Reproducible Research
Human Health & Environmental Applications

New England Association of Environmental Biologists

Water's Edge Resort & Spa. March 18-20 2009. Westbrook, CT

Hal Walker walker.henry@epa.gov

U.S. EPA - ORD.

NHEERL Atlantic Ecology Division

Narragansett, Rhode Island

How do stressors affect ecosystem services?
How can human well-being be quantified & valued?

Stressor Trends

- 2x CO_2
- 3x Reactive Nitrogen
  - ⇒ Human Health
  - ⇒ Ecosystems
- Ozone & PM-10 (NO_x and SO_x)
- Acid Deposition (NO_x & SO_x)
- Mercury

Additional information:
EPA Ecological Benefits Strategic Plan 2006.
www.epa.gov/economics

Stressors affect ecosystem structure & function, ecosystem services, and human welfare benefits

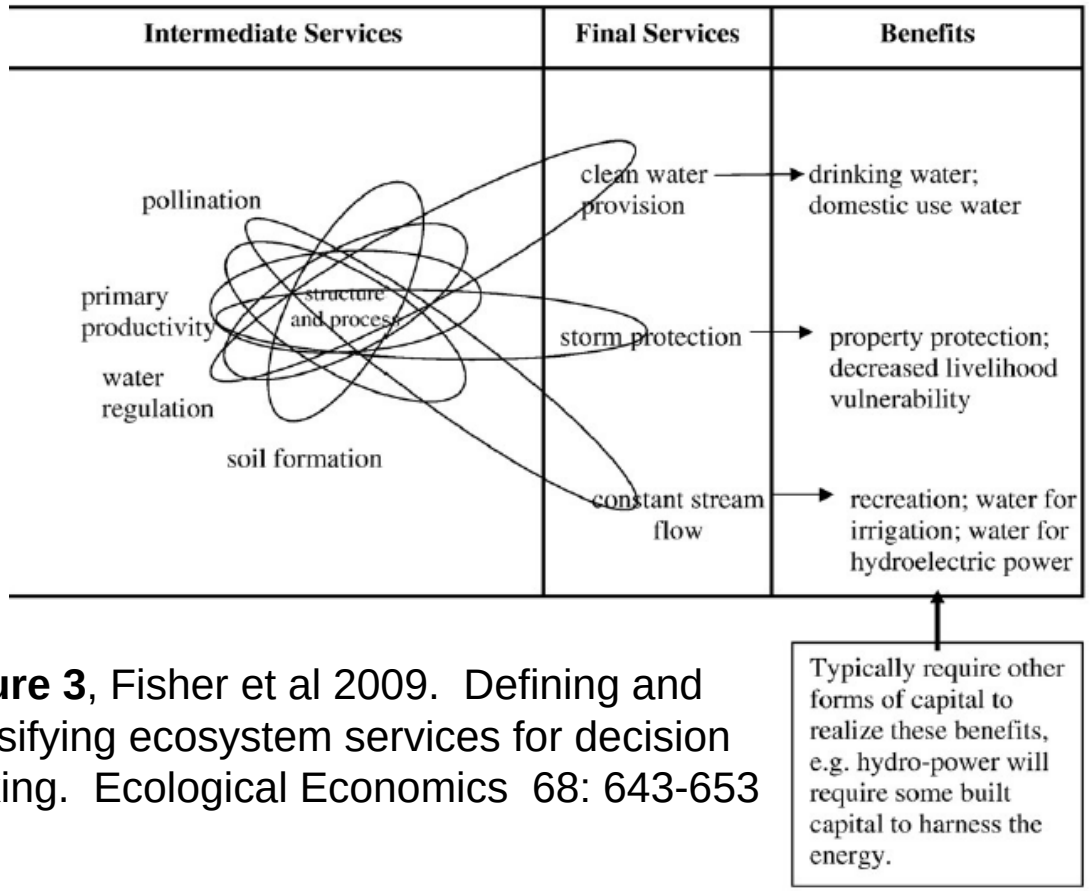


Figure 3, Fisher et al 2009. Defining and classifying ecosystem services for decision making. Ecological Economics 68: 643-653

Need for "reproducibility" in assessments of risks & in valuation of benefits trade-offs.

Environmental & human health risk assessments are based on:

1. data
2. data analysis (including computer modeling)
3. interpretation, and peer reviewed publication

To fully check the analyses, we should be able to:

4. access the data
5. access computer script used in the data analysis
6. be able to reanalyze the data, and check interpretations

Then we can agree or disagree.

Reproducible Research (RR)
Human Health & Environmental Applications

Talk Outline

- Reproducible Research. What is it?
- Two quick examples:
  - Environmental
    - Global Warming (CO_2), RR related to Mike Mann's "hockey stick".
    - Epidemiology (Human Health affected by NO_x & SO_x)

- National Lakes Assessment (preliminary RR analyses)
- NEAEB Mtg / Computational Workshop (RR analyses)

Reproducible Research (RR). What is it?

- "Reproducible Research in Scientific Discovery" R. Gentleman, 2005 http://gentleman.fhcrc.org/Fld-talks/RGRepRes.pdf

- The RR focus is on reproducibility in data analysis, and has its roots in "literate programming"(Knuth 1992), in computer sciences.

- Described a bit more in: "Beyond data: reproducible research in ecology and environmental sciences" Hollister & Walker. 2007. Frontiers in Ecology and the Environment Vol. 5, No. 1, pp.

- Reproducible Research Compendiums: Integrated Collections of text, code, data, and auxilary information (metadata), needed to replicate the data analyses (statistics and graphics).

(Gentleman 2004) http://www.bepress.com/bioconductor/paper2/

Reproducible Research (RR)

Two quick hyperlinked examples:

- Environmental Examples (CO_2, NO_x & SO_x)
  - Global Warming
  - Air Quality and Human Mortality

Global warming, natural variability or huge risk?

Northern Hemisphere Temperature Reconstruction

Mike Mann's Congressional Testimony in 2003

Is 20_{th} century variation natural or anthropogenic ?

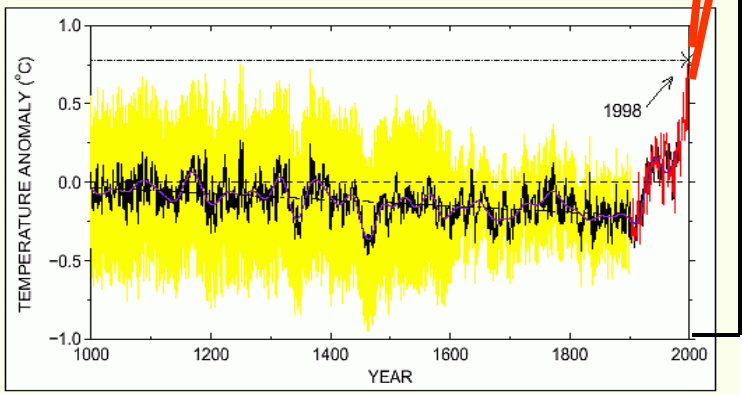


Fig.4 - The `Hockey Stick'

- At scale of the Northern Hemisphere, the "Little Ice Age" wasn't a big deal.
- By the second half of the 20th century, clear evidence of warming due to human activities

http://en.wikipedia.org/wiki/Michael_Mann_(scientist)

http://commons.wikimedia.org/wiki/File:2000_Year_Temperature_Comparison.png

Skeptical Scientists wanted to recalculate and extend analyses farther back

Mike Mann's Congressional Testimony in 2003

Northern Hemisphere Temperature Reconstruction

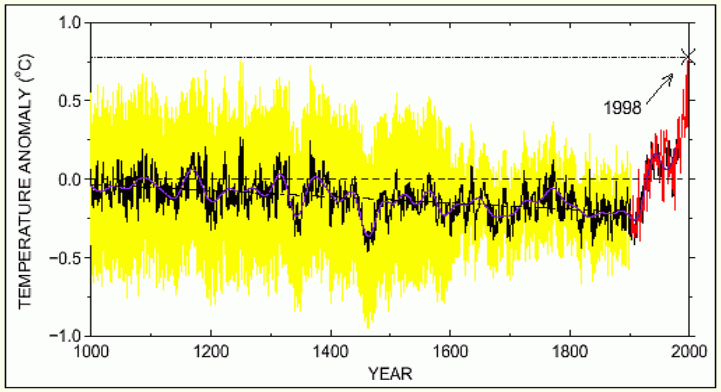
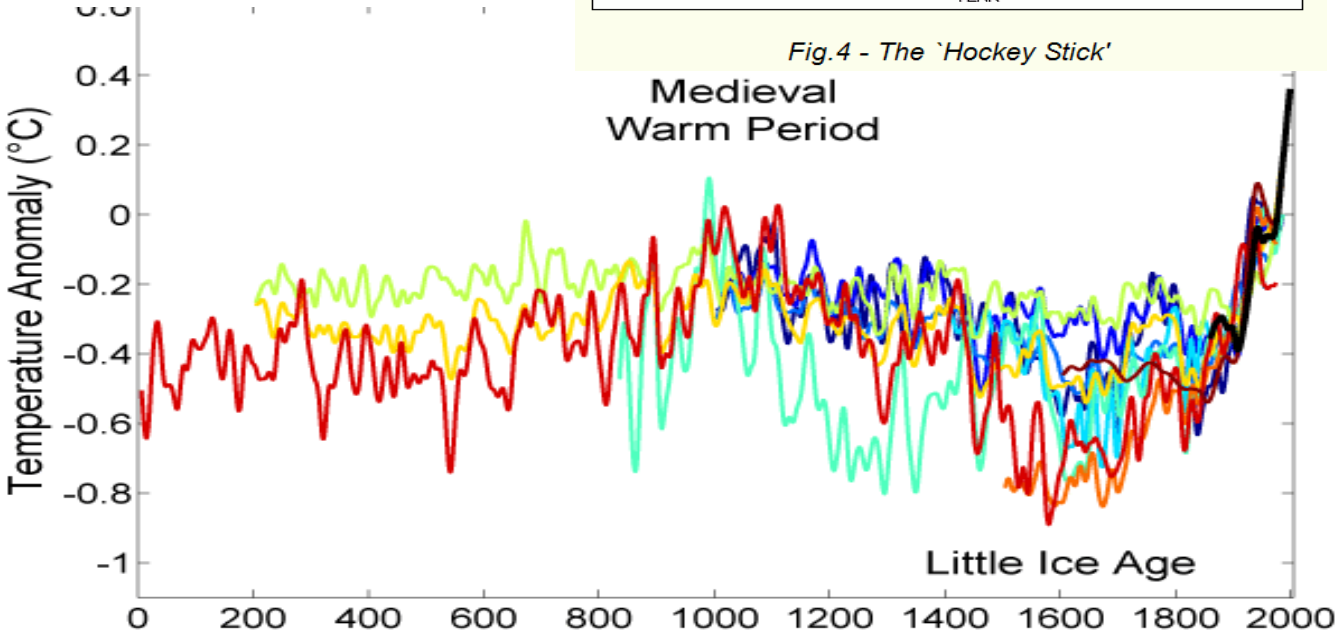


Fig.4 - The `Hockey Stick'



http://en.wikipedia.org/wiki/Michael_Mann_(scientist)

http://commons.wikimedia.org/wiki/File:2000_Year_Temperature_Comparison.png

Scientifically, the warming controversy has settled down as data and code were made available. Public Opinion hasn't settled down.

