Assignments/Dates

- Odds Are It's Wrong. Due Feb 22
  - http://www.sciencenews.org/view/feature/id/57091/title/Odds_Are,_Its_Wrong
  - Read and summarize issues about significance testing in a few paragraphs
- Chapter 4 notes due Feb 24
- Exam March 1.

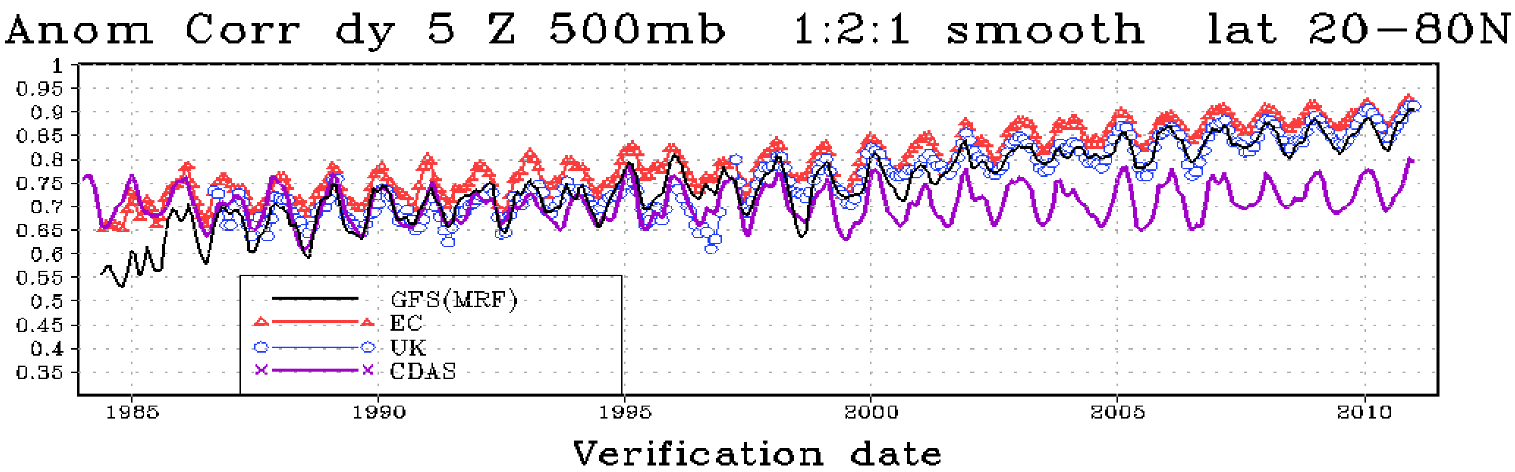
Correlating Maps Rather Than Time Series

