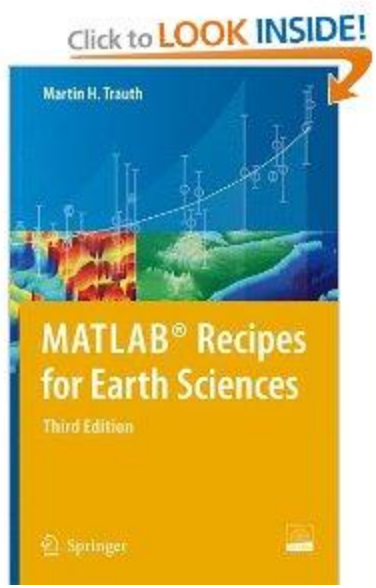
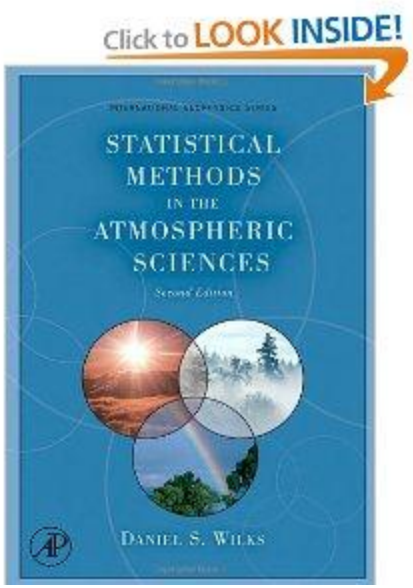
ATMOSPHERIC SCIENCES 5040/6040- Environmental Statistics

- 5040: 1.5 Units. 6040: 3 Units
- Instructors: John Horel, Court Strong, TA Levi Thatcher
- Required Text: *Matlab Recipes for Earth Sciences.*
- Recommended text for 6040: Statistical *Methods in the Atmospheric Sciences*





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
- 5040: State and demonstrate the characteristics of effective research; organize, quality control, and find relationship(s) among data
- 6040:State and demonstrate the characteristics of effective research: distill a general interest in a subject into a specific question/hypothesis that can be evaluated; organize the data; find relationship(s) among the data; and examine the significance of the results

Format

- You must read the assigned text and class notes prior to the corresponding lecture
- Each chapter of notes contains review questions that much be completed and turned in
- Other assignments will be completed as well
- 5040:/first half 6040: You will be assigned to evaluate the data collected from an automated weather station within the Salt Lake Valley during the PCAPS field
- 6040: Instead of a final exam, you will pursue a research question using statistical methods to analyze data

Class Policies and Grading

- 5040:
  - (1) note review questions (25%)
  - (2) final (25%)
  - (3) class assignments (50%)
- 6040:
  - (1) assignments (60%)
  - (2) mid-term exam (20%)
  - (3) research project (20%)

MATLAB

- It is not required that you own a copy of Matlab
- We will be using the statistics toolbox, which requires you to have an account on the CHPC network
- CHPC online form: https://www.chpc.utah.edu/apps/profile/account_request.php
  - Indicate Horel as advisor
- You may need to access Matlab on an Atmospheric Science department workstation (meteo08.chpc.utah.edu or meteo09.chpc.utah.edu)
- Levi will have lab sessions on Tuesdays after class in 7XX WBB (synoptic lab)

5040/6040 Course Outline

- Week 1. Course Overview. Basic statistical concepts. Class Notes 1. Text Chap. 1& 2.
- Week 2. Using matlab. Exploratory Univariate Data Analysis. Class Notes 2. Text Chap. 3.1-3.3
- Week 3. Exploratory Univariate Data Analysis. Continued. Text Chap. 3.4-6, 3.9.
- Week 4. Theoretical Distributions and Hypothesis Testing. Class Notes 3.
- Week 5. Theoretical Distributions and Hypothesis Testing. Continued. Exploratory Multivariate Data Analysis. Class Notes 4. Text Chap. 4.
- Week 6. Exploratory Multivariate Data Analysis. Continued.
- Week 7. A statistics sampler. Sections from Text Chapters will be assigned.
- Week 8. Mar. 1. Final: 5040. Mid-term: 6040

6040 Course Outline

- Week 9. Topics in regression: Class Notes 5, Text Chapter 4.9-4.10
- Week 10. Topics in regression continued: Class Notes 5, Text Chapter 4.5-4.6
- Week 11. Harmonic analysis: Class Notes 6, Text Chapter 5
- Week 12. Harmonic analysis continued: Class Notes 6, Text Chapter 6
- Week 13. Data reduction: Class notes 7, Text Chapter 9
- Week 14. Data reduction continued: Class Notes 7
- Week 15. April 26. Poster session

The Washington Post

The Checkup

Vaccine's link to autism is discredited again with finding of fraud in research

Tuesday, January 11, 2011

It's false. Really, seriously, false.

The research that launched an enduring but apparently erroneous belief that autism is caused by a common childhood vaccination has been debunked, disclaimed - and now debunked again.

Study links autism to babies' age difference

Risk triples for children born within a year of siblings.

