

# Expanding the Assessment Toolbox: Use of Statistical and Data-driven Techniques for Source Identification and Predictive Modeling

Diane M.L. Mas, Ph.D.

NEIWPCC Nonpoint Source  
Pollution Conference

May 2007



**FUSS & O'NEILL**  
*Disciplines to Deliver*

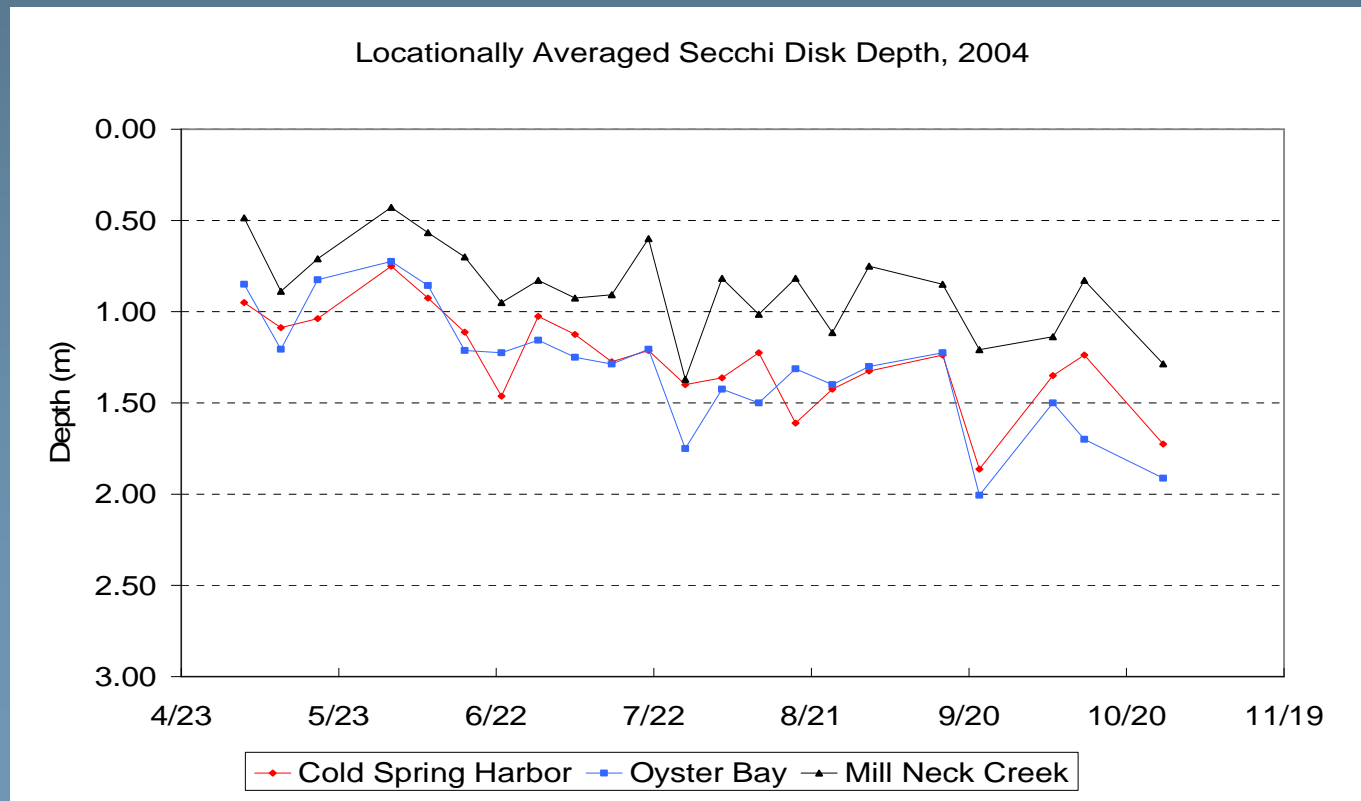
# Monitoring and Modeling

- Identifying potential sources of pollution
  - *Pathogens*
  - *Toxicity*
  - *Nonpoint Sources*
- Predicting concentrations of pollutants, especially relative to some regulatory threshold
  - *Ambient water quality standards*
  - *Recreational water quality standards*
  - *Drinking water standards*