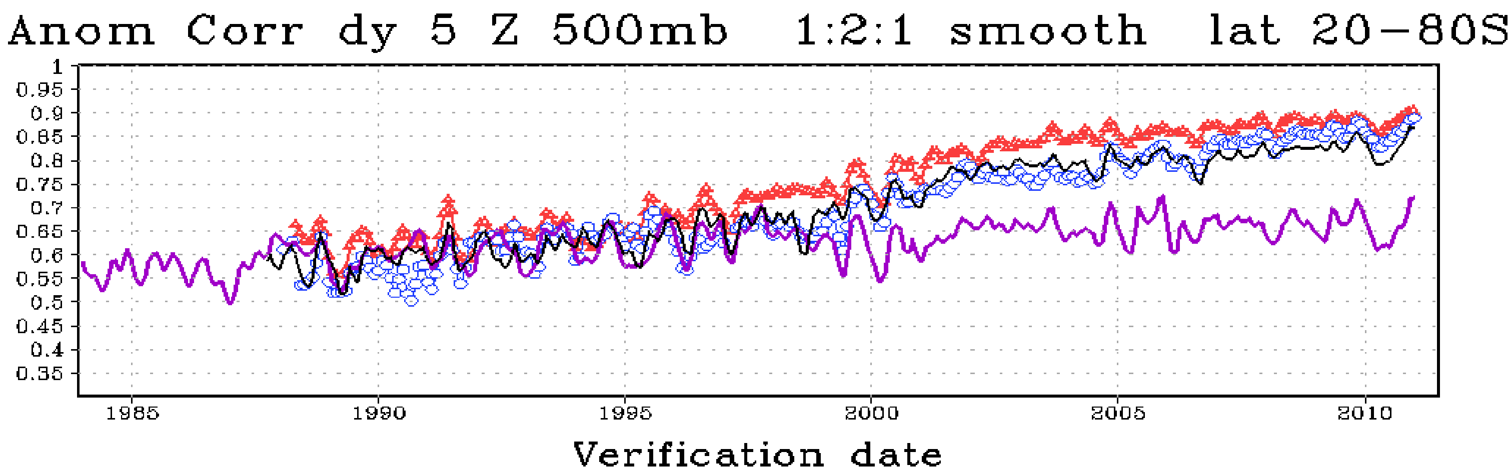
$$\hat{\vec{X}} = \begin{bmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \cdots & \hat{x}_{1,18} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \cdots & \hat{x}_{2,18} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{x}_{7,1} & \hat{x}_{7,2} & \cdots & \hat{x}_{7,18} \end{bmatrix}$$

- Comparing variability in one year over 7 locations to the variability in all of the other n= 18 years

$$\vec{S} = \hat{\vec{X}}^{*T} \hat{\vec{X}}^{*} \;/\, 7$$

Comparing Forecast Anomaly Maps to Analyses





ANOVA and significance test of linear regression

- ANOVA- analysis of variance
- how much confidence in results from linear regression?
- Common practice: assume when $r > 0.5$ or 0.6, then have at least a practical and useful association between the two variables, if a large sample
- For environmental fields with a large number of degrees of freedom, it is possible to have a linear correlation between two variates be as low as 0.1 and still potentially be judged to be statistically significant
- Such low correlation values may not have any practical significance to estimate one variable from the other
- However, they may help point out a physical relationship between the two variables that was unknown before
- The data may then be transformed, filtered or combined with other data to develop some predictive relationship

Describing the amount of variance explained by a linear relationship

$$s_y{}^2 = b^2 s_x{}^2 + \overline{e_i^2} \qquad n s_y{}^2 = n b^2 s_x{}^2 + n\overline{e_i^2}$$

- Sum of squares form:
- Total variance = explained variance + unexplained variance

ANOVA Table- Regression Form

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST=n s_y^2	n-1	n s_y^2 /(n-1)	
Regression	SSR=n $b^2 s_x{}^2$	1	MSR=n $b^2 s_x{}^2$ /1	(n-2) $b^2 s_x{}^2 /(s_y{}^2 - b^2 s_x{}^2)$
Error	SSE=n ($s_y{}^2 - b^2 s_x{}^2$)	n-2	MSE= n ($s_y{}^2 - b^2 s_x{}^2$)/(n-2)	

ANOVA

- summarize whether the variance of variable y explained by variable x is large in terms of three measures:
  - mean squared error of the regression (MSE),
  - variance explained by the regression (MSR),
  - and the F ratio that is assumed to have a known parametric form.
- We want:

1. the scatter around the line of best fit to be small, i.e., that SSE and MSE are small
2. the percent variance explained by the regression to be large (or MSR large)
3. F ratio is large, which is the ratio of the explained variance to that of the error

Degrees of Freedom

Two degrees of freedom are used up from the entire sample:

- mean value of y
- regression coefficient (b)

When looking at some statistical sources, be aware that the MSR should always be greater than the MSE unless n is small and the linear correlation is small as well.

The parametric distribution of F is determined entirely by the degrees of freedom of the larger MS (the regression) and the degrees of freedom of the smaller MS (the error).