BY LIZ SZABO
USA TODAY

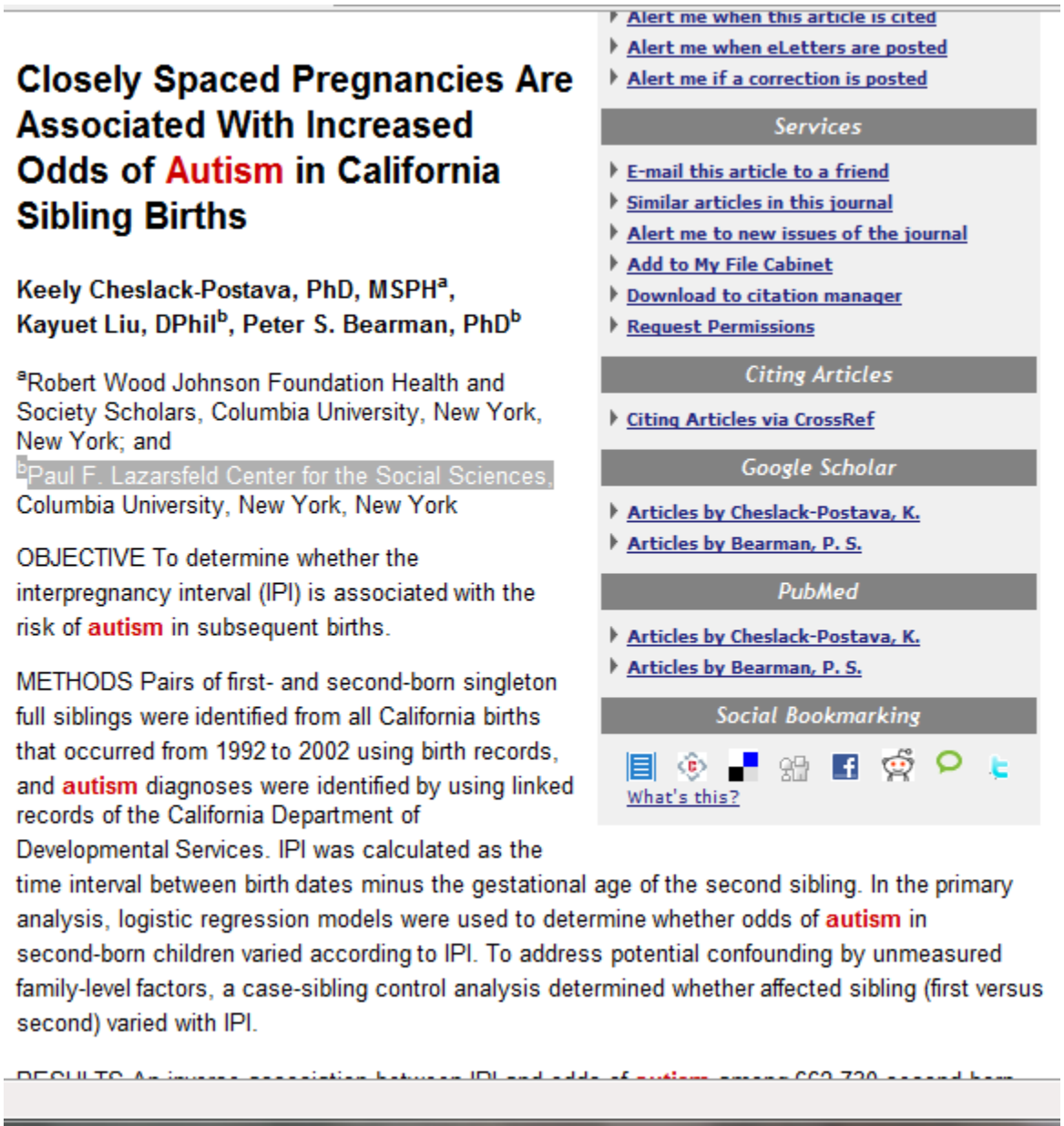
First published Jan 10 2011 06:06AM
Updated Jan 10, 2011 06:06AM

Having babies close together appears to increase the risk of autism, a new study shows, and the more closely siblings are spaced, the higher the risk that the younger one will be diagnosed with autism.

The findings, published in today's Pediatrics, are the latest clue in doctors' efforts to unravel the mystery of autism and its causes. The disorder, which affects one in 110 American children, has baffled experts and sparked controversy for years.

Researchers say autism is a family of related disorders — probaby with different

- Autism



Alert me when this article is cited
Alert me when eLetters are posted
Alert me if a correction is posted

Services

E-mail this article to a friend
Similar articles in this journal
Alert me to new issues of the journal
Add to My File Cabinet
Download to citation manager
Request Permissions

Citing Articles

Citing Articles via CrossRef

Google Scholar

Articles by Cheslack-Postava, K.
Articles by Bearman, P. S.

PubMed

Articles by Cheslack-Postava, K.
Articles by Bearman, P. S.

Social Bookmarking

What's this?

Closely Spaced Pregnancies Are Associated With Increased Odds of Autism in California Sibling Births

Keely Cheslack-Postava, PhD, MSPH[a], Kayuet Liu, DPhil[b], Peter S. Bearman, PhD[b]

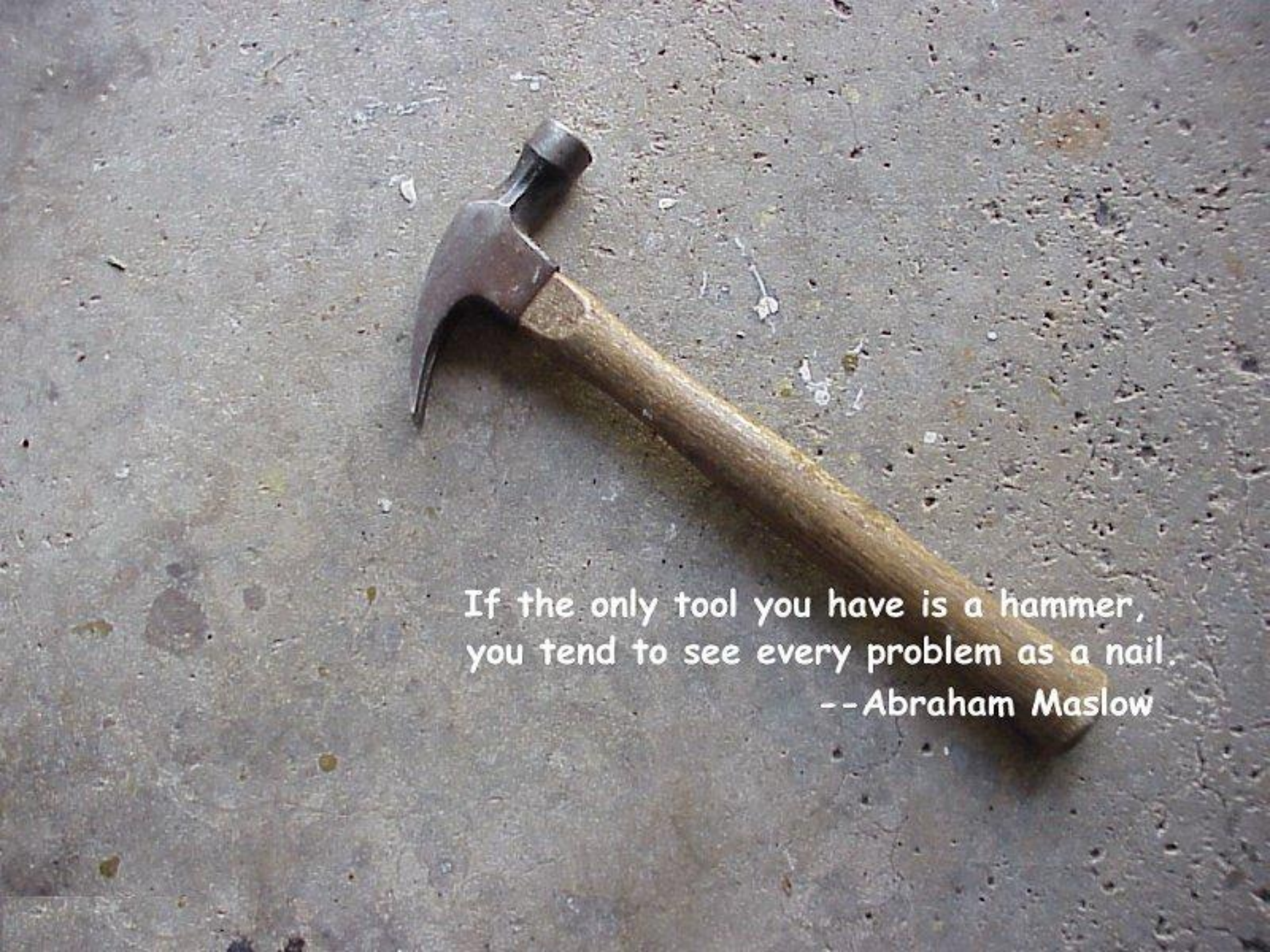
[a]Robert Wood Johnson Foundation Health and Society Scholars, Columbia University, New York, New York; and
[b]Paul F. Lazarsfeld Center for the Social Sciences, Columbia University, New York, New York

OBJECTIVE To determine whether the interpregnancy interval (IPI) is associated with the risk of **autism** in subsequent births.

