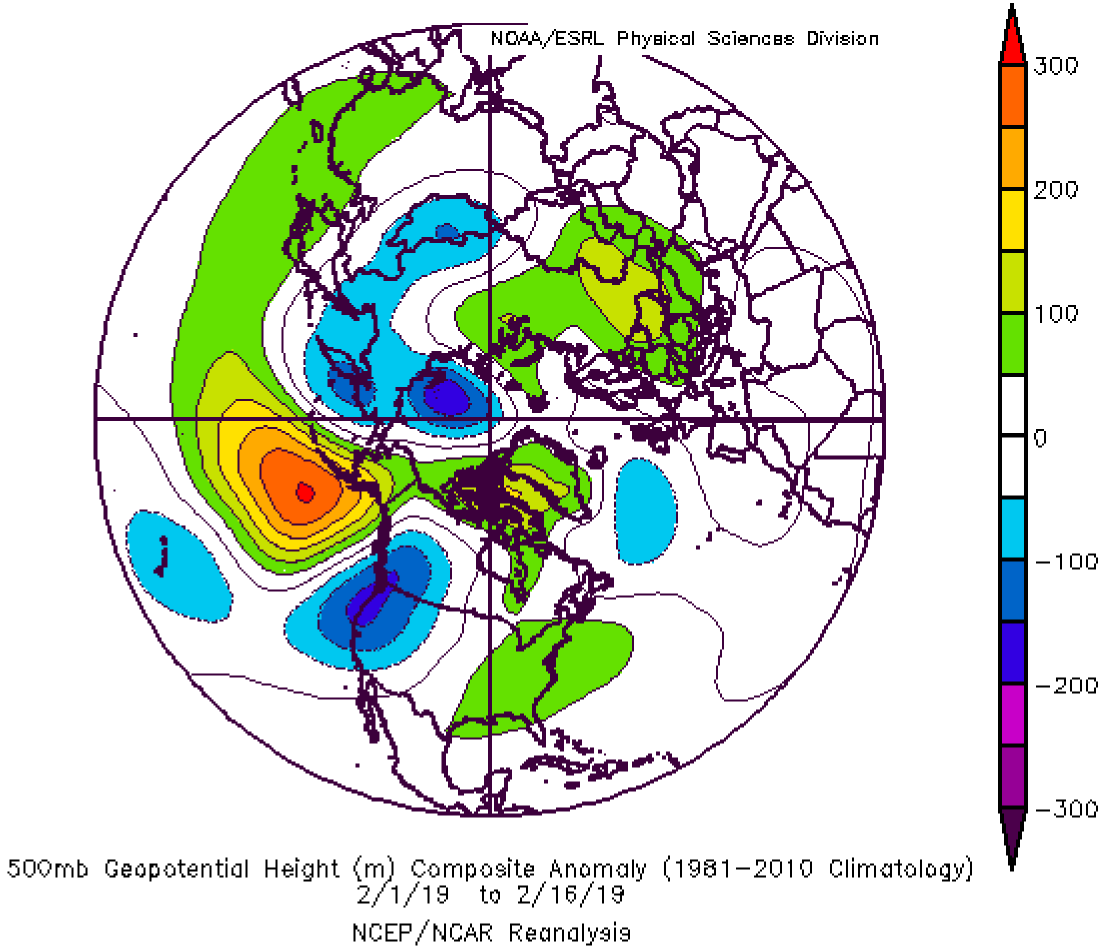
Compositing (Superposed Epoch)

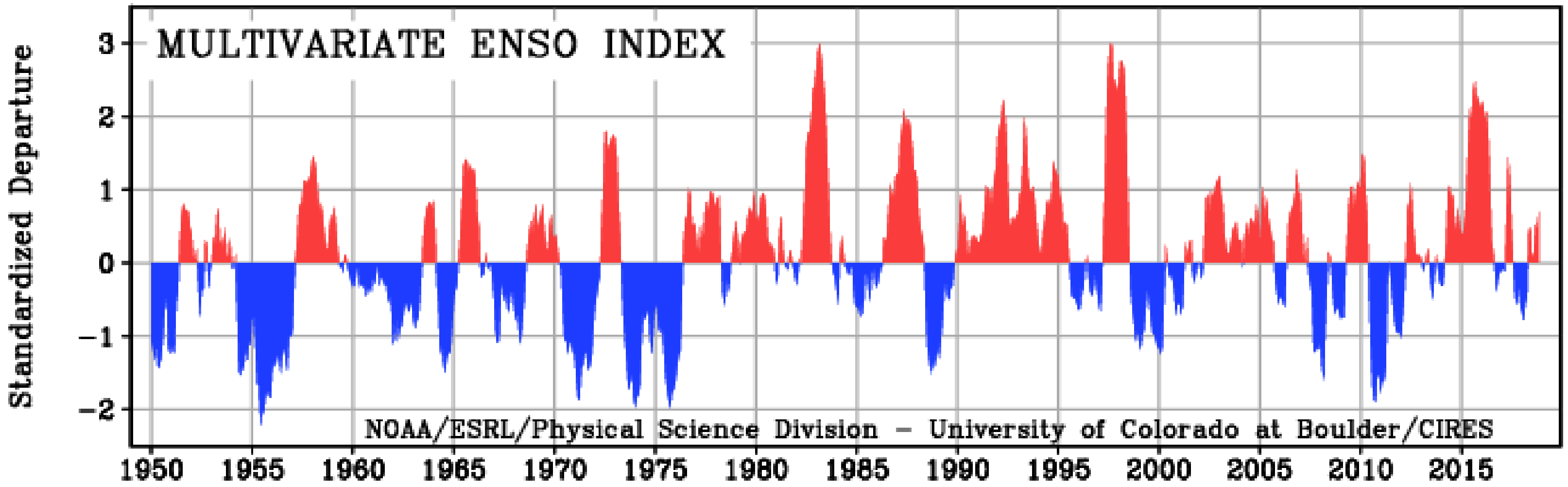
- Identify common characteristics of a sample of events
- Simplest- average conditions before, during, and after some “rare” event
- Has an advantage over linear correlation since no linear assumption necessary
- Limitation- to what extent does sample mean used in composite differ from population?
-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

Composite 500 hPa height anomaly over the first 16 days

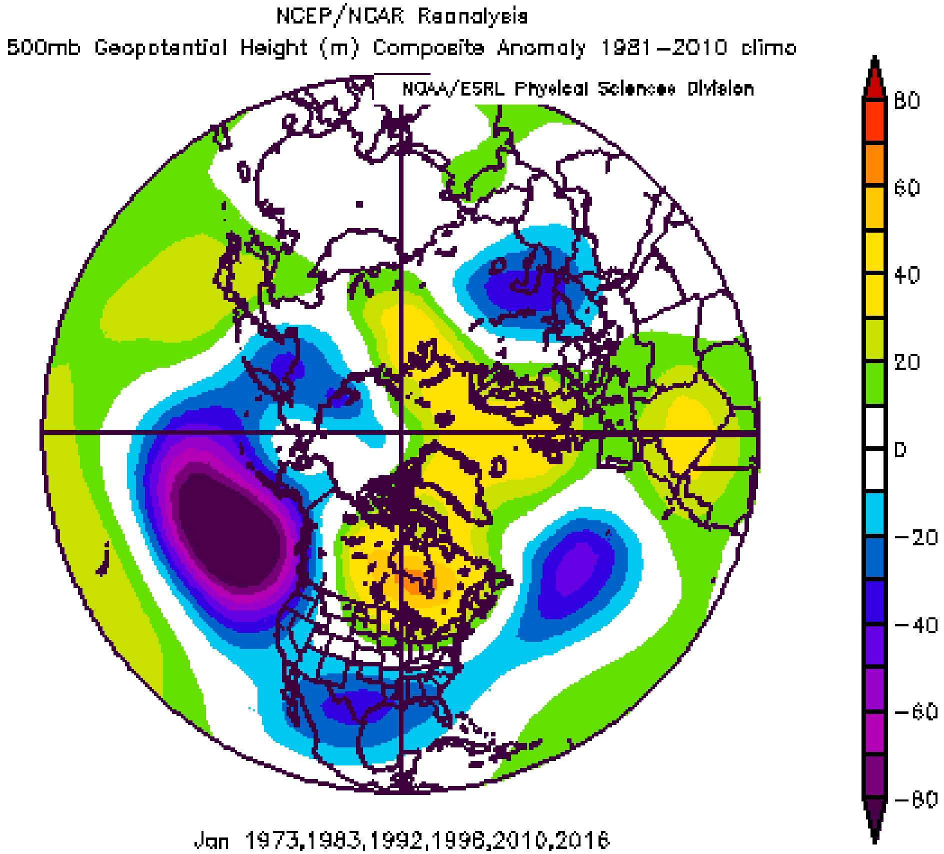
Plot link



500mb Geopotential Height (m) Composite Anomaly (1981−2010 Climatology)
2/1/19 to 2/16/19
NCEP/NCAR Reanalysis