F Test of correlation coefficient

ANOVA Table- Correlation Form

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST= n	n-1	n/(n-1)	
Regression	SSR=n r^2	1	MSR=n r^2 /1	(n-2)r ^{2} / $(1-r^2)$
Error	SSE=n (1- r^2)	n-2	MSE= n (1-r^2)/(n-2)	

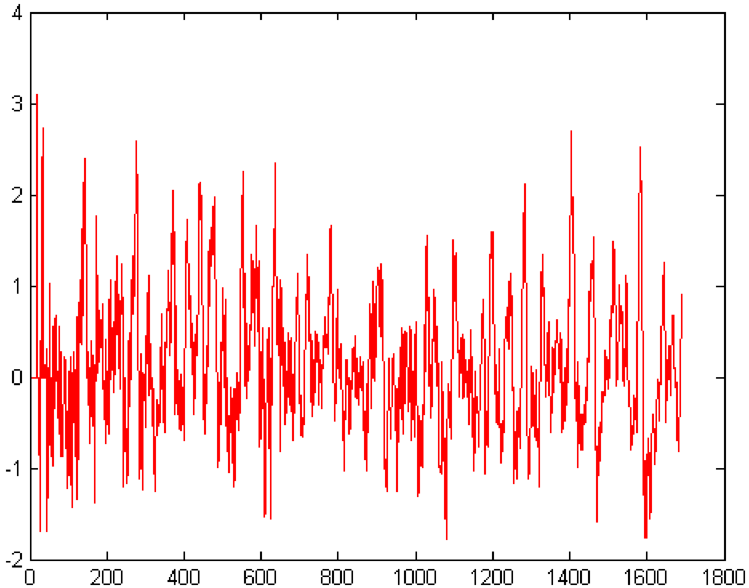
- Signal: explained variance * (n-2)
- Noise: (1-r^2)

anova.m

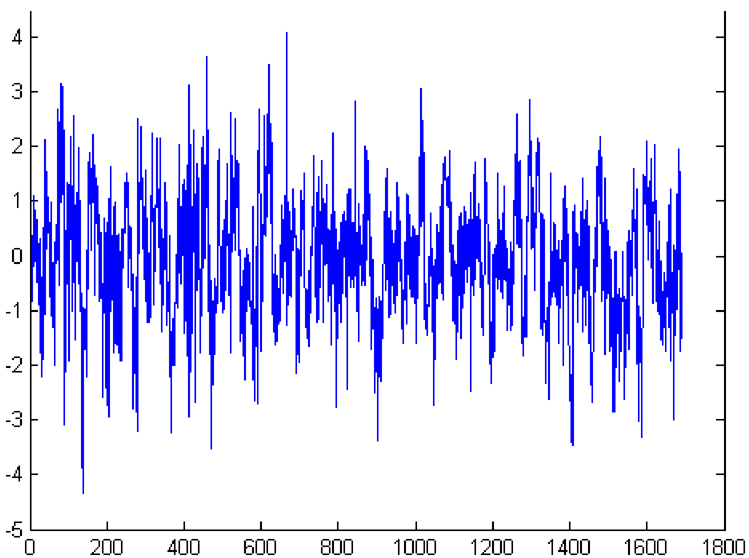
- In course matlab code
- SST and SOI data files in data subdirectory

Equatorial SST vs SOI Index

SST



r =
-0.51



SOI

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	32.15	1	32.1536	35.15	3.35368e-009
Error	3093.39	3382	0.9147		
Total	3125.54	3383			