METHODS Pairs of first- and second-born singleton full siblings were identified from all California births that occurred from 1992 to 2002 using birth records, and **autism** diagnoses were identified by using linked records of the California Department of Developmental Services. IPI was calculated as the time interval between birth dates minus the gestational age of the second sibling. In the primary analysis, logistic regression models were used to determine whether odds of **autism** in second-born children varied according to IPI. To address potential confounding by unmeasured family-level factors, a case-sibling control analysis determined whether affected sibling (first versus second) varied with IPI.

Problems with statistics?

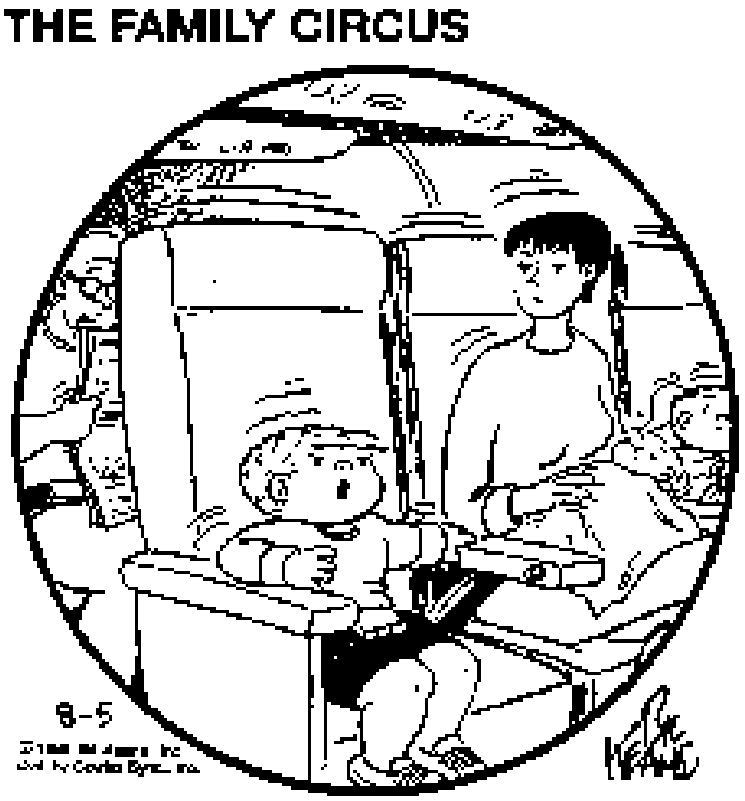
- Oriented towards confirming preconceived ideas?
- Start with a technique and look for a data set to apply it to?
- There's always two sides to every issue?
- Example
- What do you think? Other examples of poor statistics?


If the only tool you have is a hammer,
you tend to see every problem as a nail.
--Abraham Maslow

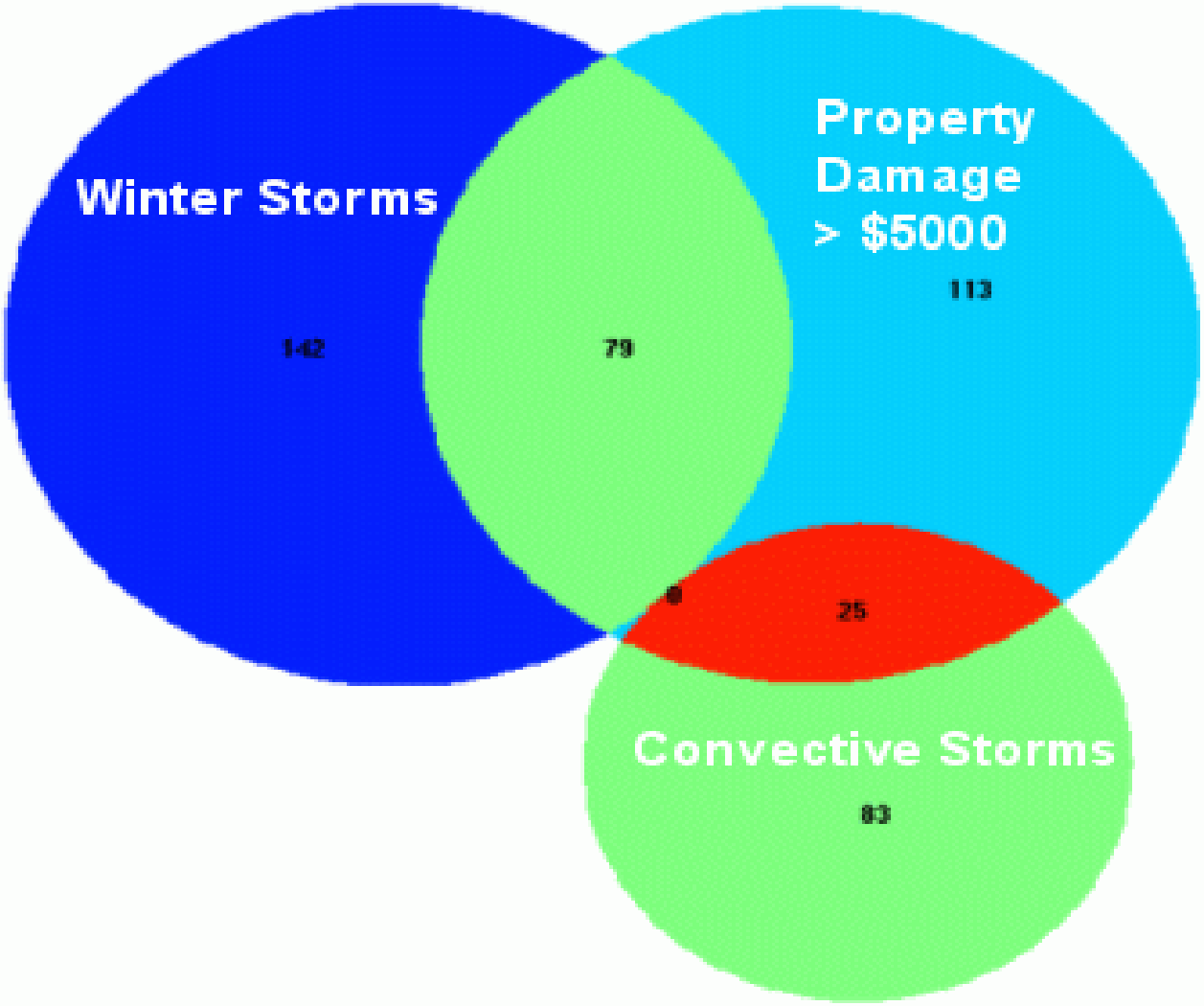
What's the Goal??