Reevaluation of the Mann-Bradley-Hughes Climate Reconstruction (1998/1999)

http://www.cgd.ucar.edu/ccr/ammann/millennium/MBH_reevaluation.html

Publications and Associated Supplemental Material :

2 °C by 2100 ?

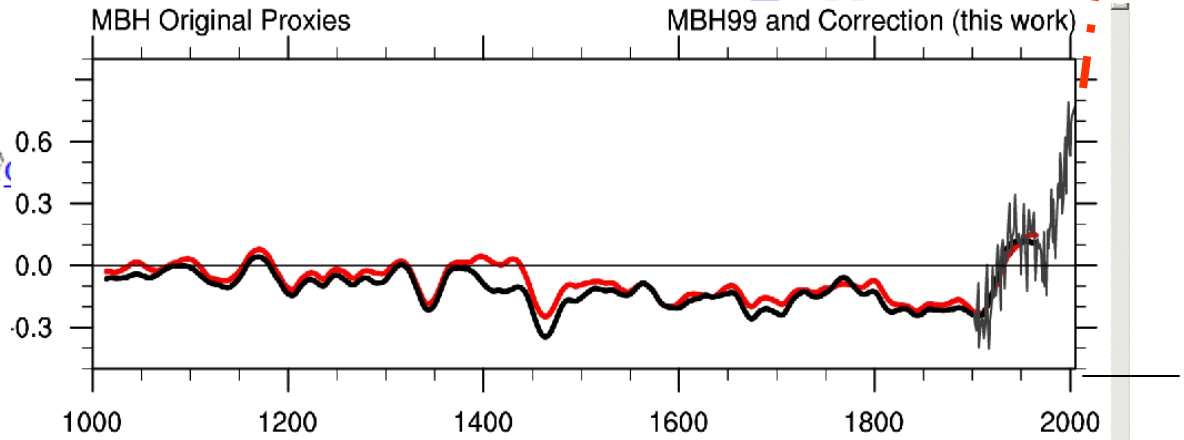
Wahl E.R. and C.M. Ammann, 2007: Robustness of the Mann, Bradley, Hughes Reconstruction of Northern Hemisphere Surface Temperatur
Examination of Criticisms Based on the Nature and Processing of Proxy Climate Evidence. Climatic Change, 85: 85:33-69.

Paper, Code and further Supplemental Material

http://www.cgd.ucar.edu/ccr/ammann/millennium/WA_supplement.html

Ammann C.M. and E.R. Wahl, 2007:
Procedures. Climatic Change, 85:71-88.

Paper and Supplemental Material and

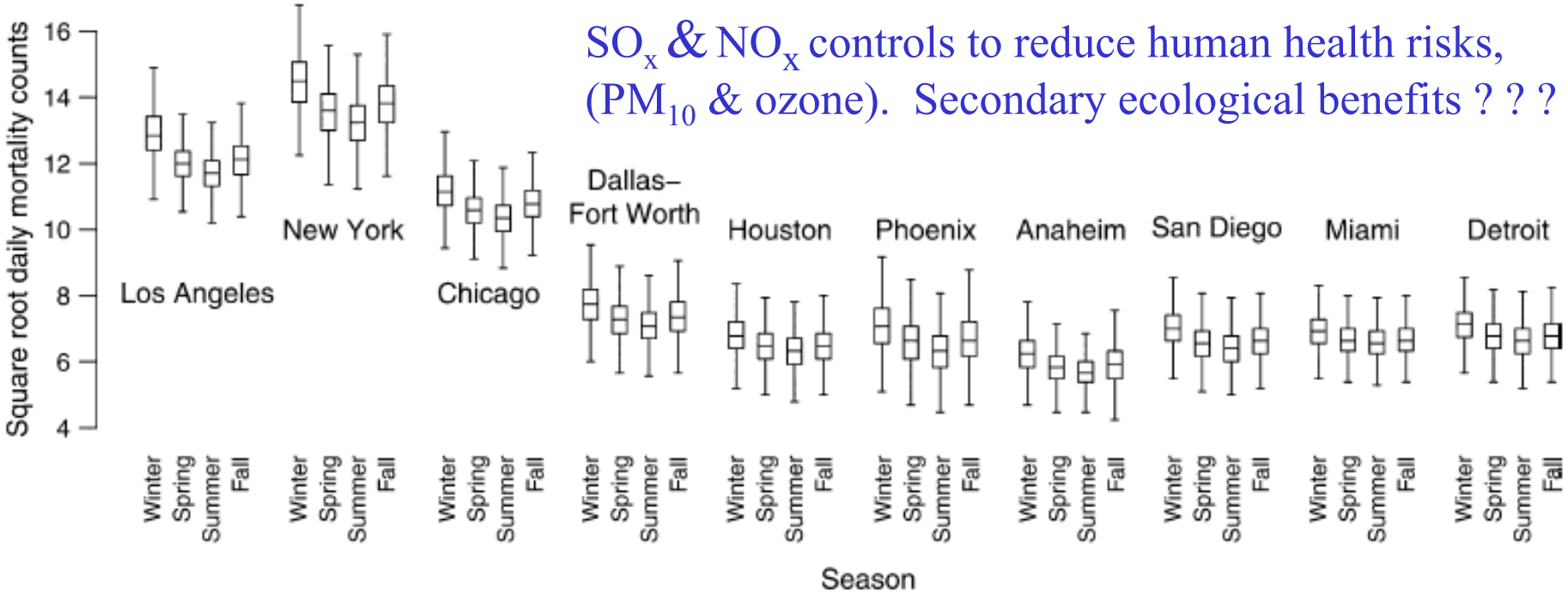


http://www.cgd.ucar.edu/ccr/ammann/millennium/AW_supplement.html

RR Example 2: How does air quality affect human health?

Seasonal Analyses of Air Pollution and Mortality in 100 US Cities

Peng et al 2005 Am J Epidemiol 2005;161:585–594



SO_x & NO_x controls to reduce human health risks, (PM_{10} & ozone). Secondary ecological benefits ? ? ?

FIGURE 1. Box plots of square-root daily mortality for the 10 largest US cities, by season, National Morbidity and Mortality Air Pollution Study, 1987–2000. Horizontal line, median; edges of box, interquartile range (25th and 75th percentiles); bars, 1.5 times the interquartile range.

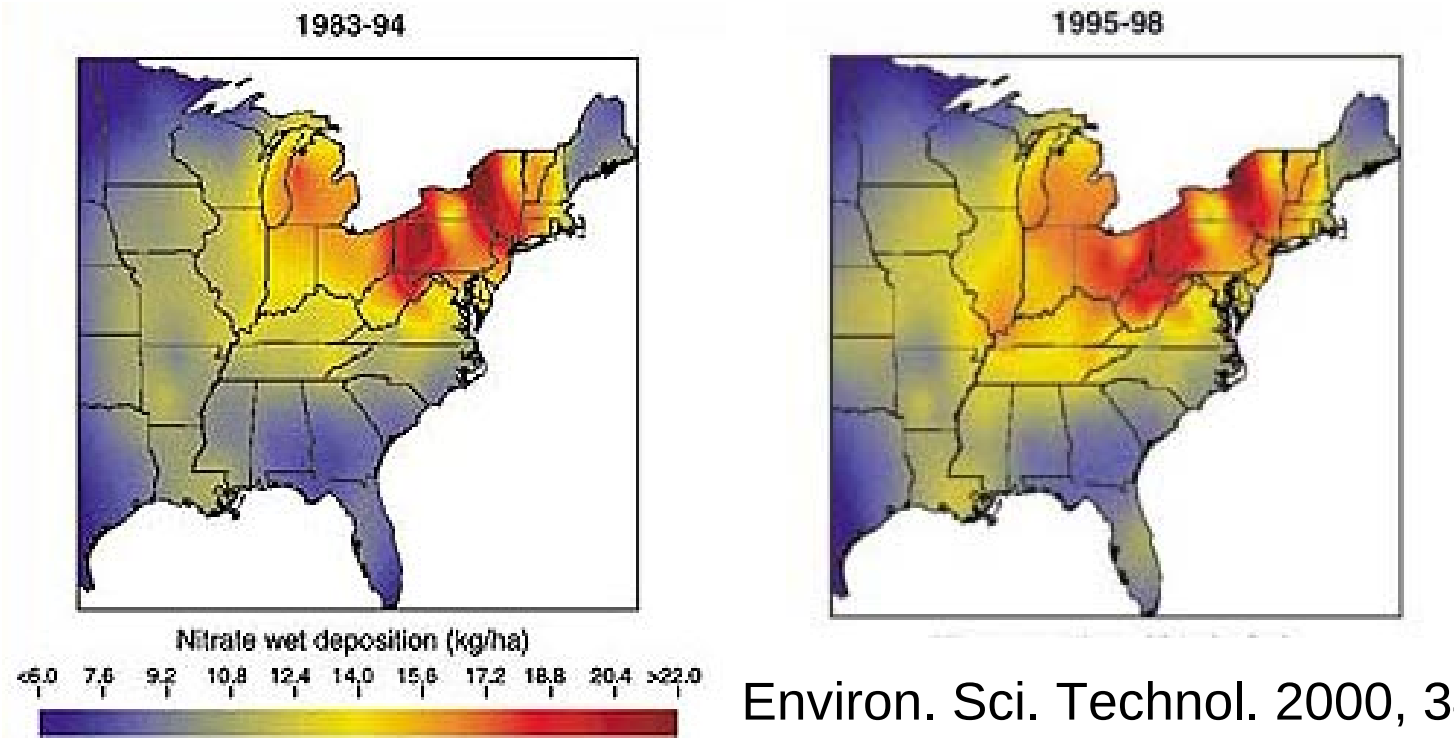
Los Angeles rated most polluted city based on ozone, and fine particulates

http://www.citymayors.com/environment/polluted_uscities.html

Peng's 2005 analysis is reproducible: http://aje.oxfordjournals.org/cgi/reprint/163/9/783

Good News: Clean Air Act ➔ 24% drop in SO_2 emissions
Bad News: Northeastern Acid Lakes Remain Impaired

Deposition of acid-forming nitrogen in the eastern United States remained virtually unchanged from 1989 through 1998



Environ. Sci. Technol. 2000, 34 (11), 248A

NO_x reductions to protect human health, will reduce acid deposition and reduce flux of nitrogen to lakes, streams & estuaries. Need to quantify secondary benefits.