F value really big, prob of this happening by chance is small

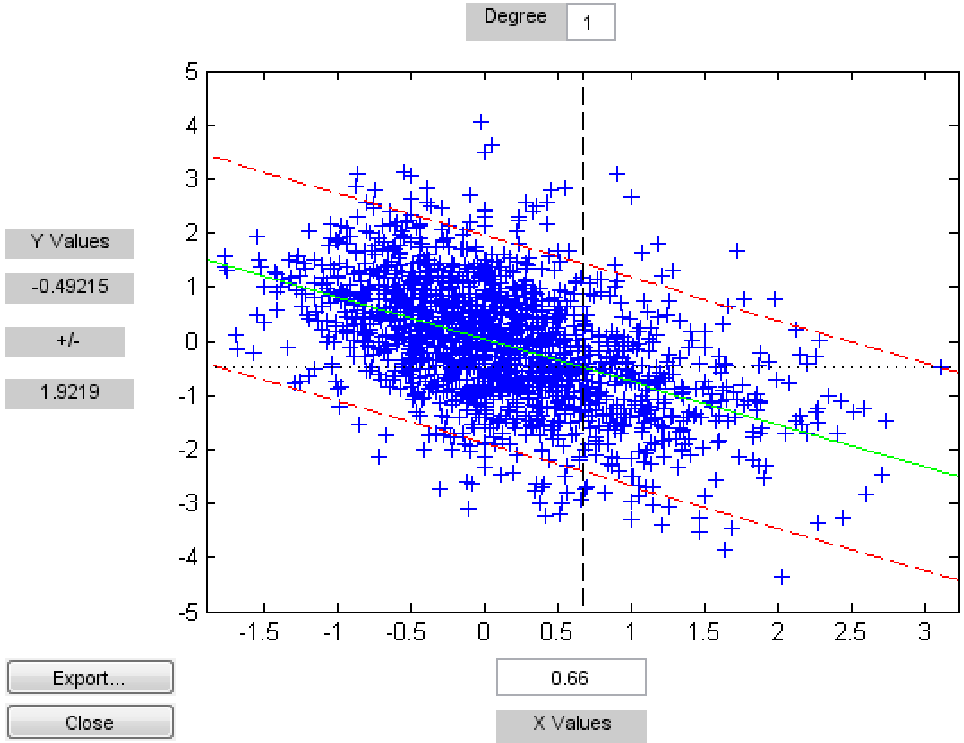
MATLAB ANOVA1: difference of means

- the variability of the data in X into two parts:
  - Variability due to the differences among the column means (variability between groups)
  - Variability due to the differences between the data in each column and the column mean (variability within groups)
- The ANOVA table in matlab has six columns
  1. source of the variability.
  2. Sum of Squares (SS) due to each source
  3. the degrees of freedom (df) associated with each source
  4. Mean Squares (MS) for each source, which is the ratio SS/df
  5. F statistic, which is the ratio of the MS's.
  6. The sixth shows the p-value, which is derived from the cdf of F

As F increases, the p-value decreases. If a significance level of .01 is chosen, then we want the p value to be less than .01 , which it certainly is in this case.