- Exploratory or descriptive statistics:
  - Organize and interpret volumes of data

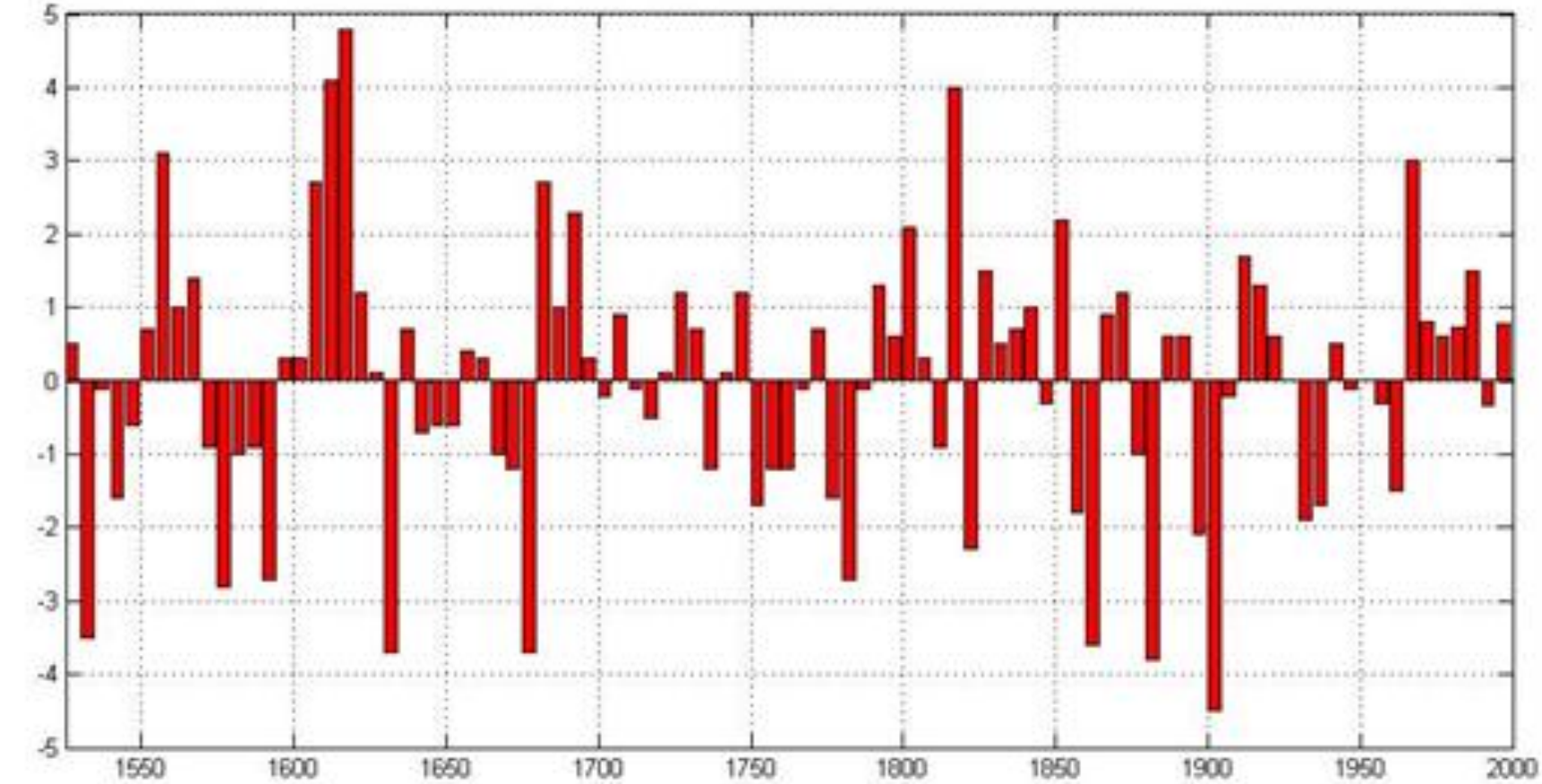
- Inferential statistics:
  - Assess the underlying physical processes that generate environmental data