In publications, we are beginning to witness a shift

Old Publication Paradigm

- Hard copy publication (e.g. Science, Nature, ES&T).
  - Amaze your friends, confuse your enemies

To the "reproducible research" community,
this isn't publication, but mere advertisement.

New Publication Paradigm

- Publication of your results in ways that others can recalculate, and refine them. Include: 1) results, 2) data, 3) computational script needed to re-compute statistics and figures.

Goal. Do this with open source or readily available software.

Reproducible Research
Human Health & Environmental Applications

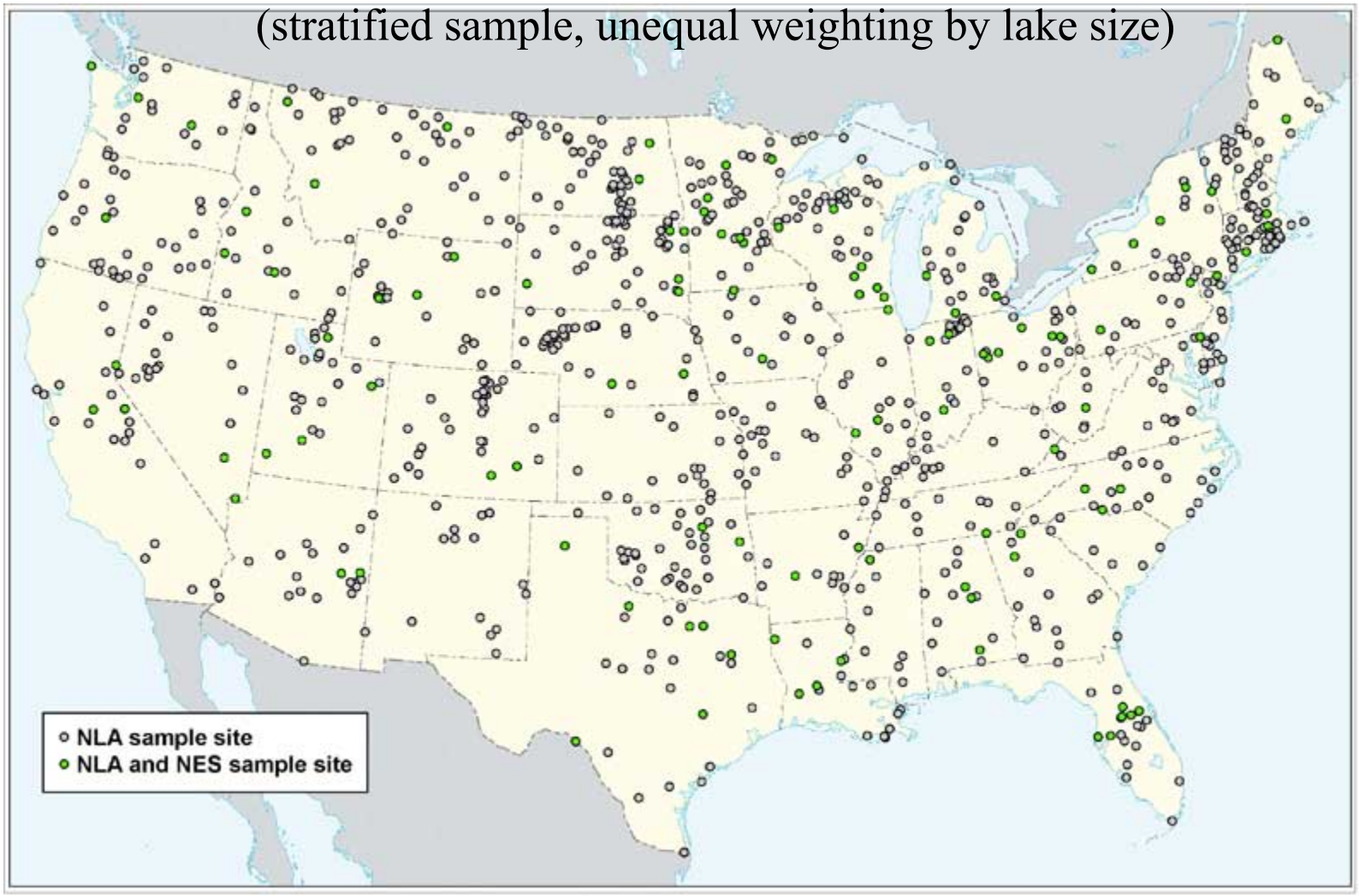
Talk Outline

- Reproducible Research. What is it?
- Examples:
  - Environmental
    - Global Warming (CO_2)
    - Epidemiology (Human Health affected by NO_x & SO_x)

- National Lakes Assessment (preliminary RR analyses)
- NEAEB Mtg / Computational Workshop (RR analyses)

Nation's Lakes Survey (NLS)
for unbiased estimates of conditions in lakes and reservoirs

(stratified sample, unequal weighting by lake size)



http://www.epa.gov/owow/lakes/lakessurvey/siteselect_factsheet.html

Excel Viewer for QA of Draft NLS Water Chemistry Data Set

Lake Depth Categories: Shallow <2.6 m < Medium < 16.3 Medium < Deep.

Select Parameter

Estimated Mean Depth (m)

Select Threshold Type

Percentiles

Table Value:

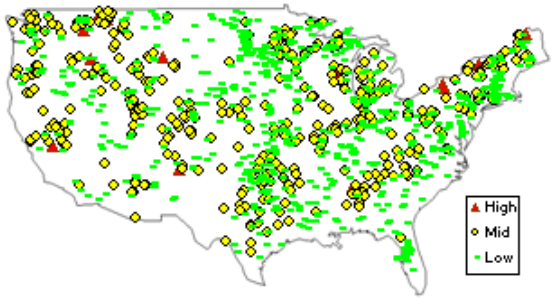
View Threshold Table

Custom thresholds are from Table
TH1 = 2.6 m
TH2 = 16 m

Estimated Mean Depth (m)

Low: < =2.6 m
Mid: 2.6 to 16 m
High: > 16 m

Min:	0.2	m
Median:	2	m
Max:	32	m
Count:	1134	



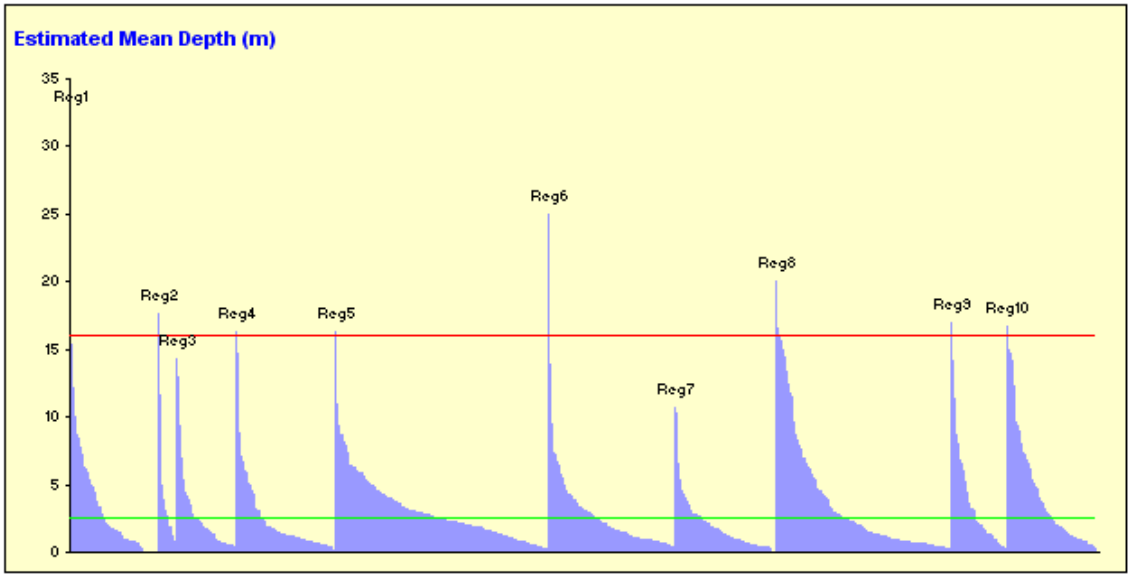
Est_Mean_z meters

Shallow < 2.6

2.6 < Medium < 16.3

Deep > 16.3

Distribution Plot: Shows all values, ordered high to low by EPA Region



Estimated mean depth.
Sorted from highest to lowest by EPA Region

Need to use station weights to estimate proportions of lakes in various categories.