NOAA/ESRL/Physical Science Division – University of Colorado at Boulder/CIRES



NCEP/NCAR Reanalysis
500mb Geopotential Height (m) Composite Anomaly 1981–2010 climo

NOAA/ESRL Physical Sciences Division

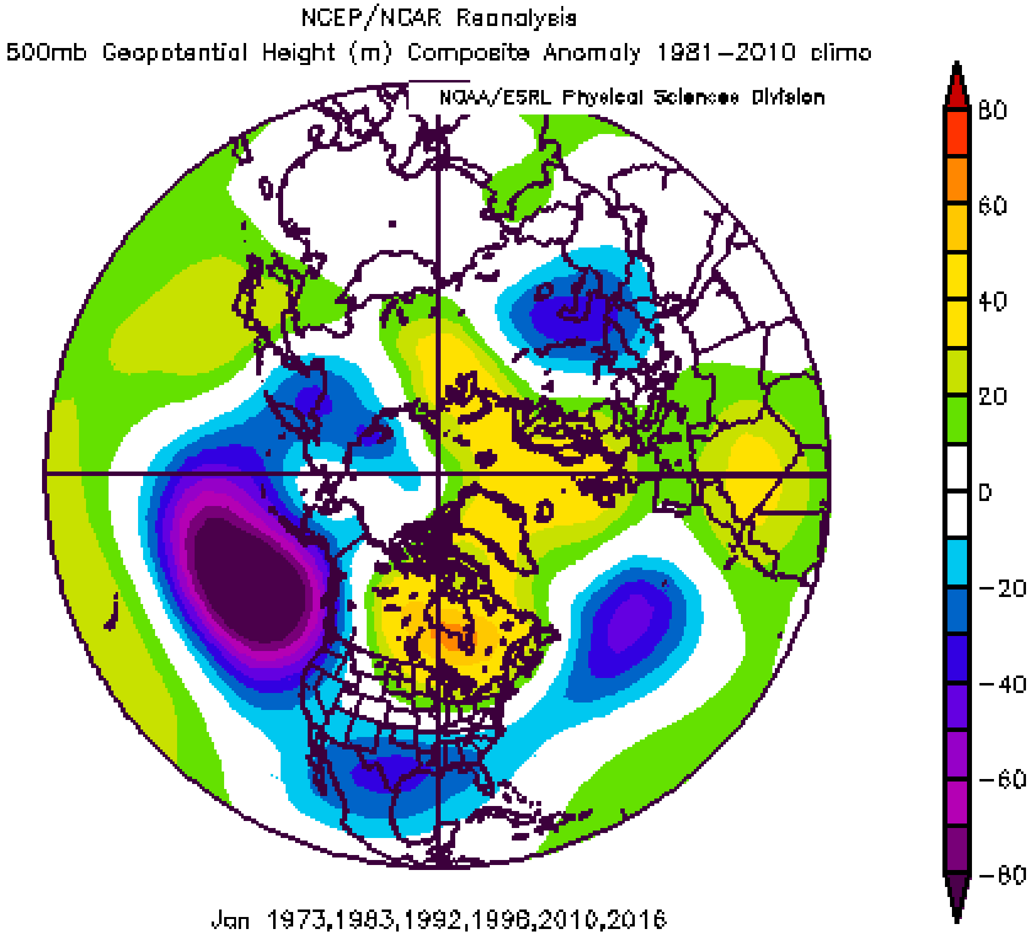
Jan 1973,1983,1992,1998,2010,2016

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


NCEP/NCAR Reanalysis
500mb Geopotential Height (m) Composite Anomaly 1981–2010 climo
NOAA/ESRL Physical Sciences Division
80
60
40
20
0
−20
−40
−60
−80
Jan 1973,1983,1992,1998,2010,2016

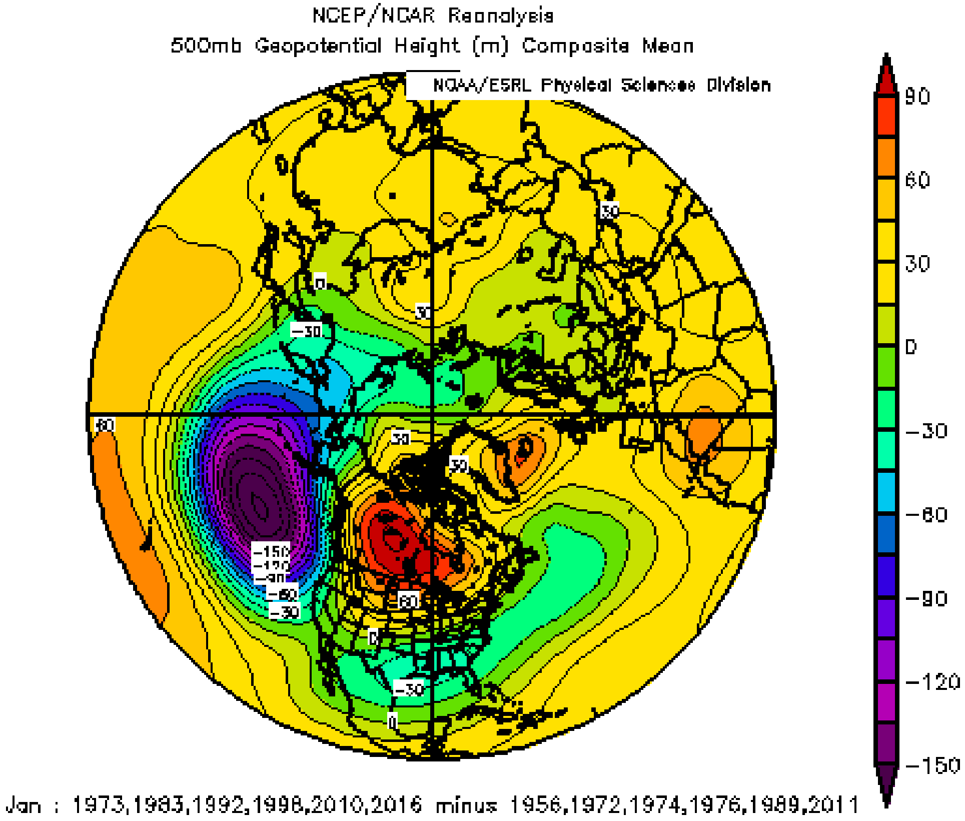
Don’t overdo it…

- 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

Create plot



NCEP/NCAR Reanalysis
500mb Geopotential Height (m) Composite Mean

NOAA/ESRL Physical Sciences Division

Jan : 1973,1983,1992,1998,2010,2016 minus 1956,1972,1974,1976,1989,2011

Run slc_time_series.m

- Comparing sample of 500 hPa height anomalies during pos MEI Jan’s to those during neg MEI Jan’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