Storm Reports: Salt Lake County 1993-2005

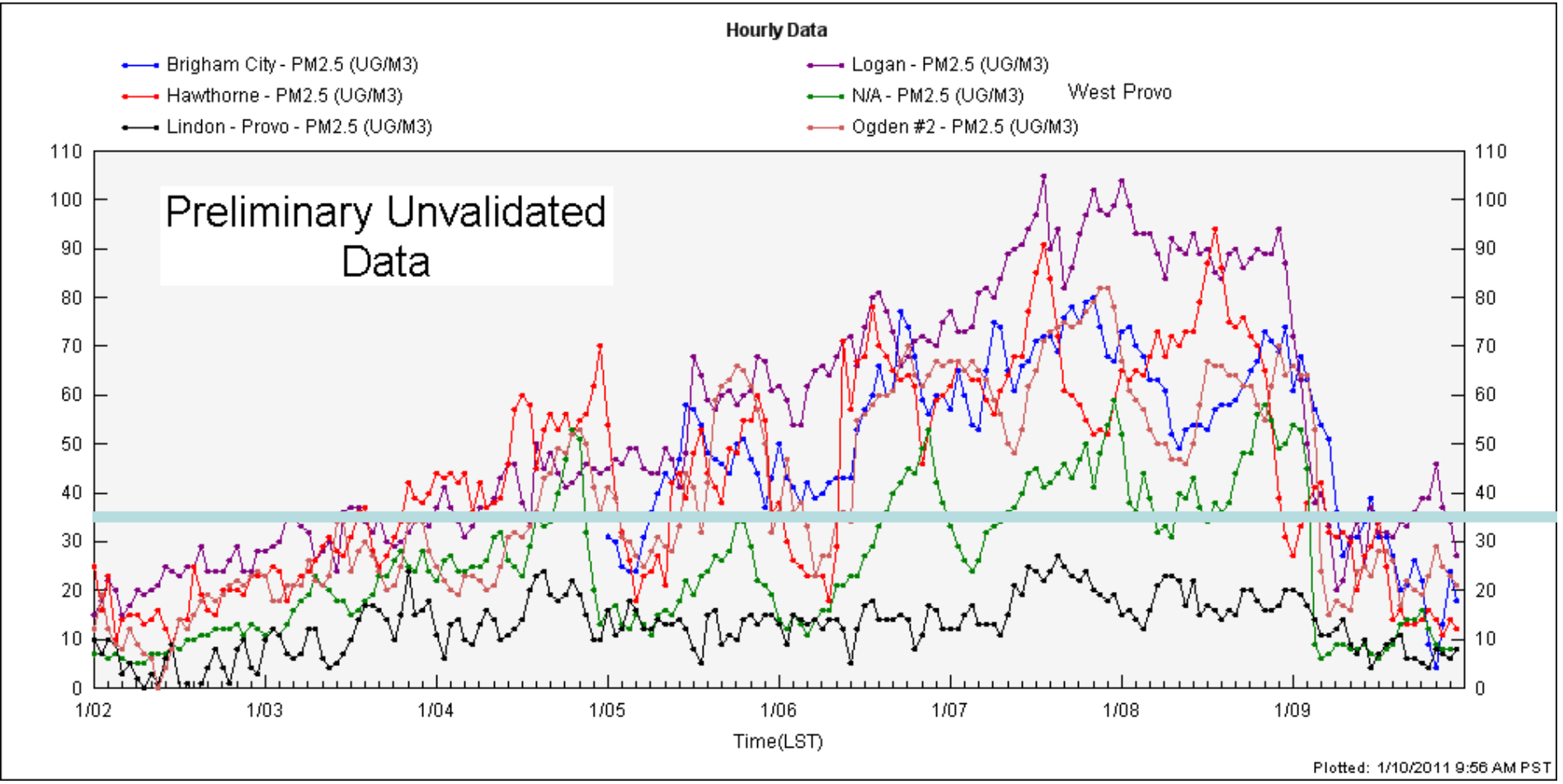


Utah Tree Ring Index



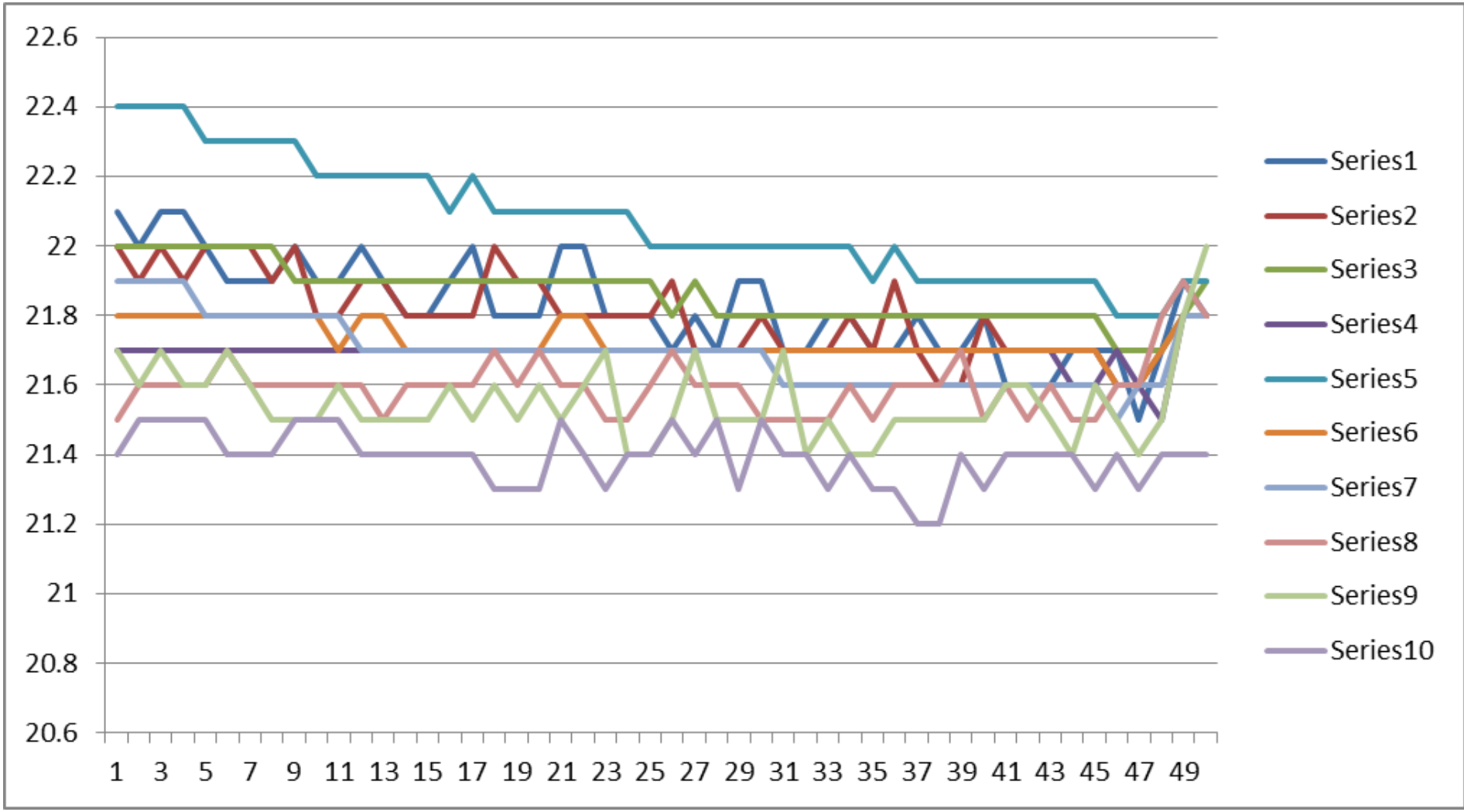
http://lwf.ncdc.noaa.gov/paleo/treering/isotope/iso-drought.html

Local Scale: PM 2.5



10 Kestrel Temperature Sensors

10 Kestrel Temperature Sensors over 25 minutes
2 samples per minute



Observations and Truth

- True value- value of a quantity sought through measurement, but unknown usually in the field
- Truth depends on application
- Assumption: average of many unbiased observations should be same as expected value of truth
- However, accurate observations may be biased or *unrepresentative* due to siting or other factors