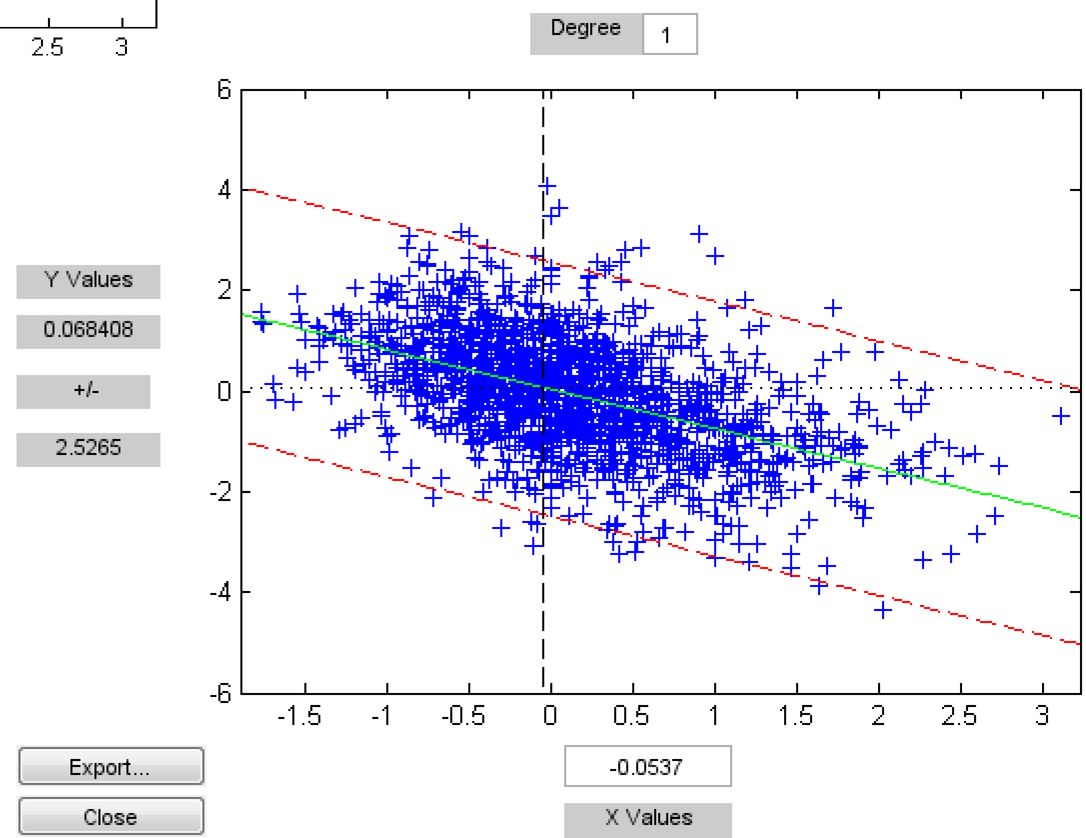
SST vs. SOI

- Anova table grossly overestimates the number of degrees of freedom since there are values each month and both indices have large persistence.
- Assume only 1 value every year is completely independent of one another: 141 degrees of freedom (don't have to subtract two more, because already done).
- Since r=-.51, then F value is 141 * .25/.75 = 47
  - degrees of freedom of the greater MS = 1
  - degrees of freedom of the lesser MS = 141
- matlab command f = finv(.99,1,141) = 6.8
  - value that 99% of the samples from a F parametric distribution should be less than for the specified numbers of degrees of freedom.
  - probability of the correlation between the SST and SOI indices occurring randomly is much less than 1%