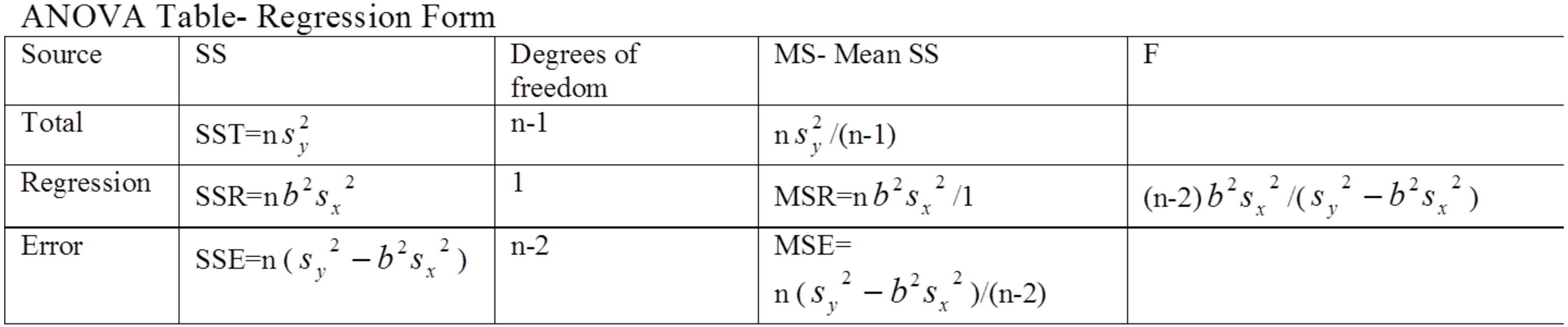
ANOVA and significance test of linear regression

- 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

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

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



ANOVA Table- Regression Form

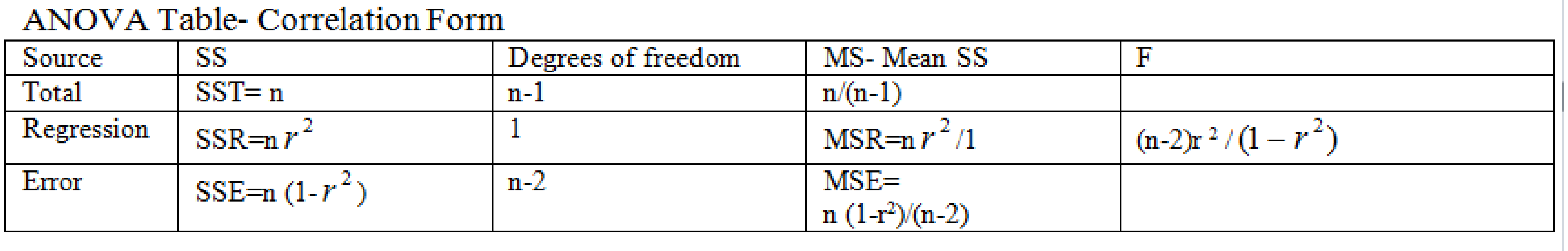
Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST=n s_y^2	n-1	n s_y^2 /(n-1)	
Regression	SSR=n $b^2 s_x^2$	1	MSR=n $b^2 s_x^2$ /1	(n-2) $b^2 s_x^2 /(s_y^2 - b^2 s_x^2)$
Error	SSE=n ($s_y^2 - b^2 s_x^2$)	n-2	MSE= n ($s_y^2 - b^2 s_x^2$)/(n-2)	

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

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)

Degrees of Freedom

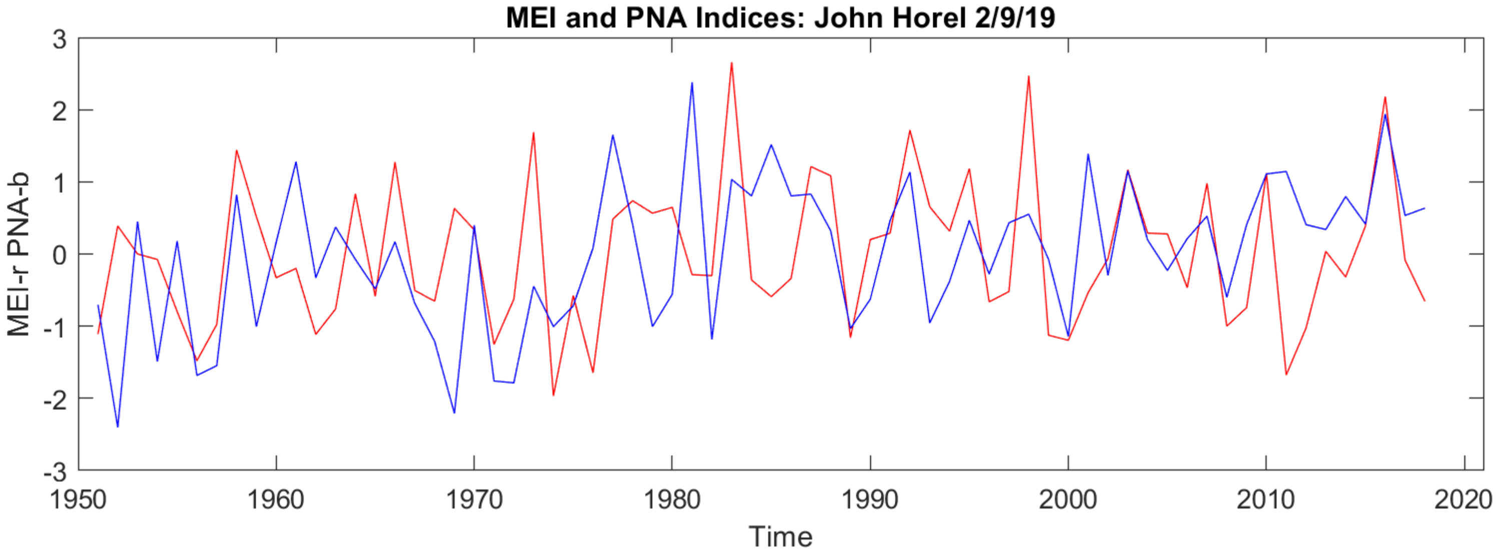
Two degrees of freedom are used up from the entire sample:

- 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).