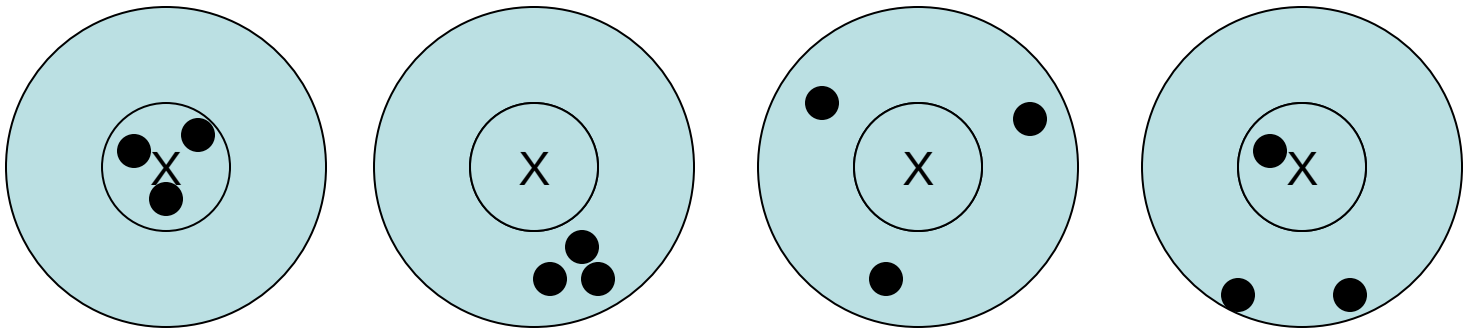
Causes of Uncertainty

- 1. we can never measure the environment with complete accuracy and precision
- 2. the environment is a chaotic system, which is a maddening combination of randomness and order arising from the characteristics of a complex nonlinear system,
- 3. our understanding of the environmental system is imperfect, so physical (and certainly statistical) models do not capture the complete behavior of the system.

Gauging Uncertainty

- Accuracy- difference in response between a standard and instrument in varying environmental conditions a measure of how close a measurement is to the “true” value

- Precision- how well repeated measurements of some quantity agree with each other. A precise instrument can be inaccurate



Systematic vs. Random Errors

- Random- that which is not precisely predictable or determinable
- Systematic- errors arising from a consistent response of a measuring device to environmental conditions or faulty characteristics of instrumentation that occurs frequently

Observations

- Observations are not perfect…
  - Gross errors
  - Local siting errors
  - Instrument errors
  - Representativeness errors

Getting a Handle on Siting Issues & Observational Errors

1. Metadata errors
2. Instrument errors (exposure, maintenance, sampling)
3. Local siting errors (e.g., artificial heat source, overhanging vegetation, observation at variable height above ground due to snowpack)
4. “Errors of representativeness” – correct observations that are capturing phenomena that are not representative of surroundings on broader scale (e.g., observations in vegetation-free valleys and basins surrounded by forested mountains)

Are All Observations Equally Good?

- Why was the sensor installed?
  - Observing needs and sampling strategies vary (air quality, fire weather, road weather)
- Station siting results from pragmatic tradeoffs: power, communication, obstacles, access
- Use common sense and experience
  - Wind sensor in the base of a mountain pass will likely blow from only two directions
  - Errors depend upon conditions (e.g., temperature spikes common with calm winds)
  - Pay attention to metadata
- Monitor quality control information
  - Basic consistency checks
  - Comparison to other stations

MesoWest Surface Weather Maps - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://mesowest.utah.edu/cgi-bin/droman/mesomap.cgi?state=NV&rawsflag=290&day1=10&month1

madden julian oscill

MesoWest Surface Weather Maps

Login | Create a U

MesoWest Region NEVADA Product Surface Weather Maps Go MesoWe

Hide Menu | Show Tables | Search | Active Fires | Map Product: Default Change

Data Selection

Region/Zone | Radius

NEVADA

Network: All Networks

Units: English

Map It!