Degree
1
Y Values
-0.49215
+/-
1.9219
Export...
Close
0.66
X Values

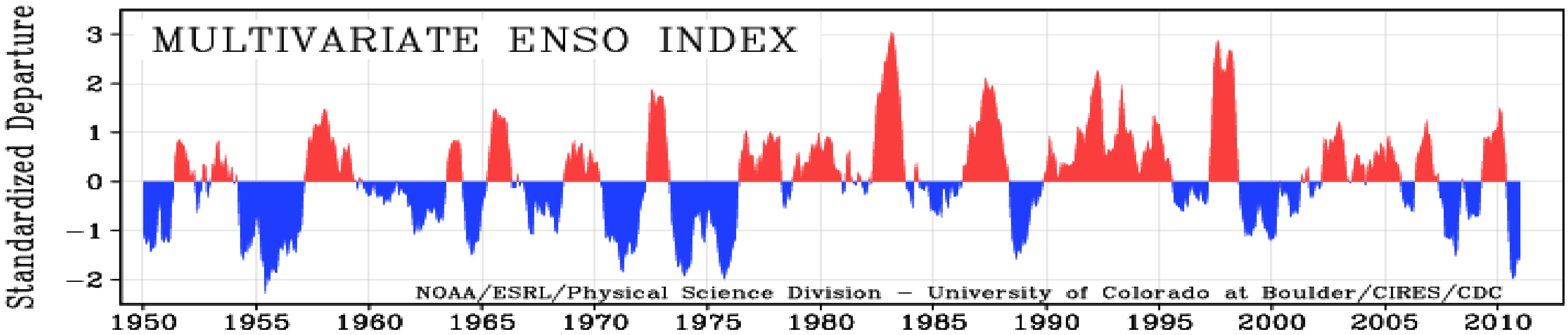
P=0.05

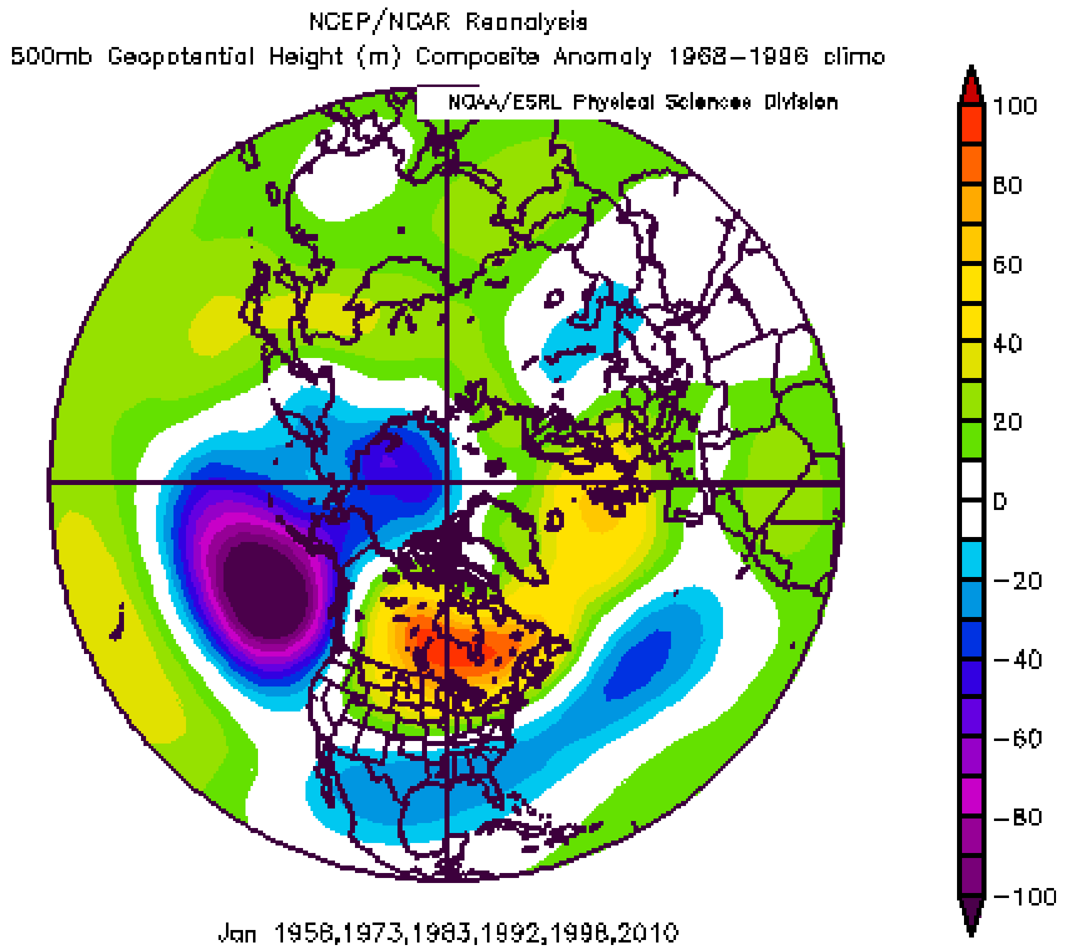
P=0.01


Degree
1
Y Values
0.068408
+/-
2.5265
Export...
Close
-0.0537
X Values

Compositing (Superposed Epoch)

- Identify common characteristics of a sample of events
- Simplest- average conditions before, during, and after some “rare” event
- Has an advantage over linear correlation since no linear assumption necessary
- Limitation- to what extent does sample mean used in composite differ from population?
- Day composites: http://www.cdc.noaa.gov/Composites/Day/
- Monthly/seasonal composites: http://www.cdc.noaa.gov/cgi-bin/data/composites/printpage.pl