MEI vs PNA Index



= 0.32

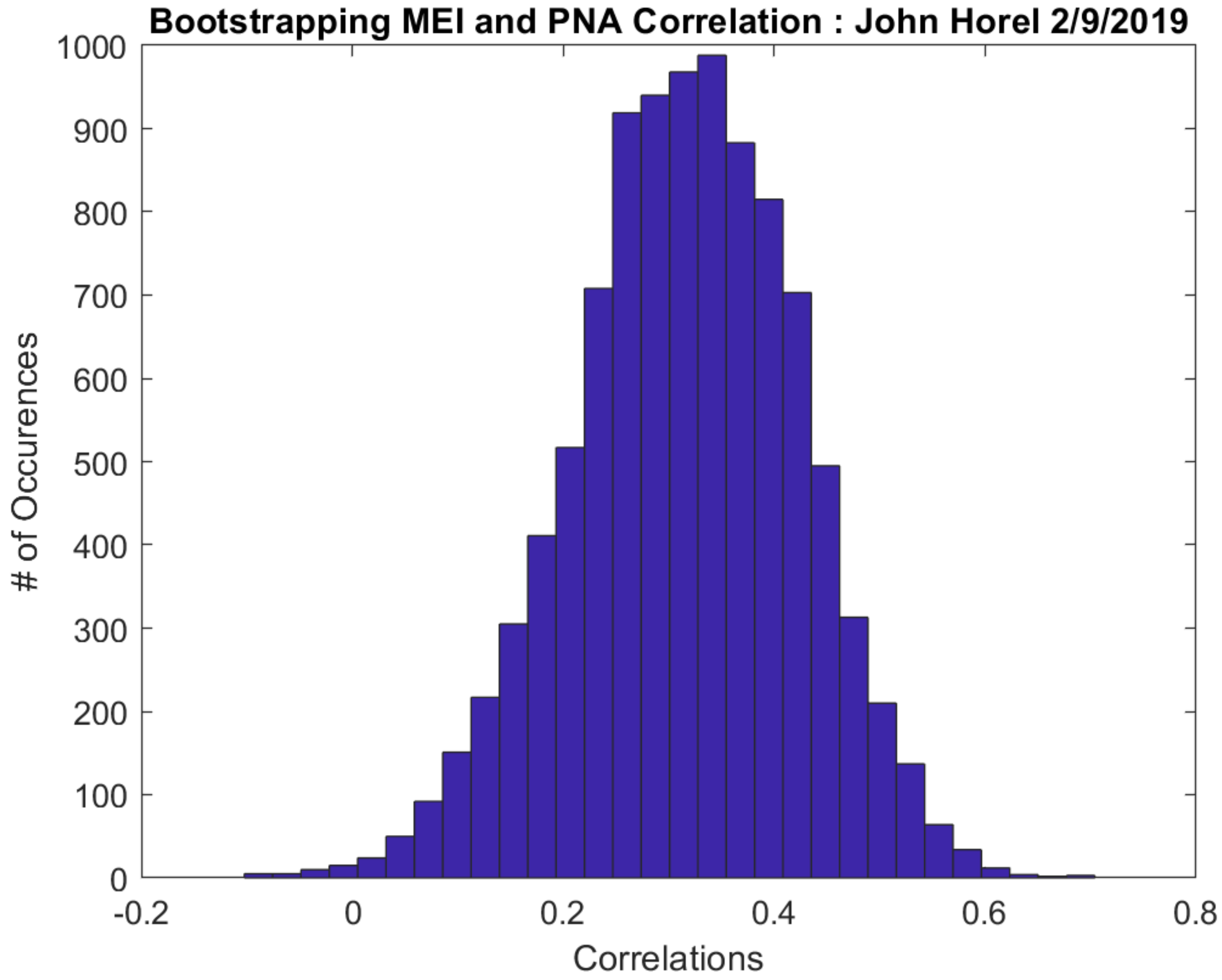
ANOVA Table- MEI vs PNA

Source	SS	Degrees of freedom	MS- Mean SS	F
Total	SST= 68	67	1.0	
Regression	SSR=7.0	1	7	7.0
Error	SSE=59	66	1.0	

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

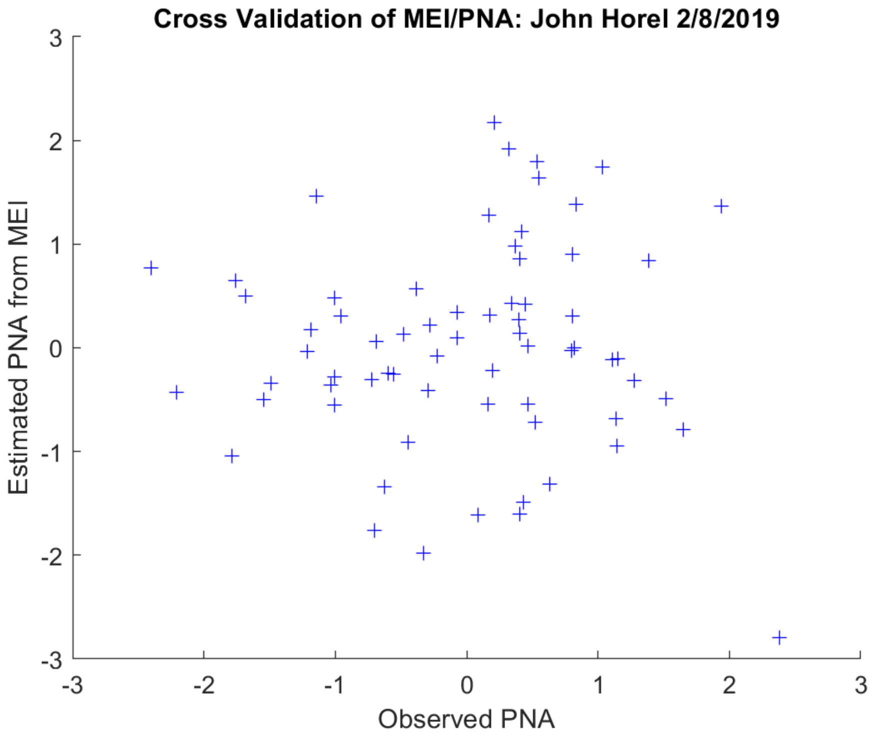
Resampling

- Resampling with replacement: bootstrap
- Correlation between MEI and PNA index = 0.32 based on all 68 years
- What if we take a sample n<68 randomly out of our 68 pairs 10,000 times?
- Each time we put all 68 years back into the hat for the next “draw”



Cross Validation

- Keep your powder dry- leave an independent sample for validation
- Cross validation: systematically remove data and recompute
- RMSE =1.3



Wilks (2016 BAMS)

- Computation of a single hypothesis test involves defining a null hypothesis H_0, which will be rejected in favor of an alternative hypothesis H_A if a sufficiently extreme value of the test statistic is observed
- Rejection of H_0 at a test level α occurs if the test statistic is sufficiently extreme that the probability- p value- of observing it or any other outcome even less favorable to H_0, if that null hypothesis is true, is no larger than α.
- If H_0 is rejected with $\alpha = 0.05$ (the most common, although an arbitrary choice), the result is said to be significant at the 5% level

Wilks (2016 BAMS)

- Most people assume that if N hypothesis tests (for example, each point on a map) then will have on average, α N erroneous rejections of the null hypothesis
- Instead use False Discovery Rate (FDR)
  - Sort in ascending order p_i values from i = 1, …, N hypothesis tests such that $p_1 \leq p_2 \leq \ldots \leq p_N$.
  - Local null hypothesis is rejected if its p_i value is no larger than a threshold level p_{FDR} that depends on the distribution of the sorted p values:
  - $p_i \leq 2\ \alpha\ i$ / N
- Run fdr.m

