ and}$$

$$\sigma_z^2 = \frac{1}{N-3}$$

… with two sample correlations, Fisher's Z can be used to test the difference in r's

$$D = z_1 - z_2 \quad \text{is } \mathrm{N}\left(0, \sigma_D^2\right), \quad \text{where}$$

$$\text{variance } \sigma_D^2 = \sigma_{z_1}^2 + \sigma_{z_2}^2$$

$$95\%\ \text{CI for } D \rightarrow \quad D \pm 1.96\sigma_D$$

r_1=0.60, N_1=40 years
r_2=0.70, N_2=60 years

- D not significant at α=0.05
- For significance, would need r_2>0.81

Effective sample size
(in context of testing for significance of correlation)

Recall that effective sample size for persistence in a single time series is

$N' = N\frac{1-r_1}{1+r_1}$, where N is sample length and r_1 is lag-1 autocorrelation

A similar adjustment applies when testing for significance of correlation beween time series x_t and y_t. The adjustment depends on BOTH lag-1 autocorrelations, $r_{1,x}$, and $r_{1,y}$:

$$N' = N\frac{1-r_{1,x}r_{1,y}}{1+r_{1,x}r_{1,y}}$$

If either series is not autocorrelated, no adjustment is made. The geosa9 script does not adjust unless both series have significant positive lag-1 autocorrelation (α=0.05, 1 tailed)

Effective sample size-- example

$$r_{1,x} = 0.6$$

$$r_{1,y} = 0.5$$

$$N = 300$$

$$N' = 300\frac{1-0.30}{1+0.30} = 300\left(0.538\right) \approx 162$$

Effect on threshold for significance at $\alpha = 0.05$

Without adjusment: $r_c = 0 \pm \frac{1.96}{\sqrt{N-2}} = 0 \pm \frac{1.96}{\sqrt{298}} = 0 \pm 0.11$

With adjusment: $r_c = 0 \pm \frac{1.96}{\sqrt{N'-2}} = 0 \pm \frac{1.96}{\sqrt{160}} = 0 \pm 0.15$

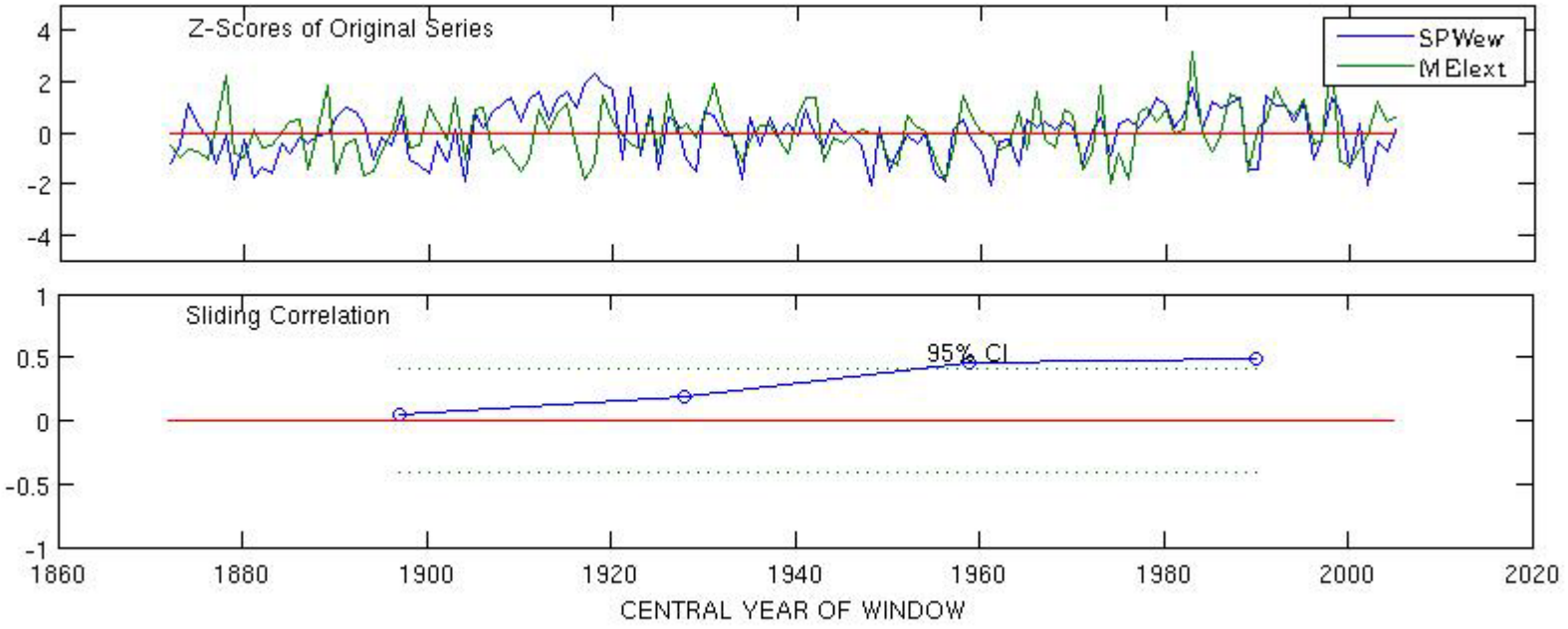
Temporal stability of relationship
windowed correlations