Display

Overlay 1: Current Temp

Overlay 2: -- None --

Highlight Data

Show Table

Other Features

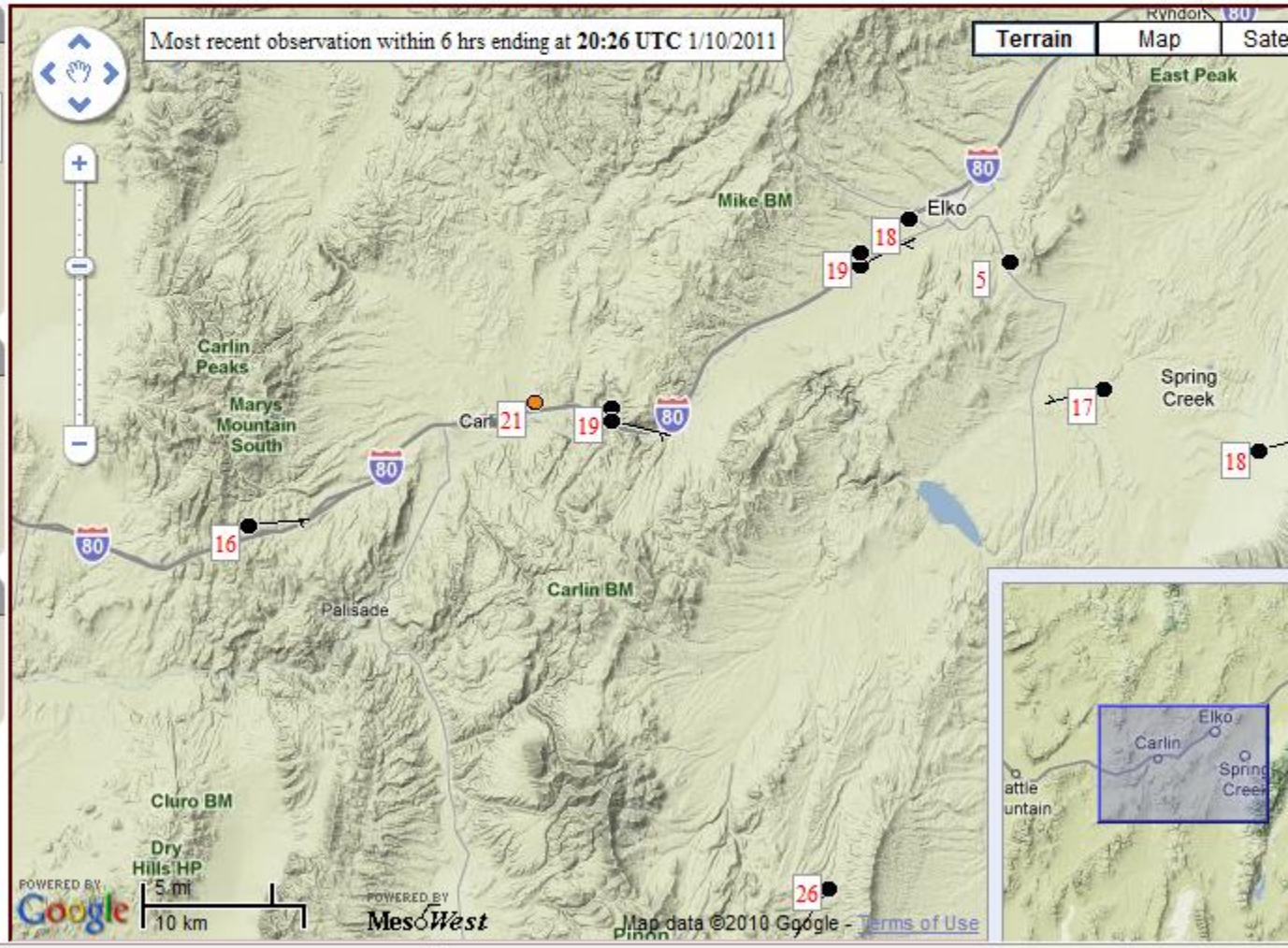
Time Options

Layer Options

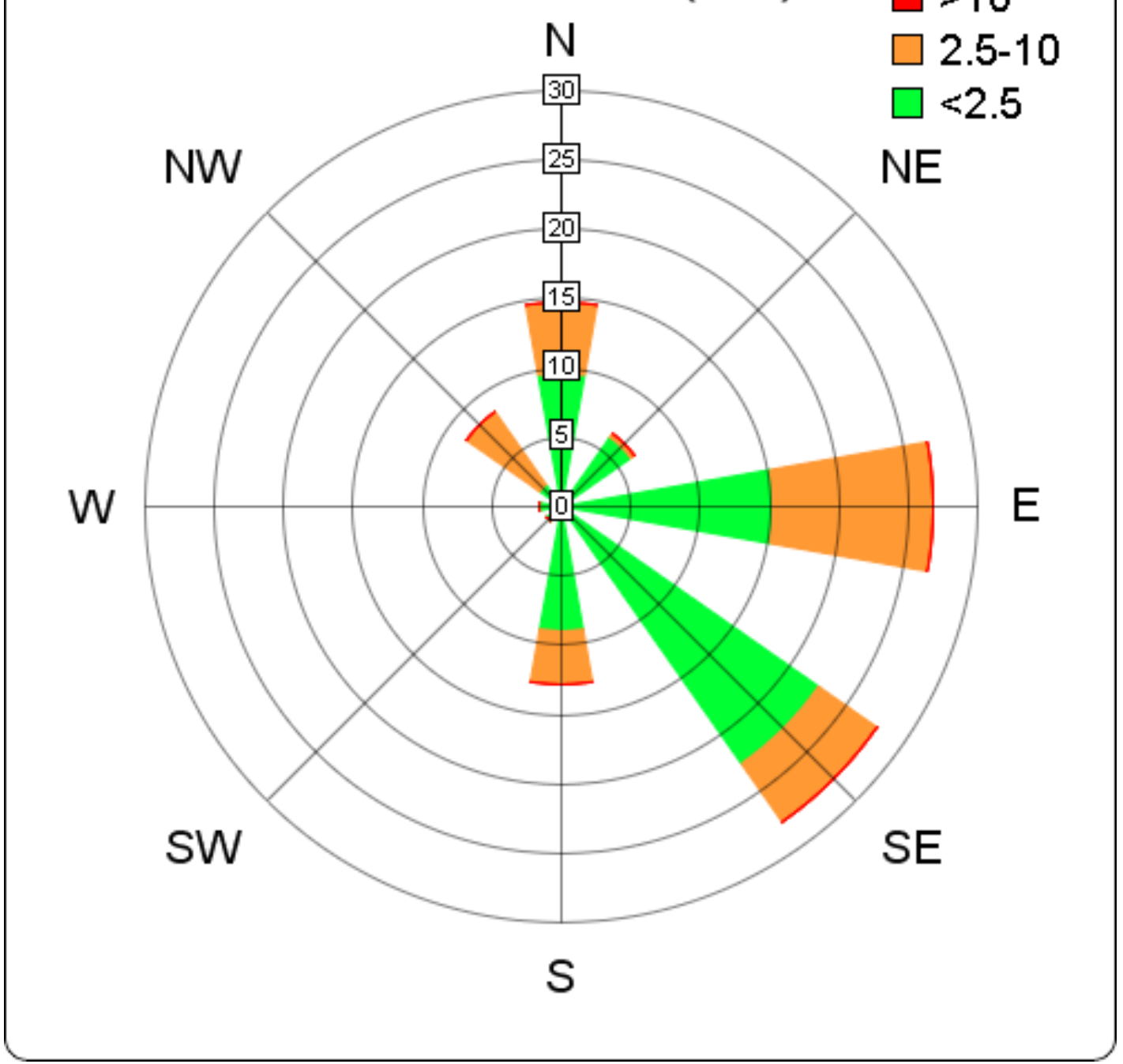
Select Language

Powered by Google™ Translate

Find us on



Waiting for mesowest.utah.edu...