False Discovery Rate

Conservative
way to estimate

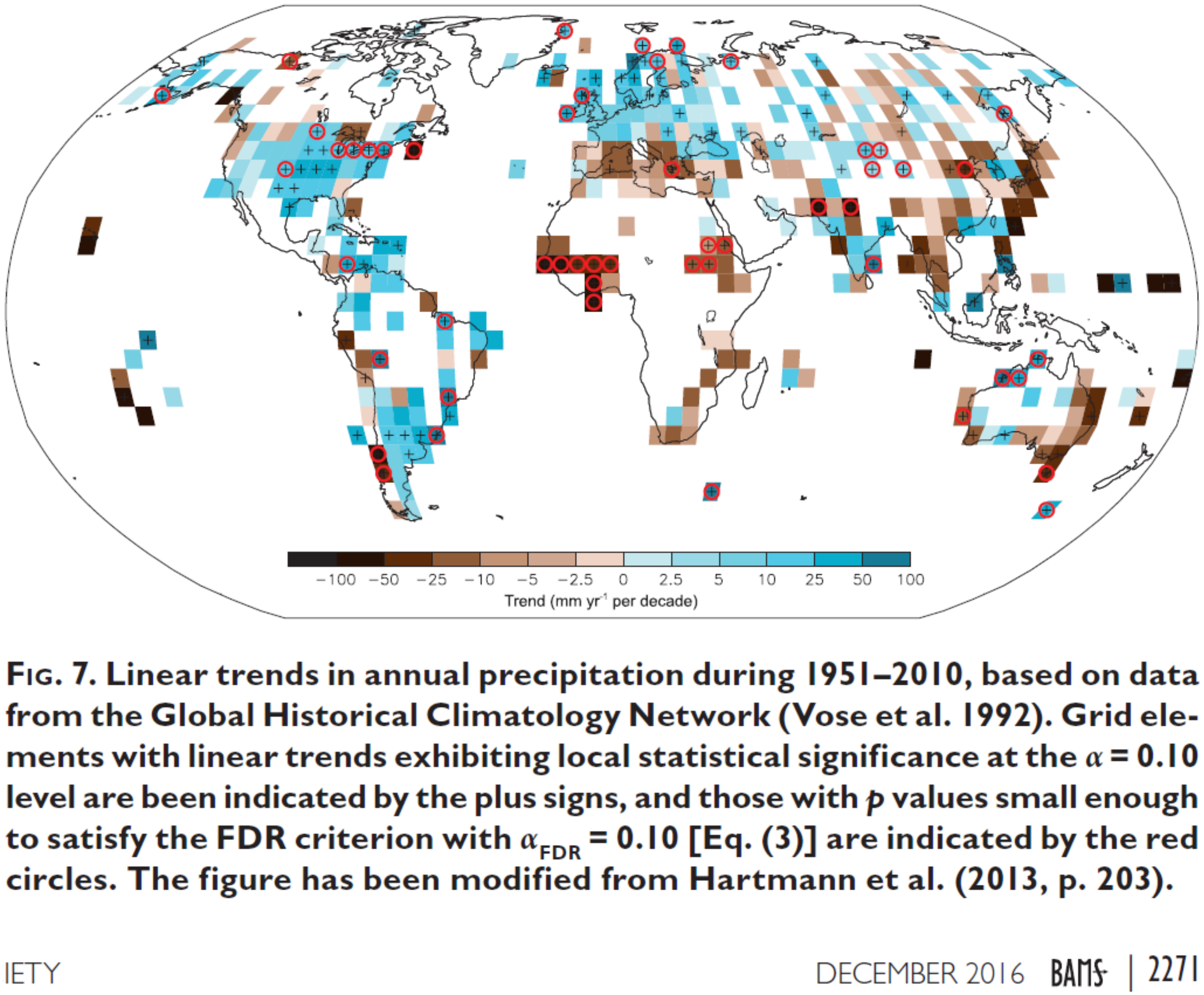
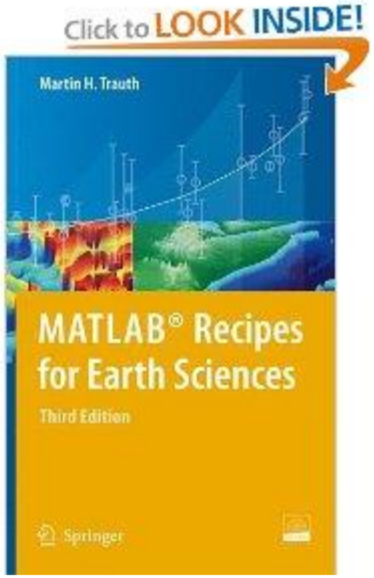
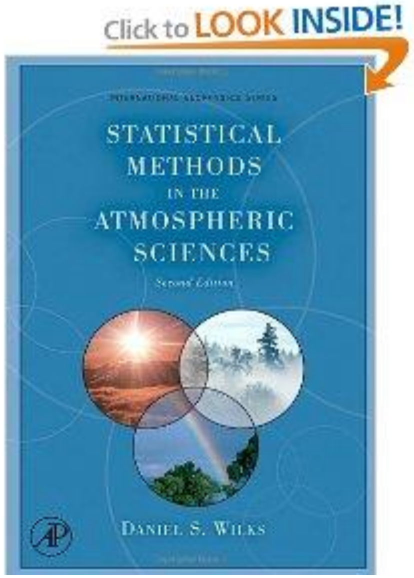


FIG. 7. Linear trends in annual precipitation during 1951–2010, based on data from the Global Historical Climatology Network (Vose et al. 1992). Grid elements with linear trends exhibiting local statistical significance at the $\alpha = 0.10$ level are been indicated by the plus signs, and those with p values small enough to satisfy the FDR criterion with $\alpha_{FDR} = 0.10$ [Eq. (3)] are indicated by the red circles. The figure has been modified from Hartmann et al. (2013, p. 203).

ATMOSPHERIC SCIENCES 5040/6040-Environmental Statistics

- Required Text: *Matlab Recipes for Earth Sciences.*
- Recommended text for 6040: Statistical *Methods in the Atmospheric Sciences*
- *Did you look at either?*





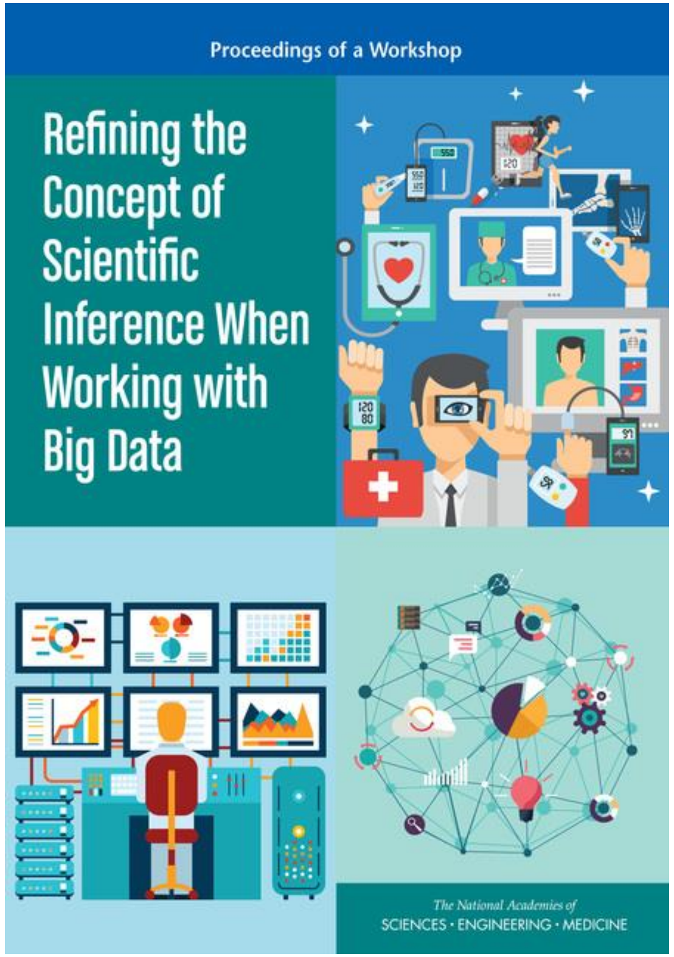
Course Learning Objectives

- State and use basic statistical metrics to analyze environmental information
- Develop proficiency to program and use MATLAB software as a tool to analyze environmental data sets
- State and demonstrate the characteristics of effective research; organize, quality control, and find relationship(s) among data

- **So…?**

You’ve been given weapons…

How are you going to use them?



Proceedings of a Workshop

Refining the Concept of Scientific Inference When Working with Big Data

The National Academies of SCIENCES · ENGINEERING · MEDICINE

June 2016

Scientific Inference

(1) big data holds both great promise and perils

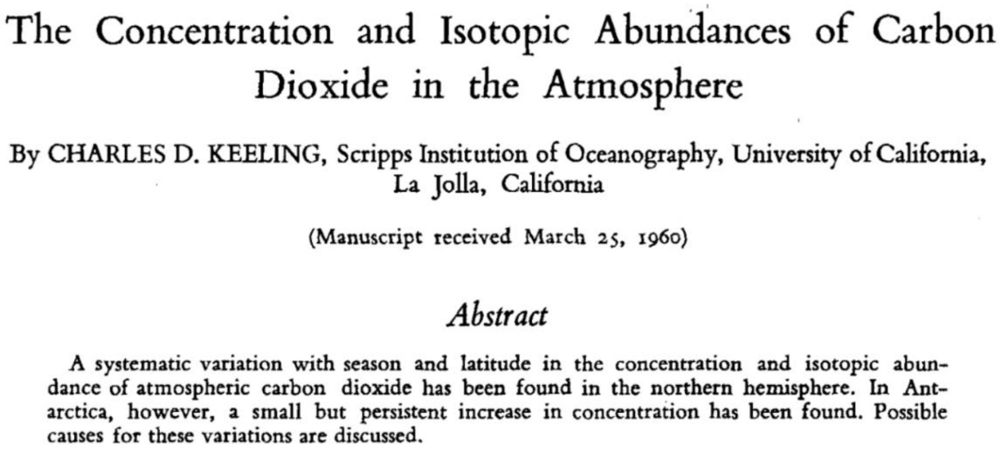
(2) inference requires evaluating uncertainty

(3) statisticians must engage early in experimental design and data collection activities

(4) open research questions can propel both the domain sciences and the field of statistics forward

(5) Opportunities exist to strengthen statistics education at all levels

Tellus 1960



The Concentration and Isotopic Abundances of Carbon Dioxide in the Atmosphere

By CHARLES D. KEELING, Scripps Institution of Oceanography, University of California, La Jolla, California

(Manuscript received March 25, 1960)

Abstract

A systematic variation with season and latitude in the concentration and isotopic abundance of atmospheric carbon dioxide has been found in the northern hemisphere. In Antarctica, however, a small but persistent increase in concentration has been found. Possible causes for these variations are discussed.