Excel Viewer for QA of Draft NLS Water Chemistry Data Set

Prototype Viewer Demonstrated at NEAEB computational session this PM

Select Parameter

Chlorophyll a(ug/L)

Select Threshold Type

Percentiles

Table Values

View Threshold Table

Custom thresholds are from Table

TH1 = 10 ug/L

TH2 = 40 ug/L

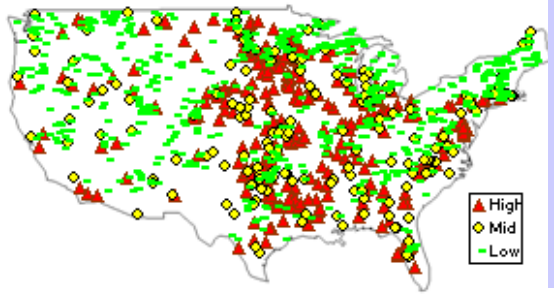
Chlorophyll a(ug/L)

Low: < =10 ug/L

Mid: 10 to 40 ug/L

High: > 40 ug/L

Min:	0.07	ug/L
Median:	7.6	ug/L
Max:	936	ug/L
Count:	1097	



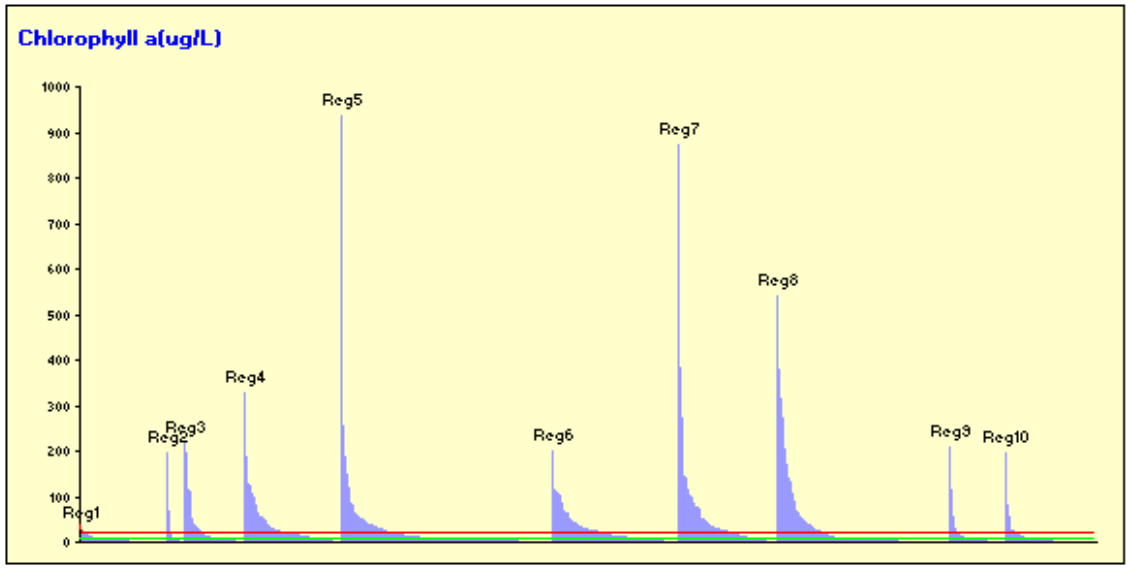
Chl-*a* ug/l

Low < 10

10 < Medium < 40

High > 40

Distribution Plot: Shows all values, ordered high to low by EPA Region



Chl-*a* Observations

Sorted from highest to lowest by EPA Region

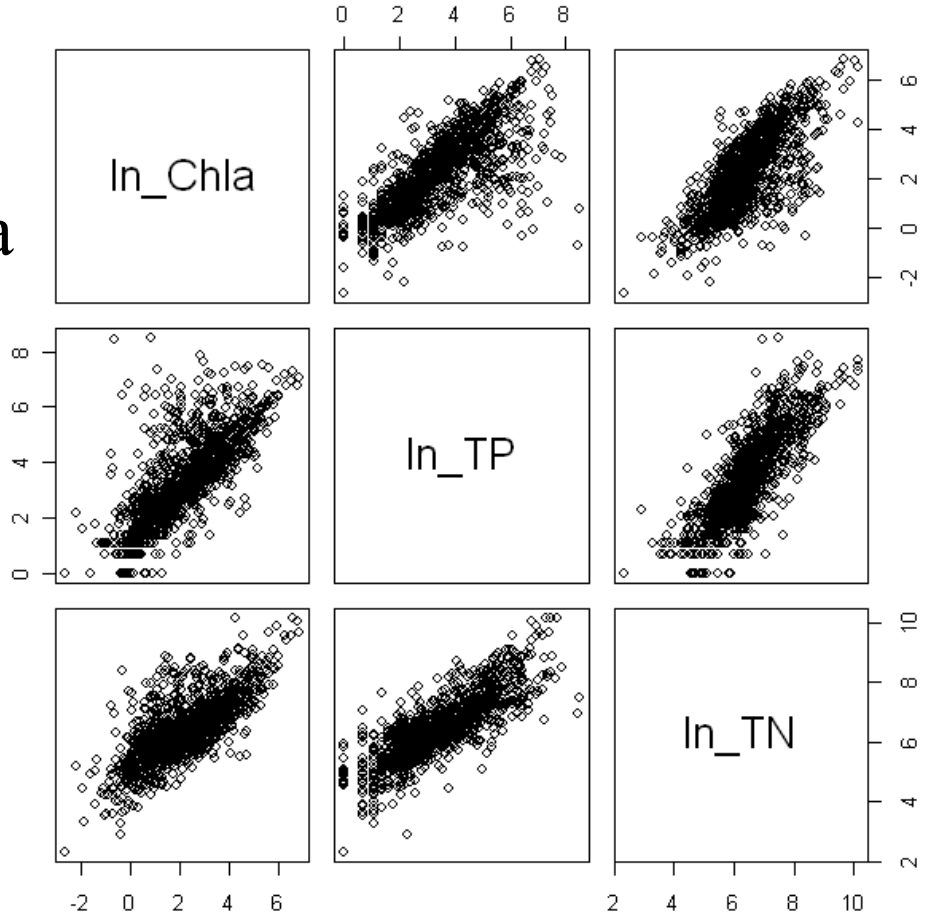
Need to use station weights to estimate proportions of lakes in various categories.

Randomly sampled lakes in EPA Region 1 reflect glaciated, nutrient poor region

Possible regression relationships between Chl-a, TP, and TN

NLS 2007
Water Chemistry Data

(Draft, not fully QA'd)



Possible model form:

$$\ln \text{Chl-}a_i \sim \alpha + \beta_1 \ln \text{TN}_i + \beta_2 \ln \text{TP}_i + \varepsilon_i$$

How do Trophic Status Indices (TSI) vary in Nation's lake & resevoirs?

Chlorophyll-*a*: $TSI(CHL) = 9.81 \ln(CHL) + 30.6$

Total Phosphorus: $TSI(TP) = 14.42 \ln(TP) + 4.15$

Influences on trophic status

- Nutrient Supply
- Climate (seasonality)
- Basin Morphometry
  - Surface area
  - Volume
  - Average depth
- Water residence Time
- Color

Trophic Status (TSI)

TSI's are "descriptive"