MULTIVARIATE ENSO INDEX
Standardized Departure
3
2
1
0
−1
−2
NOAA/ESRL/Physical Science Division − University of Colorado at Boulder/CIRES/CDC
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010

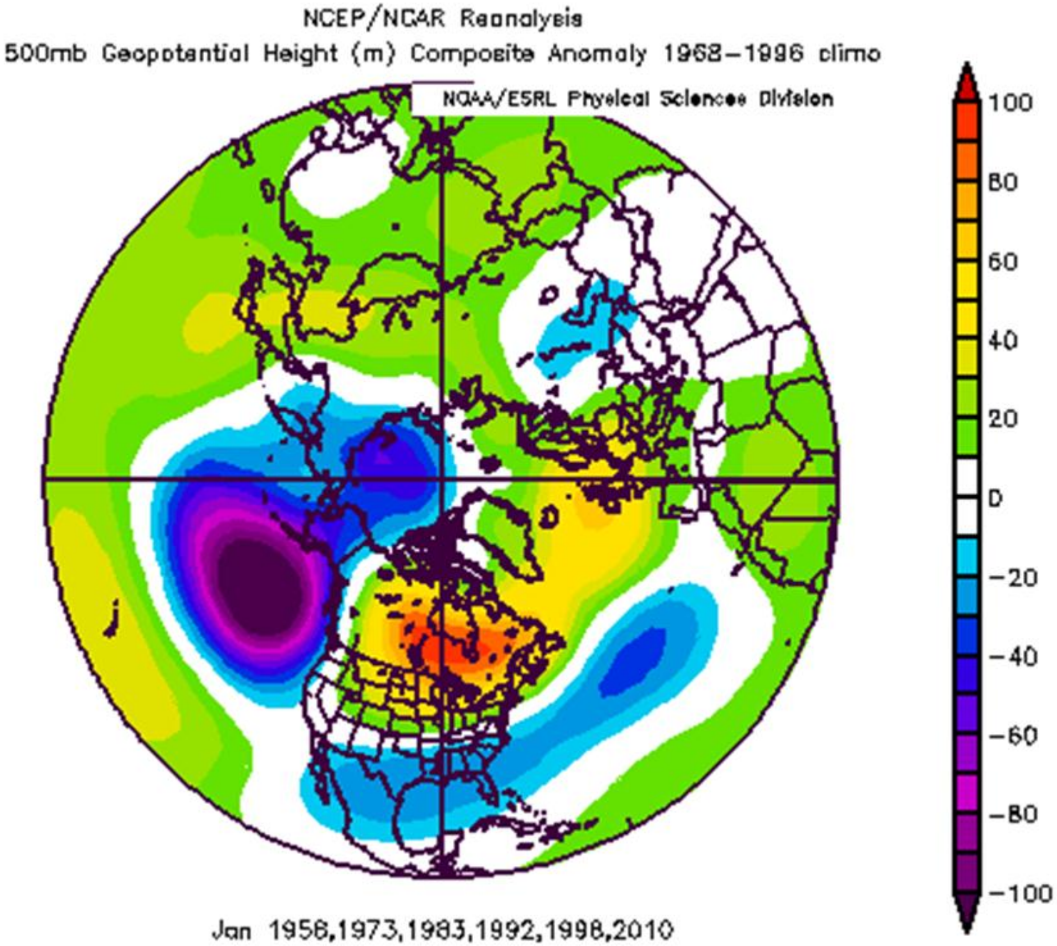

NCEP/NCAR Reanalysis
500mb Geopotential Height (m) Composite Anomaly 1968−1996 climo
NOAA/ESRL Physical Sciences Division
100
80
60
40
20
0
−20
−40
−60
−80
−100
Jan 1956,1973,1983,1992,1998,2010

Compositing Steps

- Select the basis for compositing: why are you doing it? Physical reasoning hopefully?
- Define the categories on which you define the events: above, below normal? Or …?
- Compute the means and statistics for each category (minimum is standard deviation)
- Organize and display the results
- Validate the results:
  - Significance test? t test is the bare minimum to do
  - Reproduce in an independent sample?
  - Are the results sensible in space and time?
  - Is it consistent with theory?