N
NE
E
SE
S
SW
W
NW
30
25
20
15
10
5
0
2.5-10
<2.5

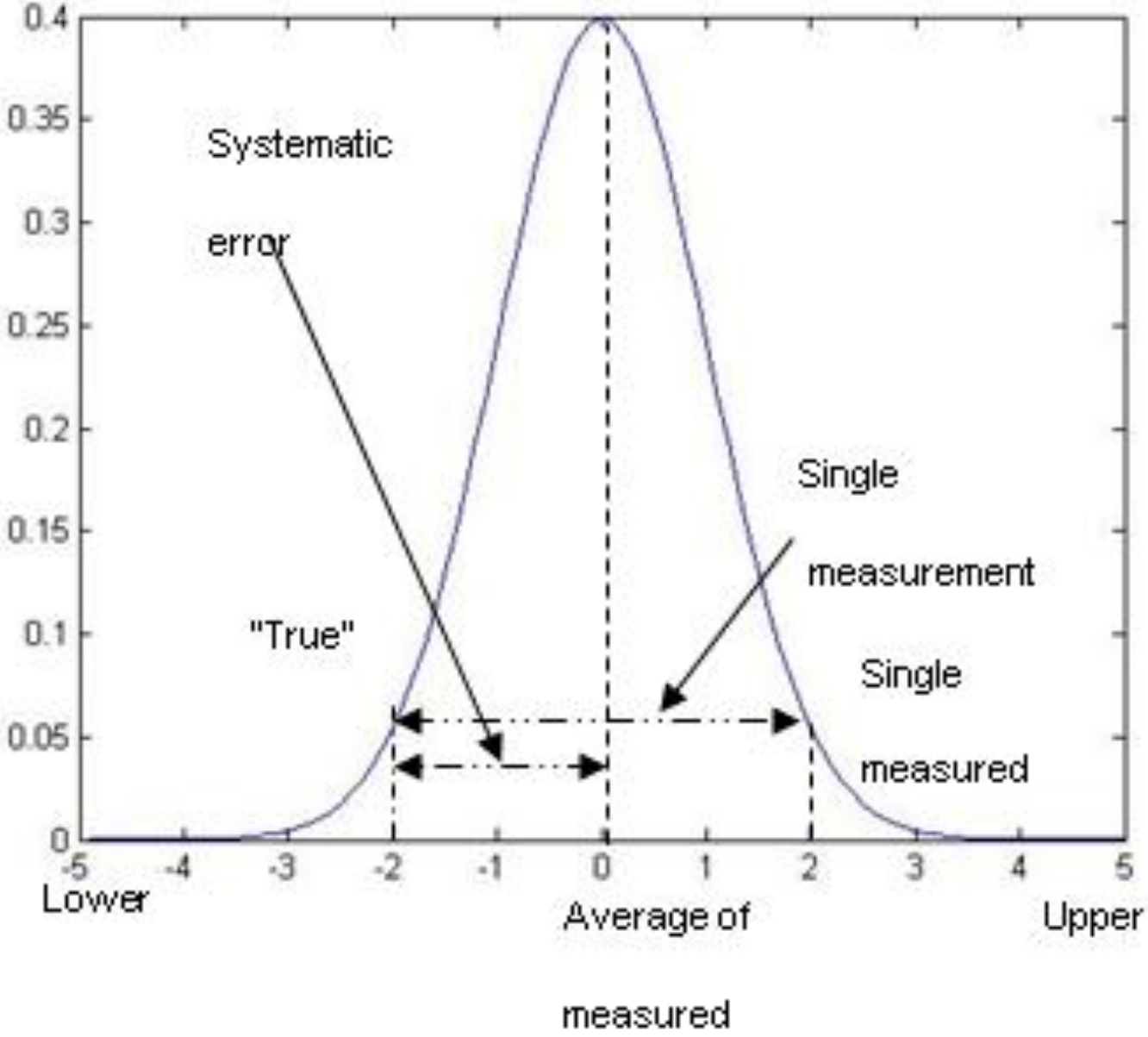

recent observation within 6 hrs ending at 06:00 UTC 1/7/2011
80
Chestnut St
5000'
5200'
5400'
POWERED BY
MesoWest
Map data ©2010 Google - Terms


Tunnel May Be
ICY
When Flashing

Population vs. Sample

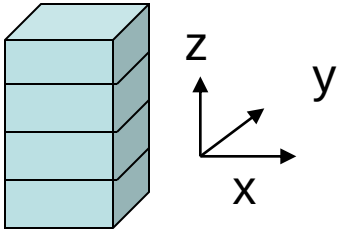
- we never know the entire population of true values as the environmental conditions change in time or space.
- We hope that we choose a sample of observations for analysis such that each element in the population has an equal chance to be selected.
- Sampling issues
  - Trends
  - serial dependence of environmental data
  - model sample tend to be less variable than observed samples

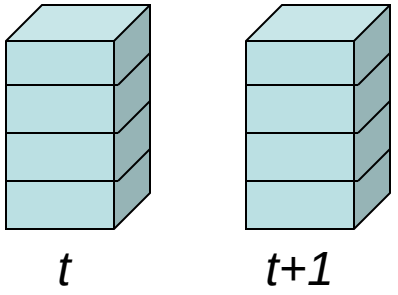
- Selecting the sample for analysis is a critical aspect of organizing the data and depends on the question to be addressed by the study
- rule of thumb: sample should be large enough to capture the phenomenon of interest many times
- “Degrees of freedom”: number of independent elements in the sample;
  - usually much smaller than the total number of members in the sample in environmental data sets
- Keeping your powder dry- saving data for an independent sample to evaluate and confirm your results.
- Tendency to assume sample is drawn randomly from the population, when sample grossly underestimates the variability inherent in the population


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
Systematic
error
"True"
Single
measurement
Single
measured
Lower
Average of
measured
Upper

Large Dimensionality of Geophysical Data Sets

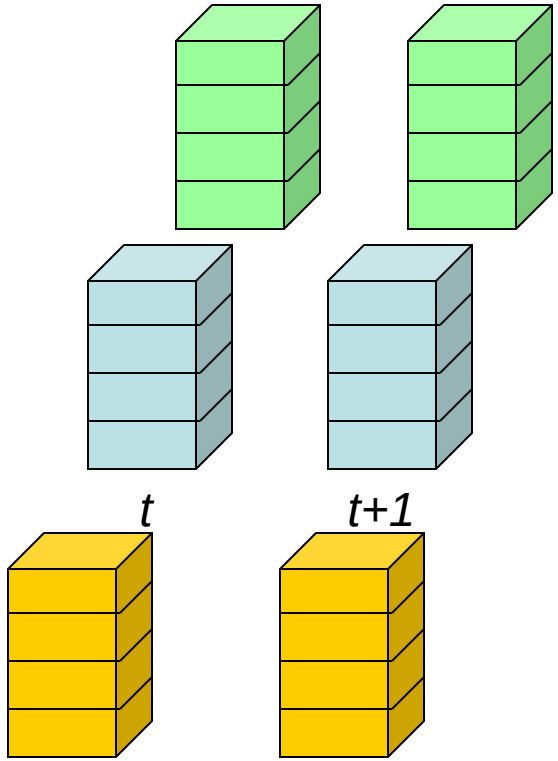
•Space: x, y, z





•Time: time (t) and forecast time (t_f)

•Parameter & Source (temperature, winds, different models, measuring systems, perturbations)



Common Goal is to Synthesize and Reduce Dimensionality

- Statistical analysis of environmental data typically involves reducing the dimensionality of the data to a manageable size.
- Which variable(s) do we need to consider? Can we consider one variable (univariate analysis) or must we consider multiple variables (multivariate analysis). What time scales are we interested in? Hours, days, months, years? And, what region (local, regional, globally) or level in the vertical (surface, subsurface, upper air)?
- Are the data available on a spatial grid or at specific points?

10 Kestrel Temperature Sensors

Diving into some Data

- Going to take 10 kestrels outside for ~15-20 minutes
- ~3 people per kestrel
- 5 different environmental conditions assigned randomly
  - Aspirate the temperature sensor vigorously by swinging the kestrel at a fast rate
  - Aspirate the temperature sensor weakly by swinging the kestrel at a slow rate
  - Hold the kestrel steady away from your body
  - Place the kestrel on an object such as the pavement, bench, or rock
  - Place the kestrel in your pocket, then remove it, then put it back in on the time scale of a few minutes each time
  - Do you have a better idea?
- Data will be downloaded. You won't know which conditions are associated with which record
- Will we be able to detect which ones are which? Stay tuned…