Oligotrophic < 40

40 < Mesotrophic < 50

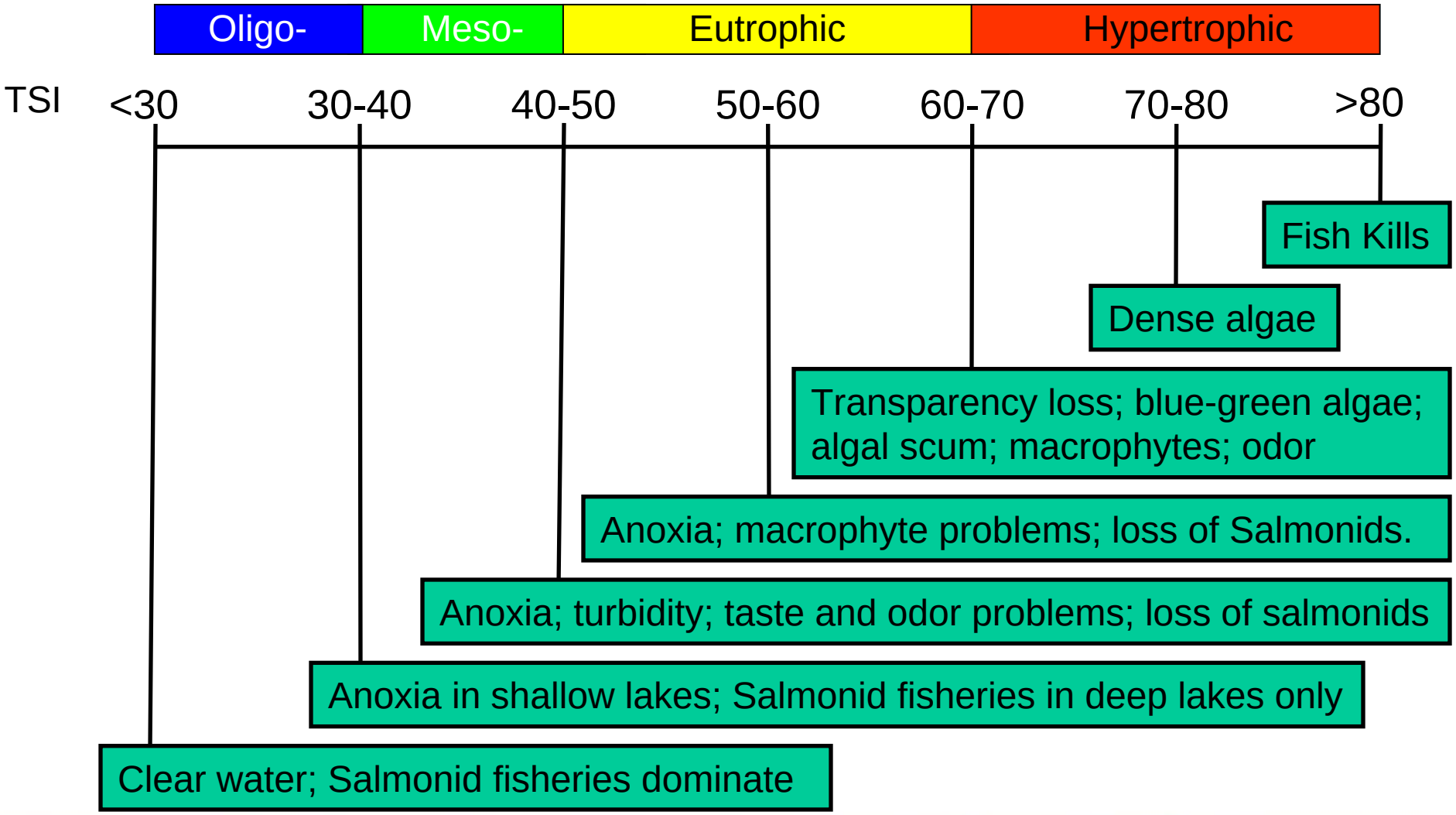
50 < Eutrophic < 70

70 < Hyper-Eutrophic

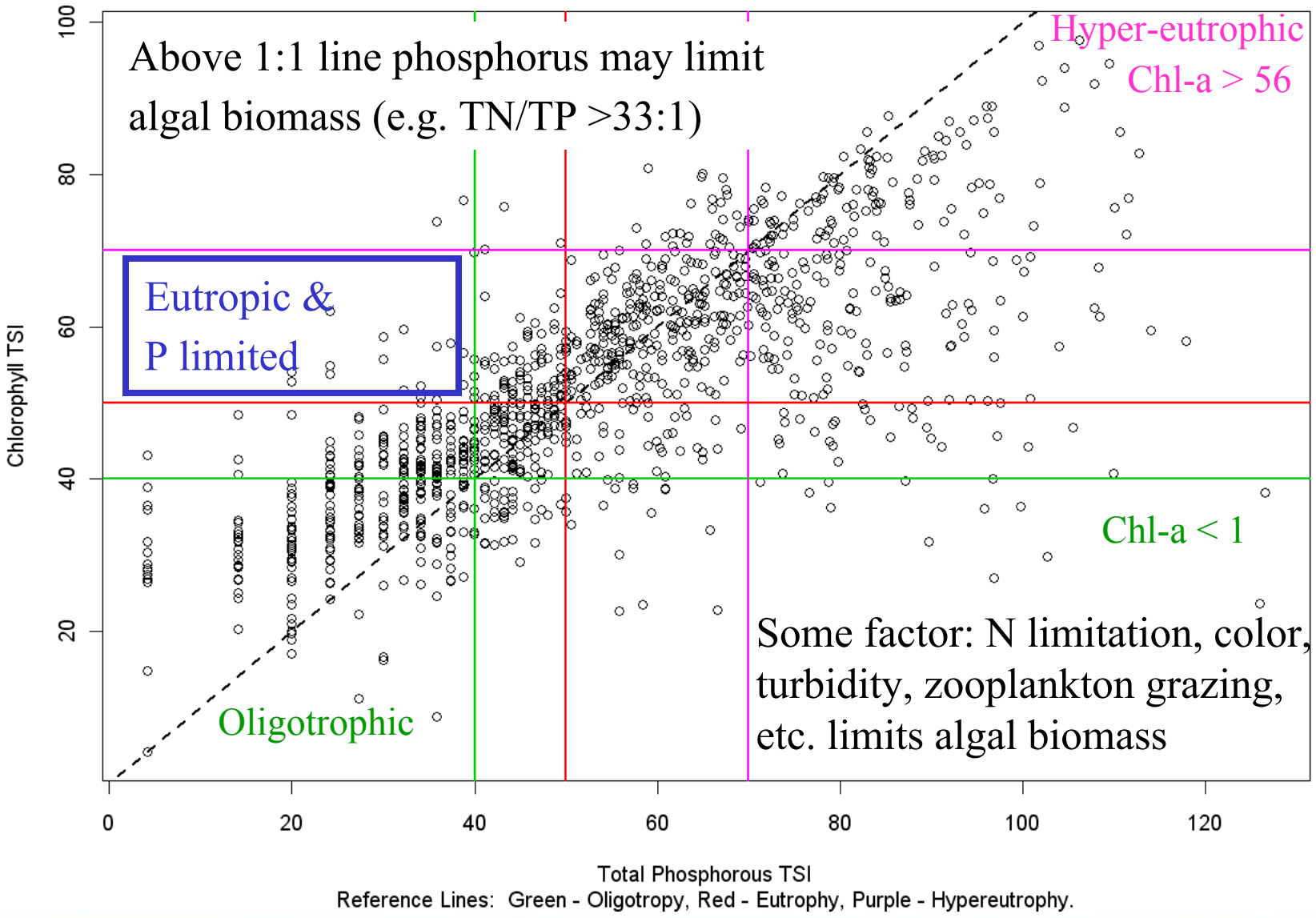
Trophic Status Indicators & possible ecological effects

TSI(TP) = 14.42 ln(TP) + 4.15 or TSI(CHL) = 9.81 ln(CHL) + 30.6

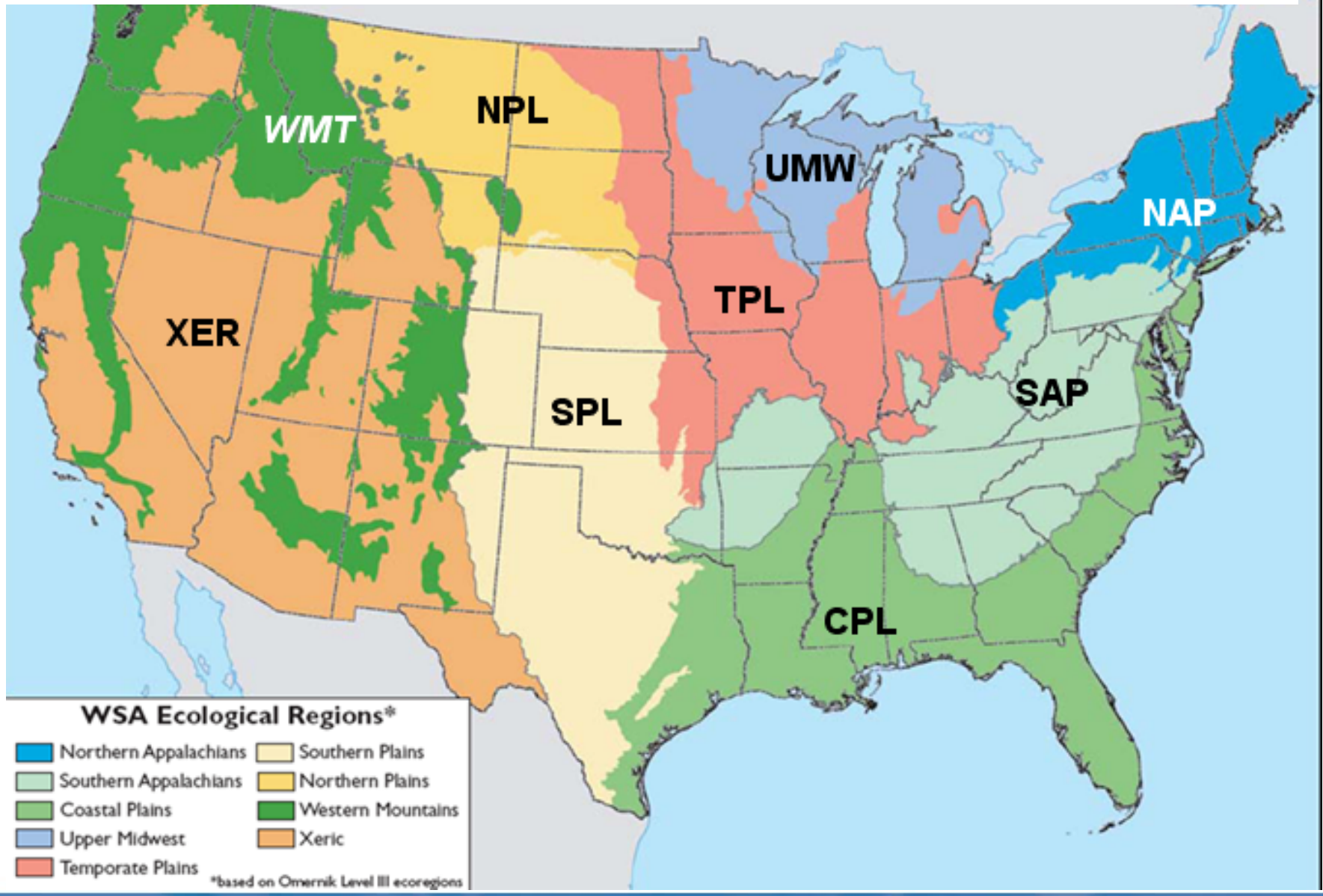
http://castlehs.com/users/johlsen/ecology/A%20Trophic%20State%20Index.doc

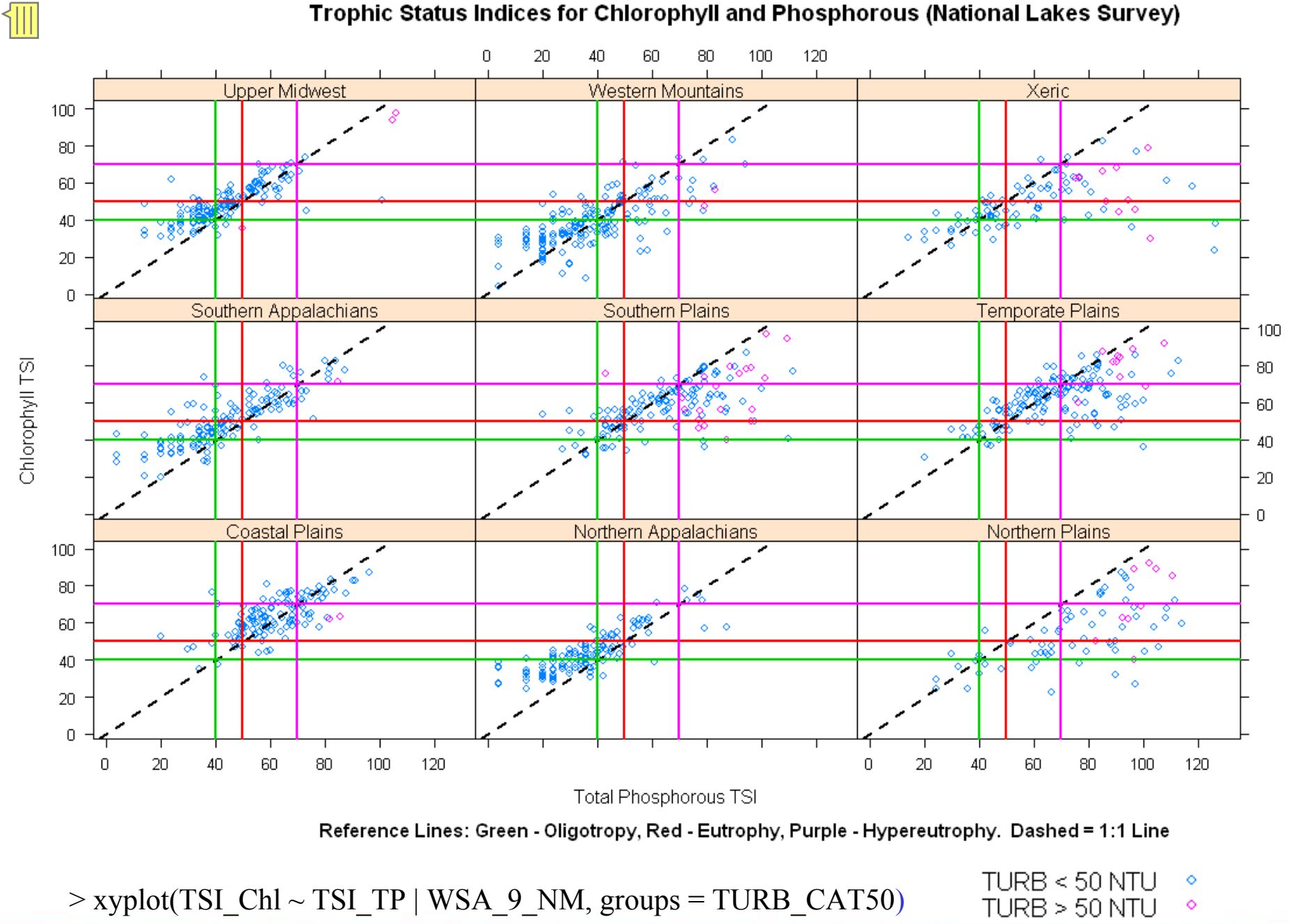


Trophic Status Index for Chl-*a* & Total Phosphorous (draft NLS data)