How many spatial degrees of freedom?

- Count the anomaly blobs

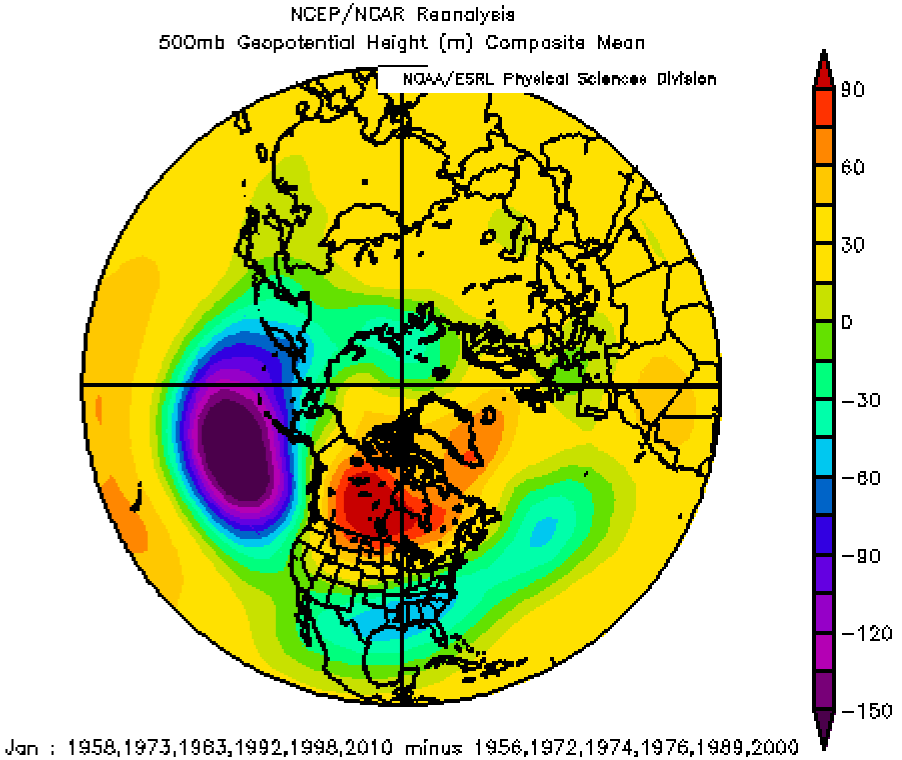


Don’t overdo it…

- Was there a reason a priori to expect the relationship?
- How arbitrary was the choice for defining the composites?
- How subjective and biased was your analysis? Did you tweak your approach to get better results?
- Do the results make sense?
- Are there simpler explanations possible?

Composite Difference Between Two Samples

- High MEI January’s vs. Low MEI January’s



Significance test of difference between two sample means

- common compositing approach is to contrast circulation features associated with two extremes of an index: wet (dry) years in Utah precipitation or El Nino/La Nina seasons.
- Sample means can be computed from the same population or completely different batches of data
- An appropriate null hypothesis is that the population means are the same.
- 2-tailed test IF looking at both positive and negative differences.
- $$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(n_1 s_1{}^2 + n_2 s_2{}^2)(1/n_1 + 1/n_2)/(n_1 + n_2 - 2)}$$
- Signal is the difference between the two sample means with degrees of freedom n_1 and $n_{2,}$
- s_1 and s_2 are sample standard deviations.
- Don't use matlab command ***ttest2:*** can be used only in the situation where the two sample standard deviations are the same