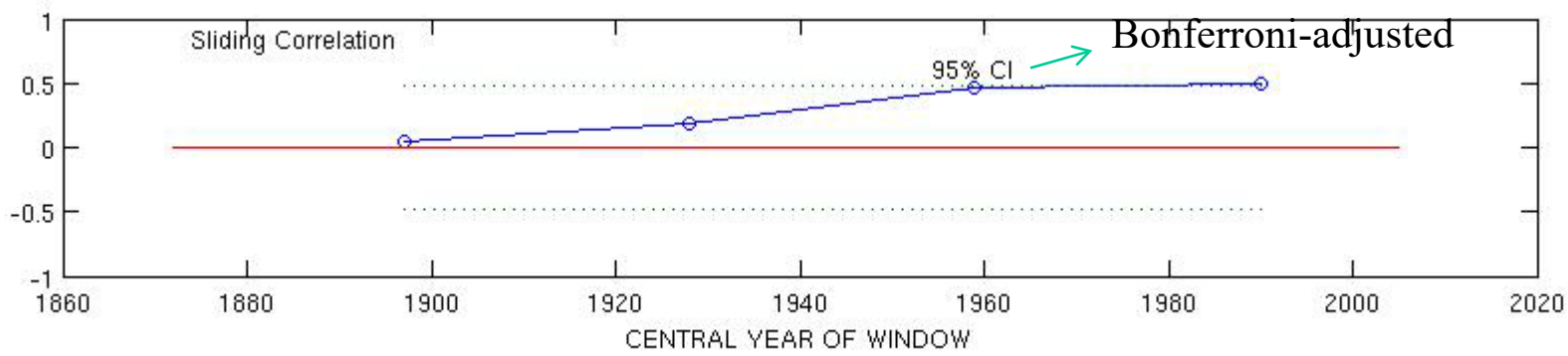
- Compute correlations in a sliding time window
- Plot correlations to summarize temporal variability of relationship between the series
- Considerations:
    1. Persistence – as affects significance of r
    2. Window size – affects critical level of r
    3. Changing means and standard deviations
    4. Multiple comparisons: increased chance of “bogus” significance

Example:

- White-fir earlywood index, San Pedro Martir
- Multivariate ENSO index
- 1982-2005
- 31-year centered window

Example:


Z-Scores of Original Series
SPWew
MElext
Sliding Correlation
95% CI
CENTRAL YEAR OF WINDOW


Sliding Correlation
95% CI
Bonferroni-adjusted
CENTRAL YEAR OF WINDOW

"Universal" null hypothesis
(in context of multiple comparisons)

"All of the evaluated correlation coefficients are zero"

If the universal null hypothesis is what is of interest, the Bonferroni adjustment of alpha-level is appropriate

α' = alpha-level for an individual test

k = number of tests evaluated

$\alpha = \alpha' / k$ appropriate alpha-level for testing universal H0

Contrasting hypotheses

Multiple tests evaluated individually

$$H0: \quad \rho = 0$$

$$H1: \quad \rho \neq 0$$

Multiple tests evaluated jointly

$$H0: \quad \text{all } \rho = 0$$

$$H1: \quad \text{at least one } \rho \neq 0$$

Universal
Null hypothesis