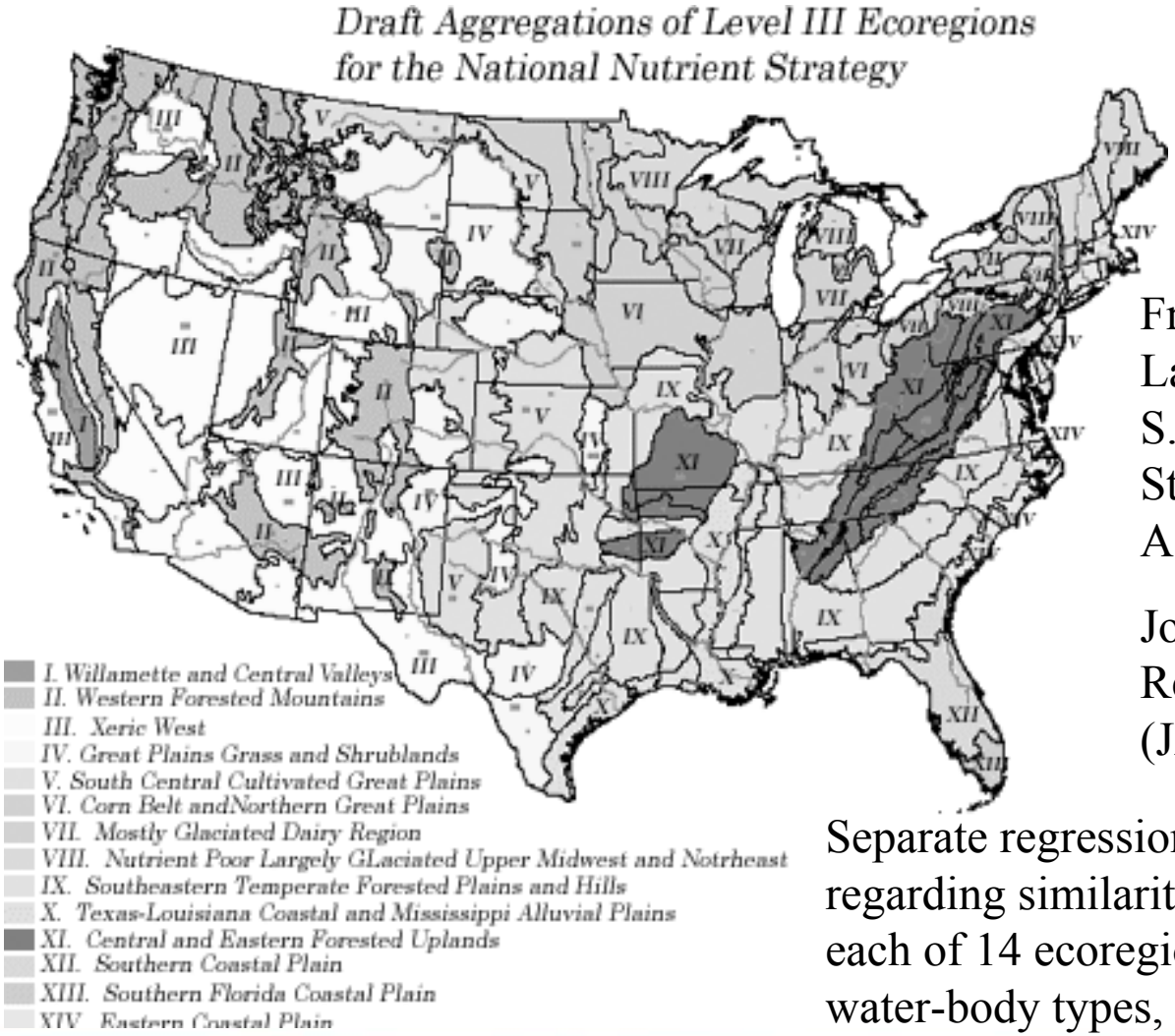
A quick descriptive look at trophic status variations by ecoregion.





> xyplot(TSI_Chl ~ TSI_TP | WSA_9_NM, groups = TURB_CAT50)

Moving from descriptive TSI statistics, to modeling [Chl-*a*] based on [TN], [TP], etc.



From:
Lamon III, E. Conrad and Song S. Qian, 2008. Regional Scale Stressor-Response Models in Aquatic Ecosystems.

Journal of the American Water Resources Association (JAWRA) 44(3):771-781.

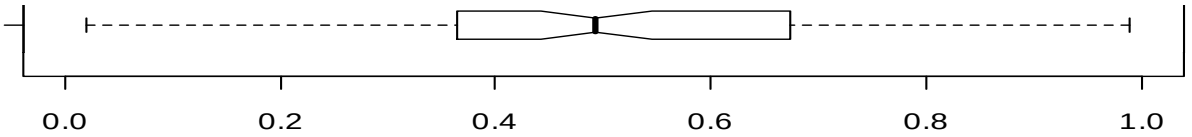
Separate regression coefficients for inference regarding similarities and differences between each of 14 ecoregions, and between the two water-body types, lakes/ponds and reservoirs.

Using Lamon and Qian (2008) regressions for 14 nutrient ecoregions & observed TN & TP from NLS to predict log (Chl-*a*)

Nation Nutrient Strategy: Nutrient Region 8 = the nutrient poor, glaciated upper Midwest and Northeast

Predicted distribution log (Chl-*a*) = X

compared to observed log(Chl-*a*) = x

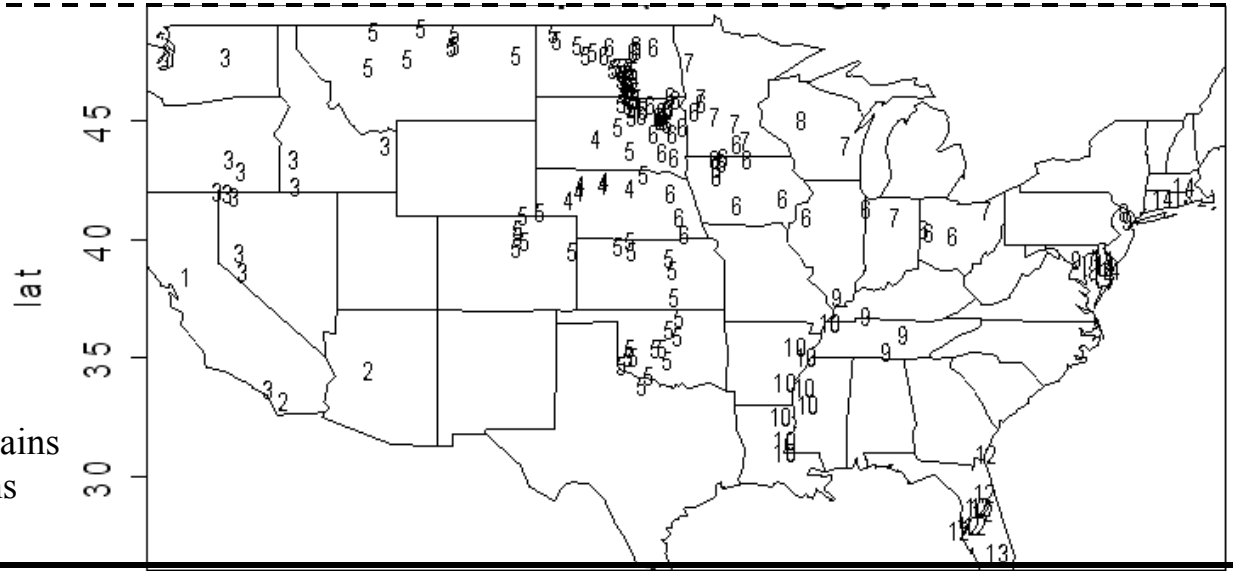


NLS predicted vs observed log(Chl-*a*) Pr (X < x)

Prob >= 50 % Lake (Chl-*a* >= 40 Ug/L)

Numbers 1 thru 14 = National Nutrient Strategy "aggregated level III ecoregions"

Lamon and Qian 2008 equations (unadjusted), using NLS [TN], [TP] to predict locations with prob > 50% that Chl-*a* >= 40 ug/l.



5 = South Central Cultivated Great Plains
6 = Corn Belt & Northern Great Plains
7 = Mostly Glaciated Dairy Region

Personal Communication: E. Conrad Lamon III, Statistical Ecology Associates LLC Canyon Lake, TX Phone: 830-964-3155. Uses open source R, and WinBugs http://en.wikipedia.org/wiki/WinBugs

Another example: Chlorophyll-*a* standards for Finnish Lakes

(x): Excellent = 4 ug/l, Good = 10 ug/l, Satisfactory = 20 ug/l, Poor = 50 ug/l.