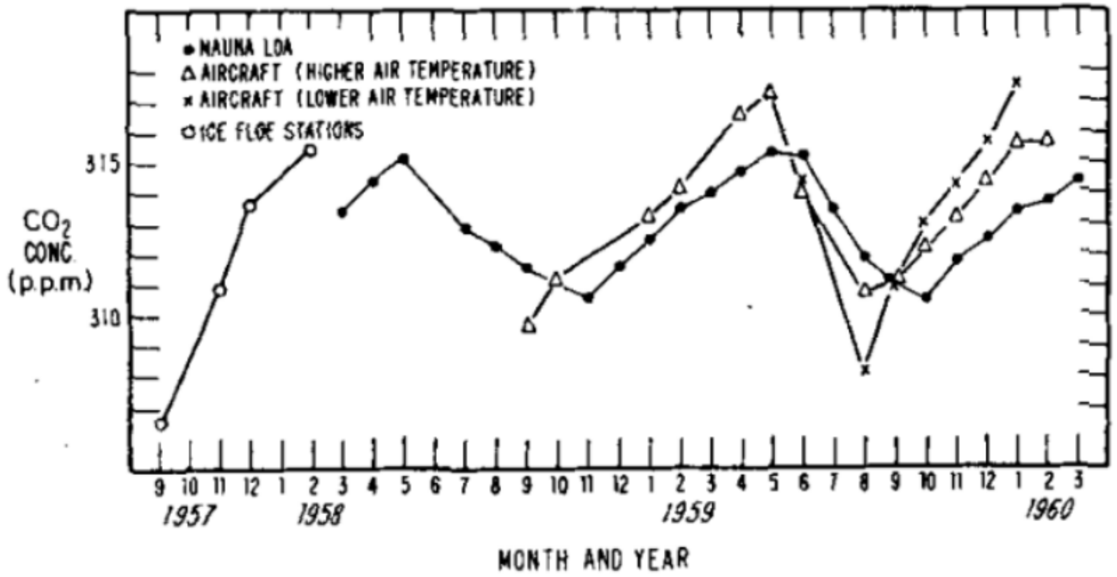
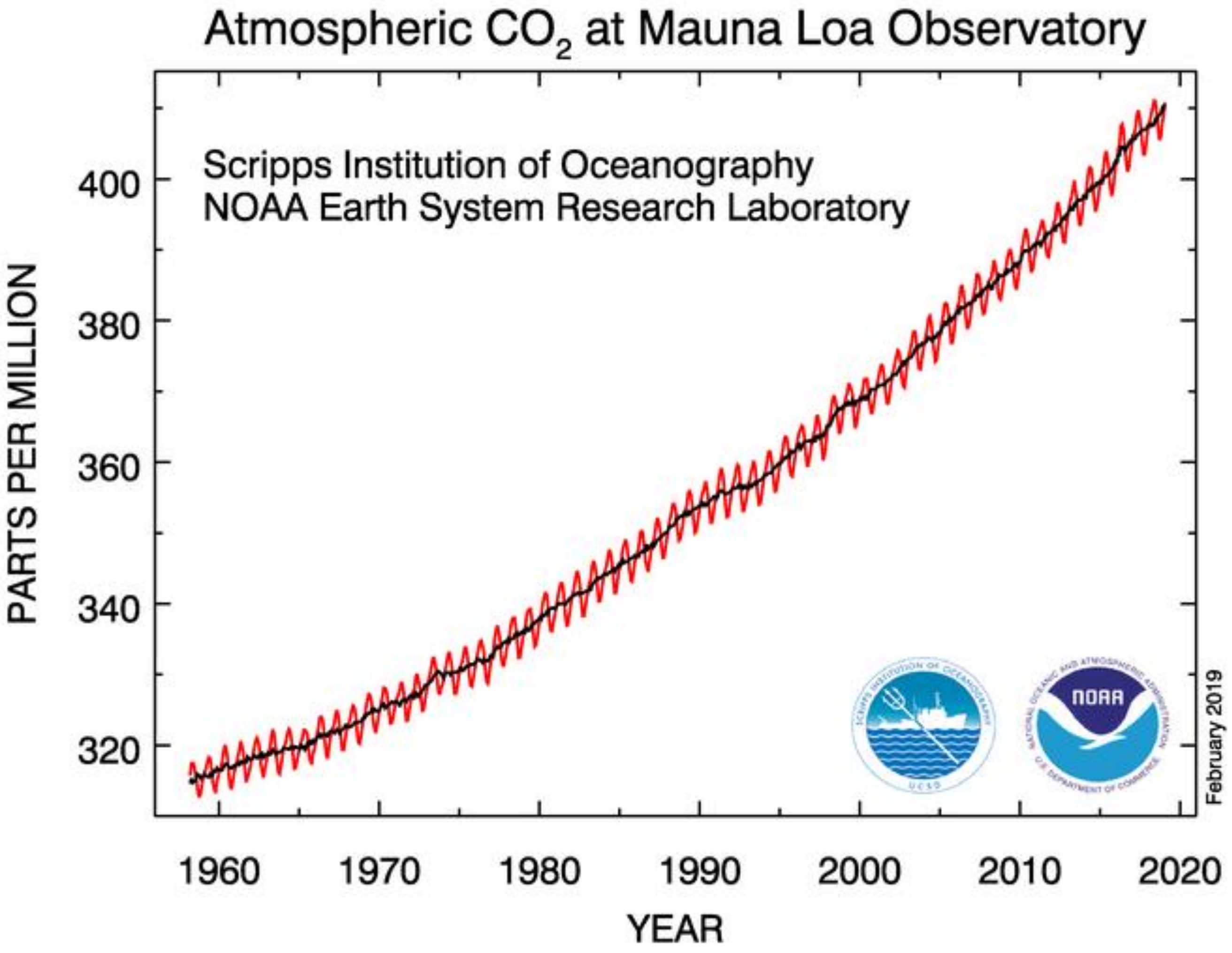
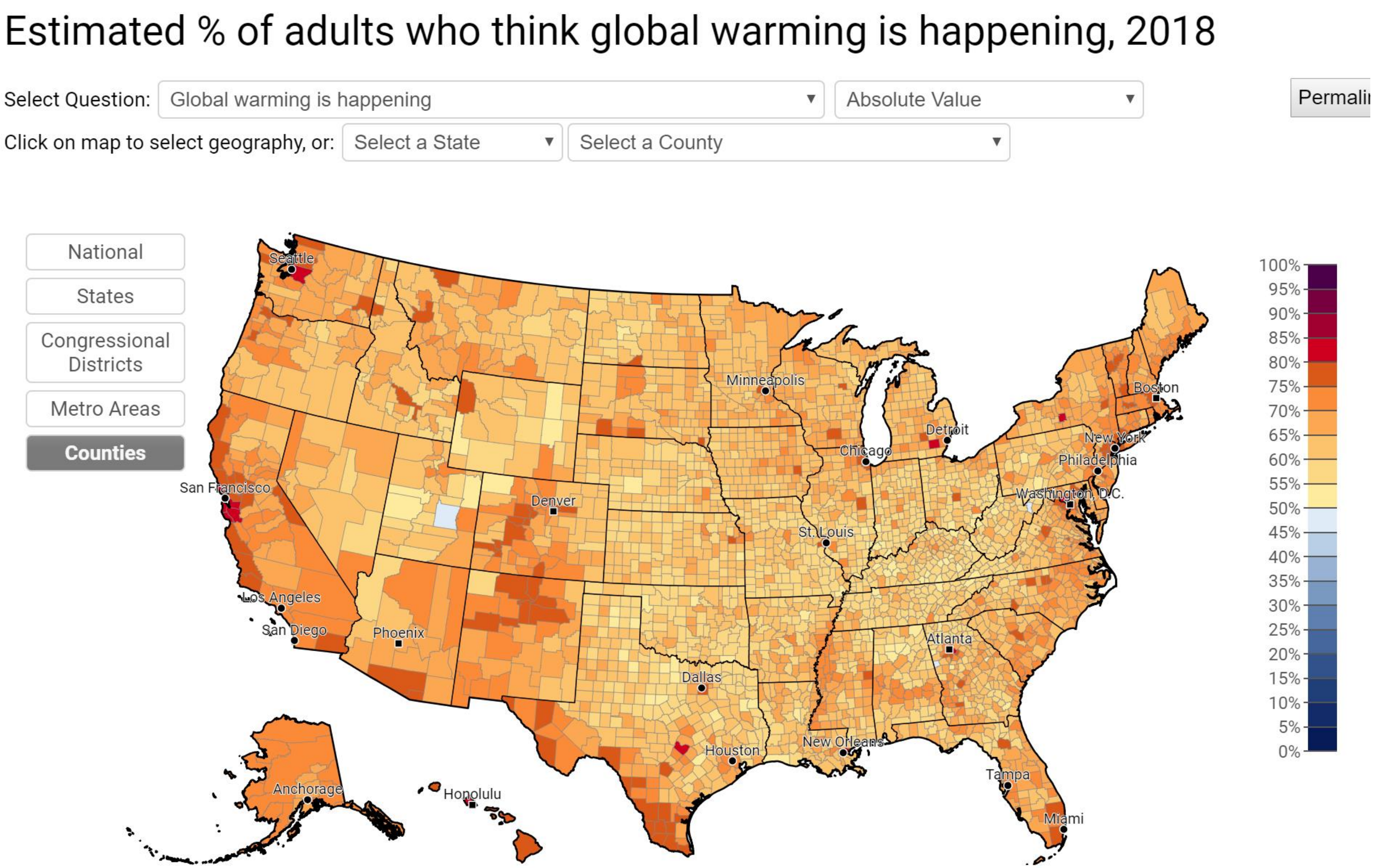