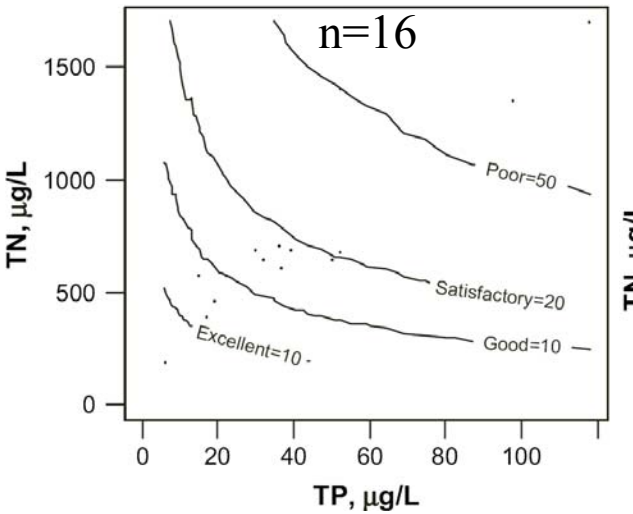
Prob (growing season average Chl-*a* ug/l will be > x) < 10 %

Used a Bayesian TREED Regression (Chipman et al 2002)* to partition the data based on: geographic, morphometric, and chemical properties of lakes.

* Open source algorithm: Chipman et al 2002. Mach. Learn. 48, 299-320

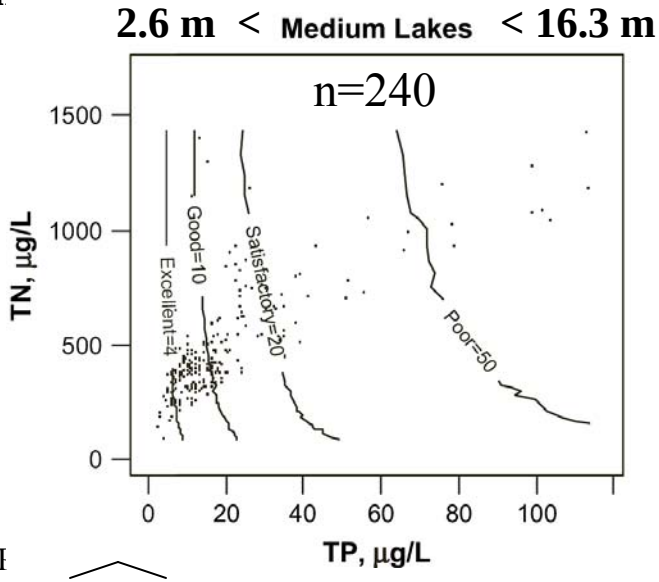
Model 1:

Shallow Lakes < 2.6 m



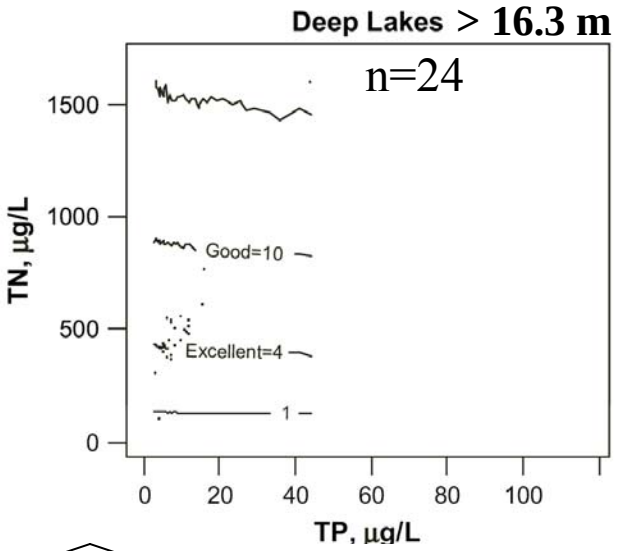
$\widehat{\ln Chla} = 0.074 + 0.646\ \ln TN + 0.418\ \ln TP$

2.6 m < Medium Lakes < 16.3 m



$\widehat{\ln Chla} = -0.002 + 0.128\ \ln TN + 0.651\ \ln TP$

Deep Lakes > 16.3 m

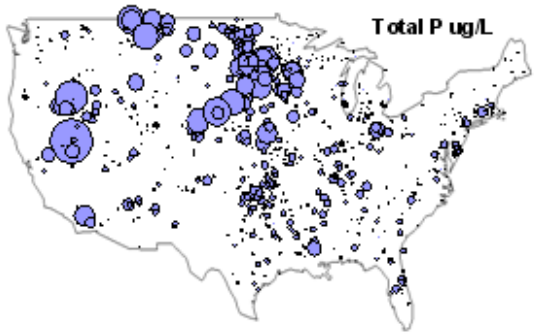


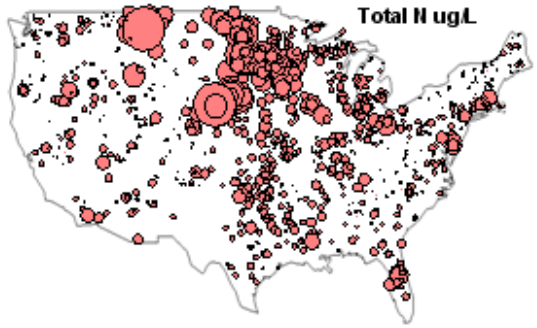
$\widehat{\ln Chla} = -0.129 + 0.646\ \ln TN + 0.027\ \ln TP$

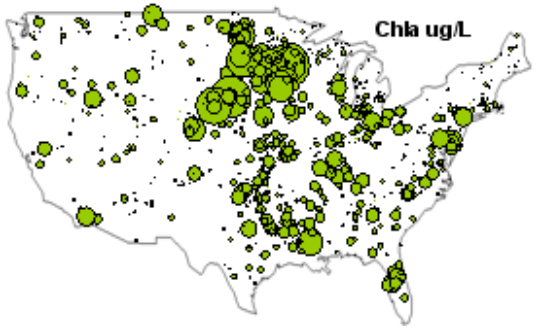
E.C. Lamon III et al. / Environmental Modeling & Software 23 (2008) 938-947

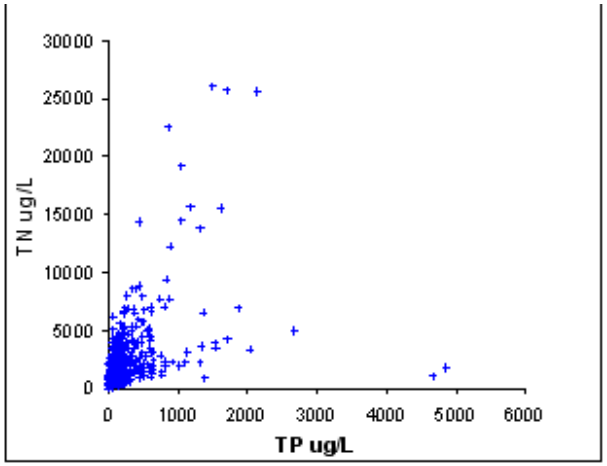
Finnish lakes results might be useful in analysis of U.S. lakes.

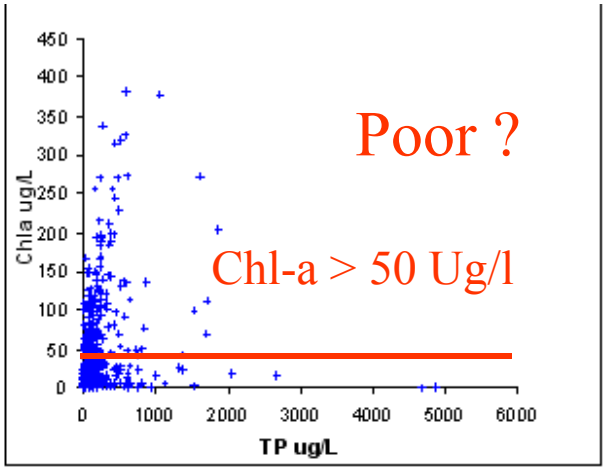
TP, TN & Chl-a in National Lake Survey (NLS). All Lakes

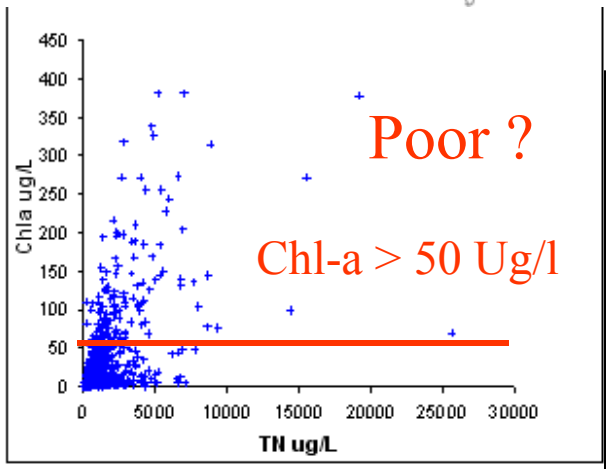










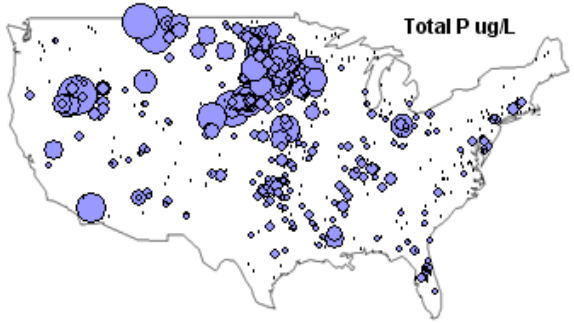


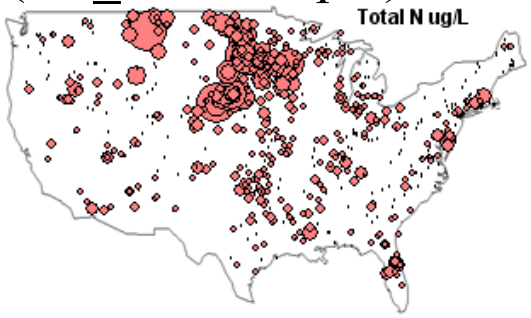
What are the spatial patterns in three depth categories ?

Shallow < 2.3 m < Medium < 16.3 m < Deep

Spatial patterns: TP, TN, & Chl-*a* for three lake depth categories

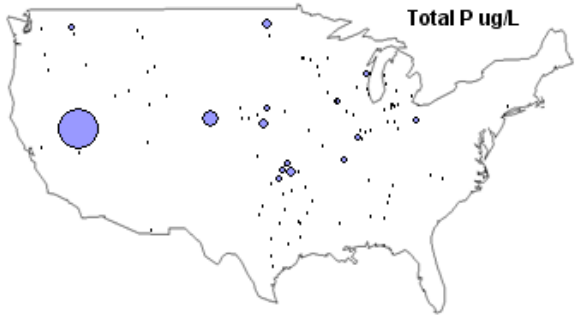
Shallow (Est_Mean Depth) < 2.6 m

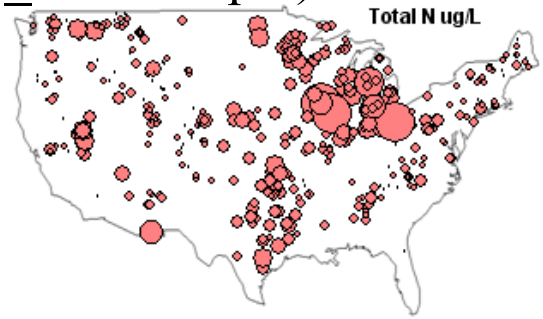




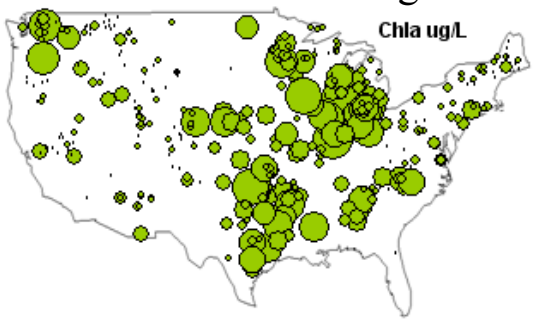
Chl-*a* 1.0 - 400.0 ug / l

2.6 m < Medium (Est_Mean Depth) = < 16.3 m

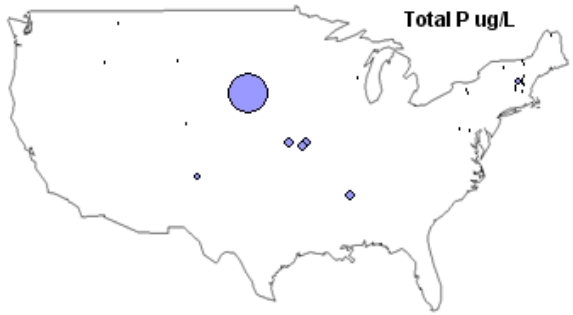


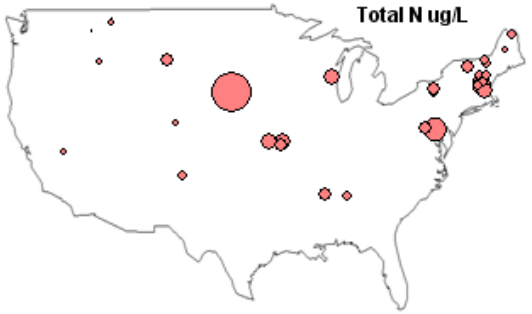


Chl-*a* 1.0 - 100.0 ug / l

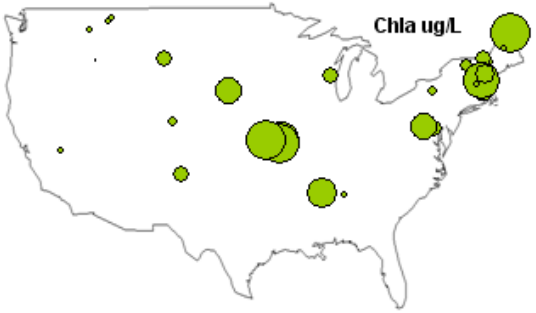


Deep (Est_Mean Depth) > = 16.3 m





Chl-*a* 0.66 - 17 ug / l

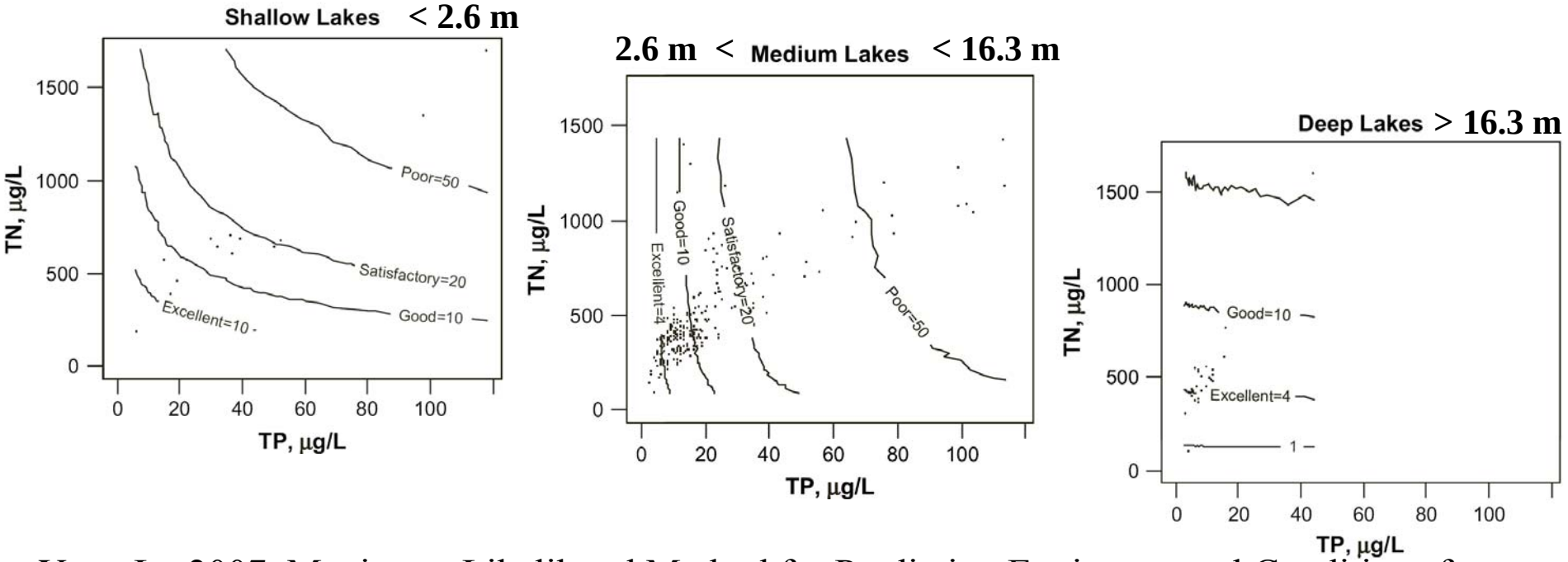


For different lake types, how to quantify & value benefits trade-offs?

How do changes in land-use relate to [TN], [TP] and [Chl-*a*]. What are human welfare benefits trade-offs?

=> Changes in land-use related to [TN], & [TP] and [Chl-a] and benefits trade-offs

=> Shifts in diatom assemblages (top and bottom of cores) could document Δs.



Yuan L. 2007. Maximum Likelihood Method for Predicting Environmental Conditions from Assemblage Composition: The R Package bio.infer Journal of Statistical Software, vol. 22, issue i03

http://econpapers.repec.org/article/jssjstsof/22_3Ai03.htm

This can get complicated. Good to plan for reproducibility.

For effective technology transfer of computational methods, need for to plan for access to: results, data and script.

Most scientific research is based on:

1. data
2. data analysis
3. interpretation, leading publication (aka advertisement)

To make more rapid research progress, we should be able to:

4. access the data used in publications
5. access computation script to replicate the data analyses
6. be able to re-compute statistics and graphics

Then we can adapt and refine the analyses.

Reproducible Research
Human Health & Environmental Applications

Talk Outline

- Reproducible Research. What is it?
- Examples:
  - Environmental
    - Global Warming (CO_2)
    - Epidemiology (Human Health affected by NO_x & SO_x)
- National Lakes Assessment (preliminary analyses)
- NEAEB Mtg / Computational Workshop (RR analyses)

NEAEB Mtg Computational Session

Conditional probability and change point analysis.
Data visualization and analysis with Excel Viewer, GIS Servers, and EDAS2

01:00 - 03:00	**Session 5B (Salon B)** **COMPUTATIONAL SESSION I**
01:00 - 01:30	*Conditional probability analysis and change-point analysis: Background and Introduction. John Paul USEPA RTP NC*
01:30 - 02:00	*Use of Conditional Probability in development of Nutrient Criteria in VT Lakes. Neil Kamman United States Environmental Protection Agency, Office of Water*
02:00 - 02:30	*Setting Sediment Benchmarks Using Multiple Analytical Methods. Ben Jessup TetraTech*
02:30 - 03:00	*Finding the Story in the Data: Using Excel to Visualize and Analyze Lake and Estuarine Monitoring Data. John Kiddon USEPA, AED*
03:00 - 03:30	
03:30 - 05:00	**Session 6B (Salon B)** **COMPUTATIONAL SESSION II**
03:30 - 04:00	*Using CProb in R and Excel to conduct conditional probability analysis. Jeffrey W. Hollister USEPA, AED*
04:00 - 04:30	***GIS servers for visualization & computation. Jane Copeland SRA***
04:30 - 05:00	*EDAS2, Environmental Data Analysis System. Jeffrey S. White TetraTech*

Acknowledgements

- Ernie Pizzuto for encouraging me to give this talk.
- Bryan Milstead's preliminary thinking relating conditions in New England lakes, to ecosystem services & benefits
- EPA OW, and States participating in the National Lakes Survey (NLS)
- John Kiddon,EPA Atlantic Ecology Division (Quick Excel views) Calculations with NLS data demoed in NEAEB session this afternoon
- E. Conrad Lamon III, for some preliminary follow-on analyses related to "Regional Scale Stressor-Response Models in Aquatic Ecosystems" Lamon and Qian (2008) JAWRA 44(3):771-781 using draft NLS data.

Reproducible Research
Human Health & Environmental Applications

New England Association of Environmental Biologists

Water's Edge Resort & Spa. March 18-20 2009. Westbrook, CT

Hal Walker walker.henry@epa.gov

U.S. EPA - ORD.

NHEERL Atlantic Ecology Division

Narragansett, Rhode Island