Fig. 1. Variation in concentration of atmospheric carbon dioxide in the Northern Hemisphere.



Atmospheric CO$_2$ at Mauna Loa Observatory

Yale



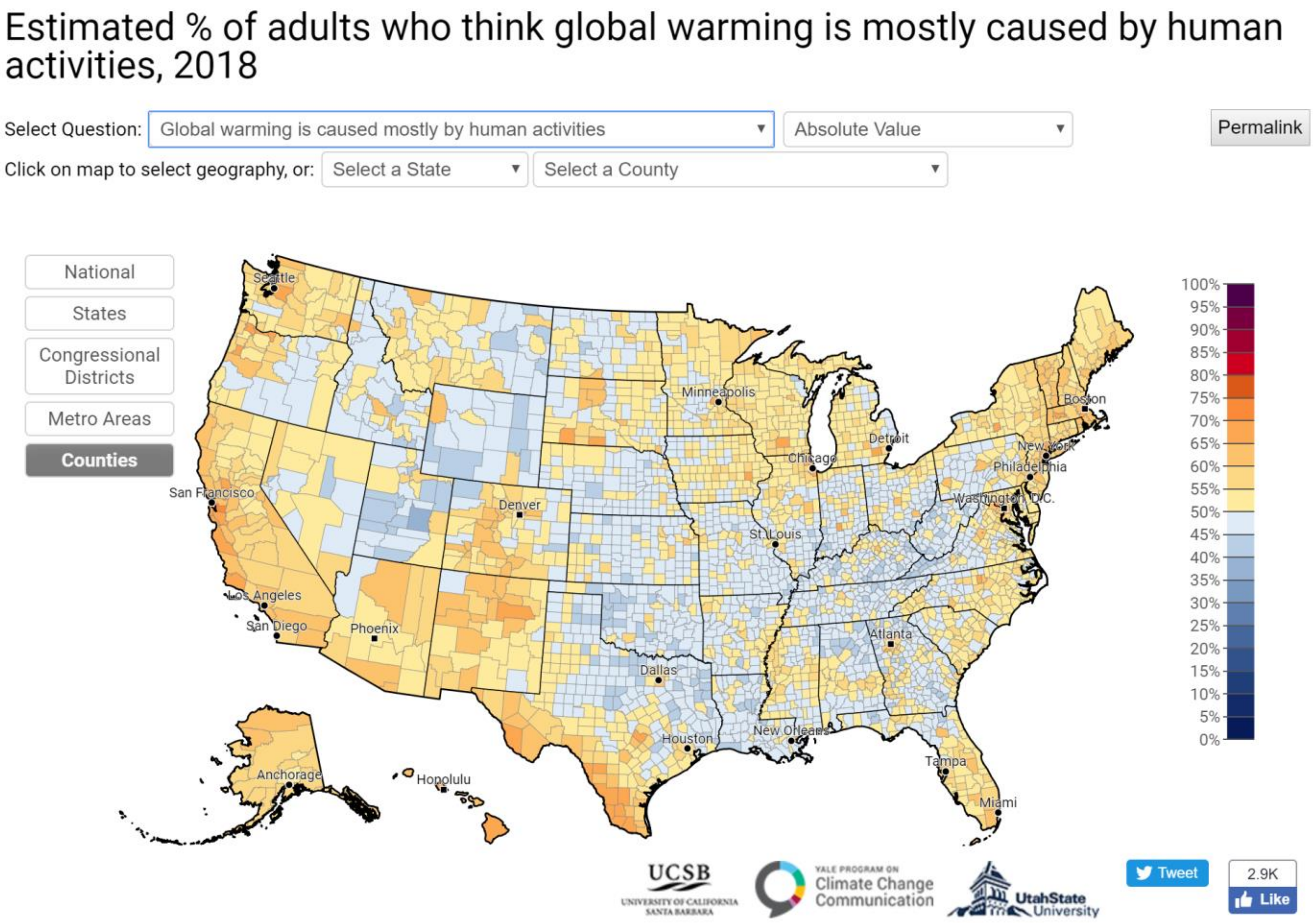
Estimated % of adults who think global warming is happening, 2018

Select Question: Global warming is happening ▾ Absolute Value ▾

Click on map to select geography, or: Select a State ▾ Select a County ▾

Most Americans -- in nearly every county across the United States-- understand the world is warming, according to Yale University research released

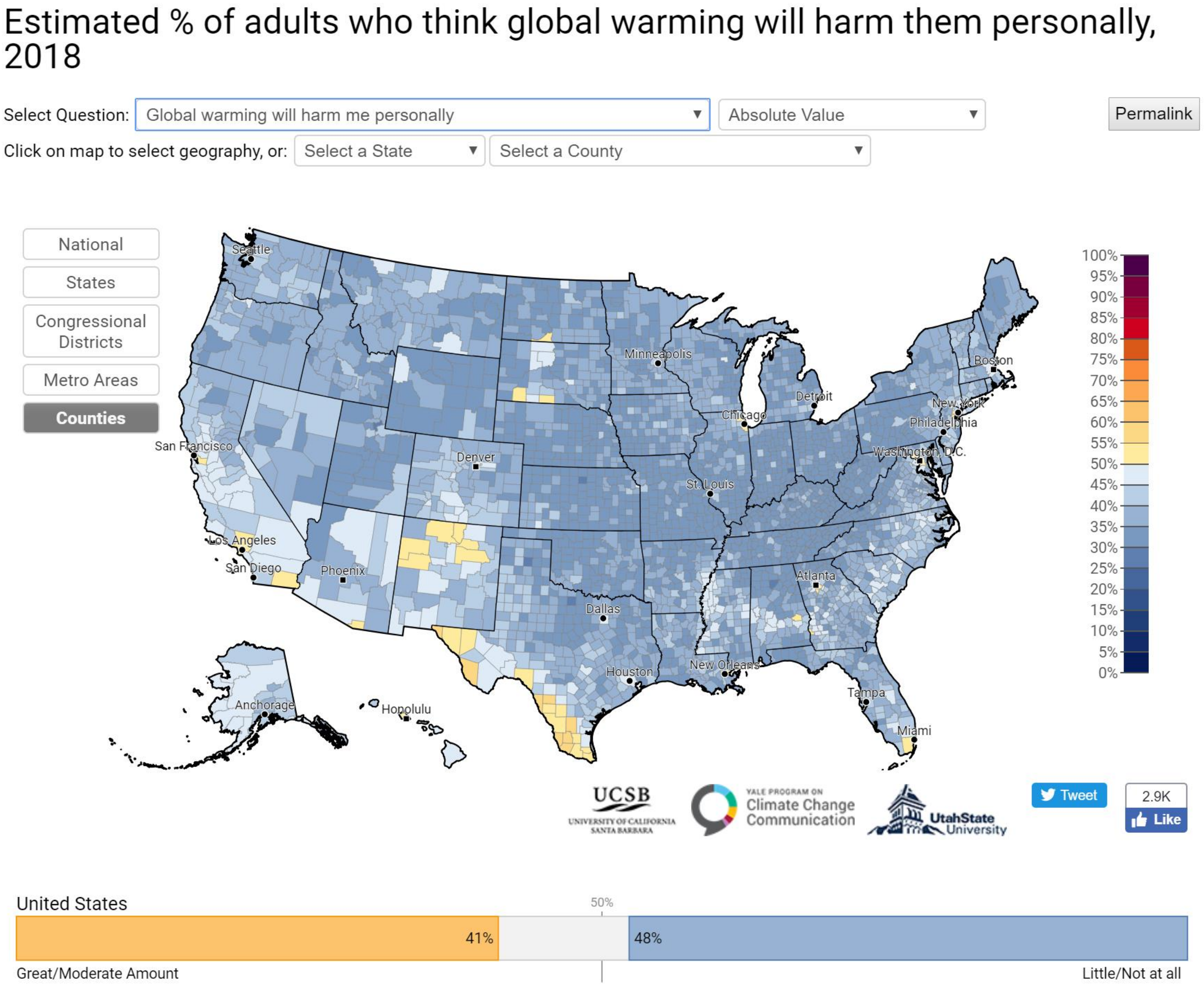
57 percent of Americans understand that humans are causing global warming



Estimated % of adults who think global warming is mostly caused by human activities, 2018

Select Question: Global warming is caused mostly by human activities ▾ Absolute Value ▾

Click on map to select geography, or: Select a State ▾ Select a County ▾



Estimated % of adults who think global warming will harm them personally, 2018

Select Question: Global warming will harm me personally ▾ Absolute Value ▾ Permalink

Click on map to select geography, or: Select a State ▾ Select a County ▾

National

States

Congressional Districts

Metro Areas

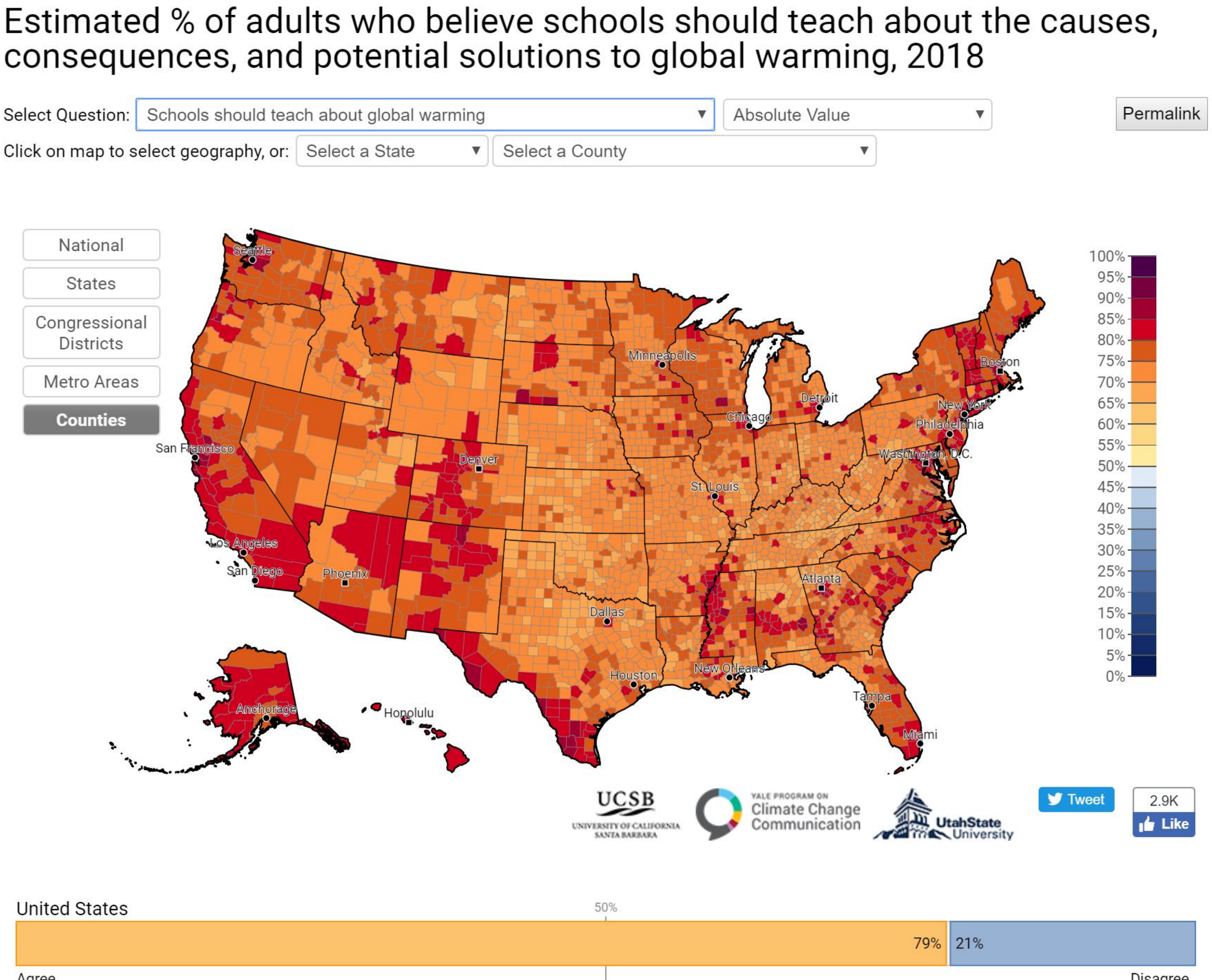
Counties

United States

50%

41% | 48%

Great/Moderate Amount | Little/Not at all



Estimated % of adults who believe schools should teach about the causes, consequences, and potential solutions to global warming, 2018

Select Question: Schools should teach about global warming ▾ Absolute Value ▾

Permalink

Click on map to select geography, or: Select a State ▾ Select a County ▾

National

States

Congressional Districts

Metro Areas

Counties

UCSB UNIVERSITY OF CALIFORNIA SANTA BARBARA

YALE PROGRAM ON Climate Change Communication

UtahState University

Tweet

2.9K Like

United States

What makes for meaningful environmental statistics?

Rank (1-high; 4 low) individually what you consider to be the most important and then discuss:

_____ records (high’s/low’s, extreme events)

_____ have high societal impact

_____ help to explain physical phenomena

_____ can be used to make empirical forecasts for commercial benefit (commodity or weather trading)

Discuss whether environmental statistics can be used effectively to influence Utah public policy? And, if so, how?

- To prepare for rare events (mud slides, flash floods, high wind damage, earthquakes, etc.)

- To modify personal behavior that affects public resources (clear air, water, etc.)

- To plan for future water availability

Assignments: wrapping things up

- Online/take home final for undergrads/grads posted
- Undergrads could use MLIB Wednesday IF you don’t have a class in this time slot during the 2nd half
- It’s ok to discuss coding issues of exam with others; not the answers to any questions, work independently on those