# Monitoring Data

- Assess conditions
- Share Information

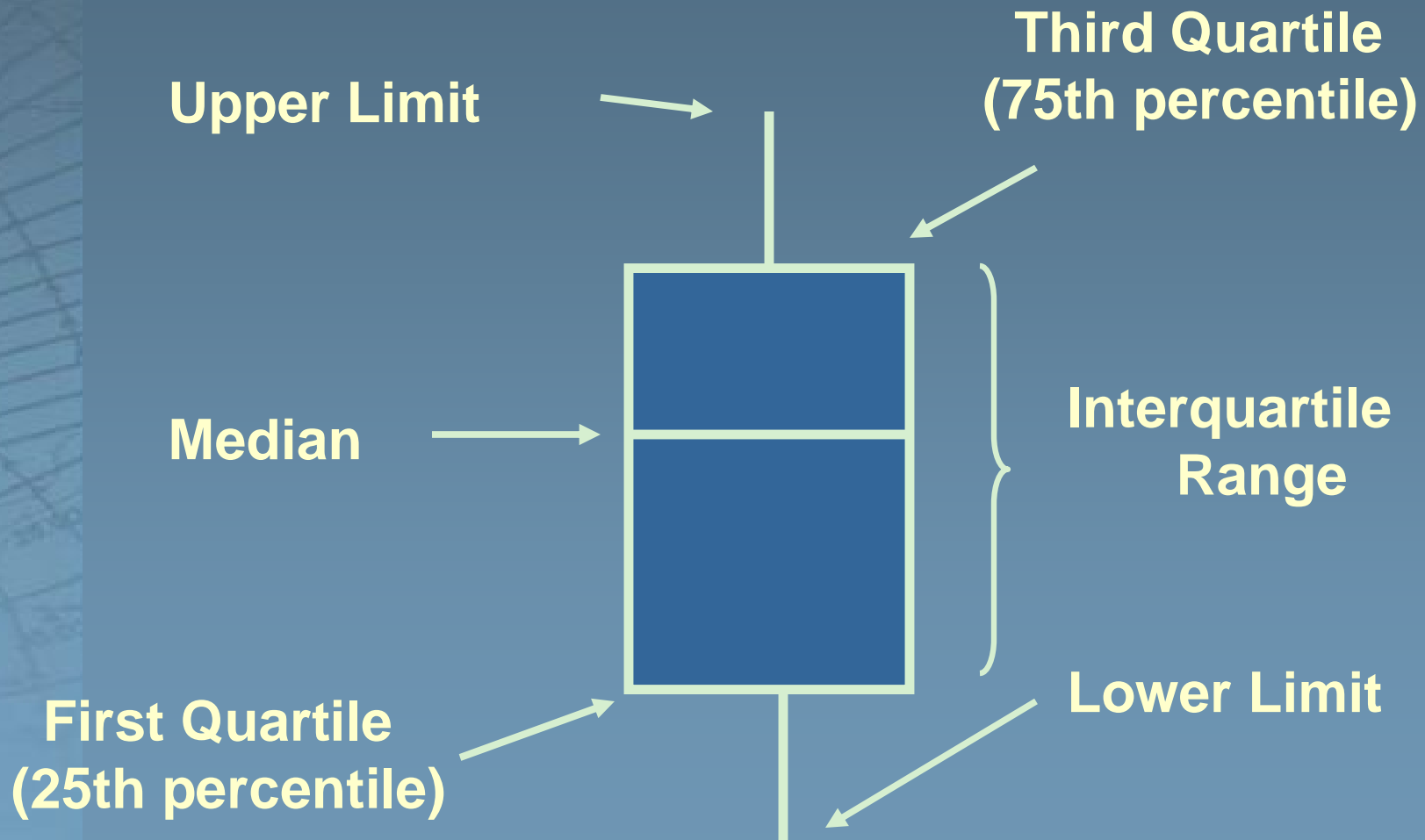


# Data Analysis and Presentation

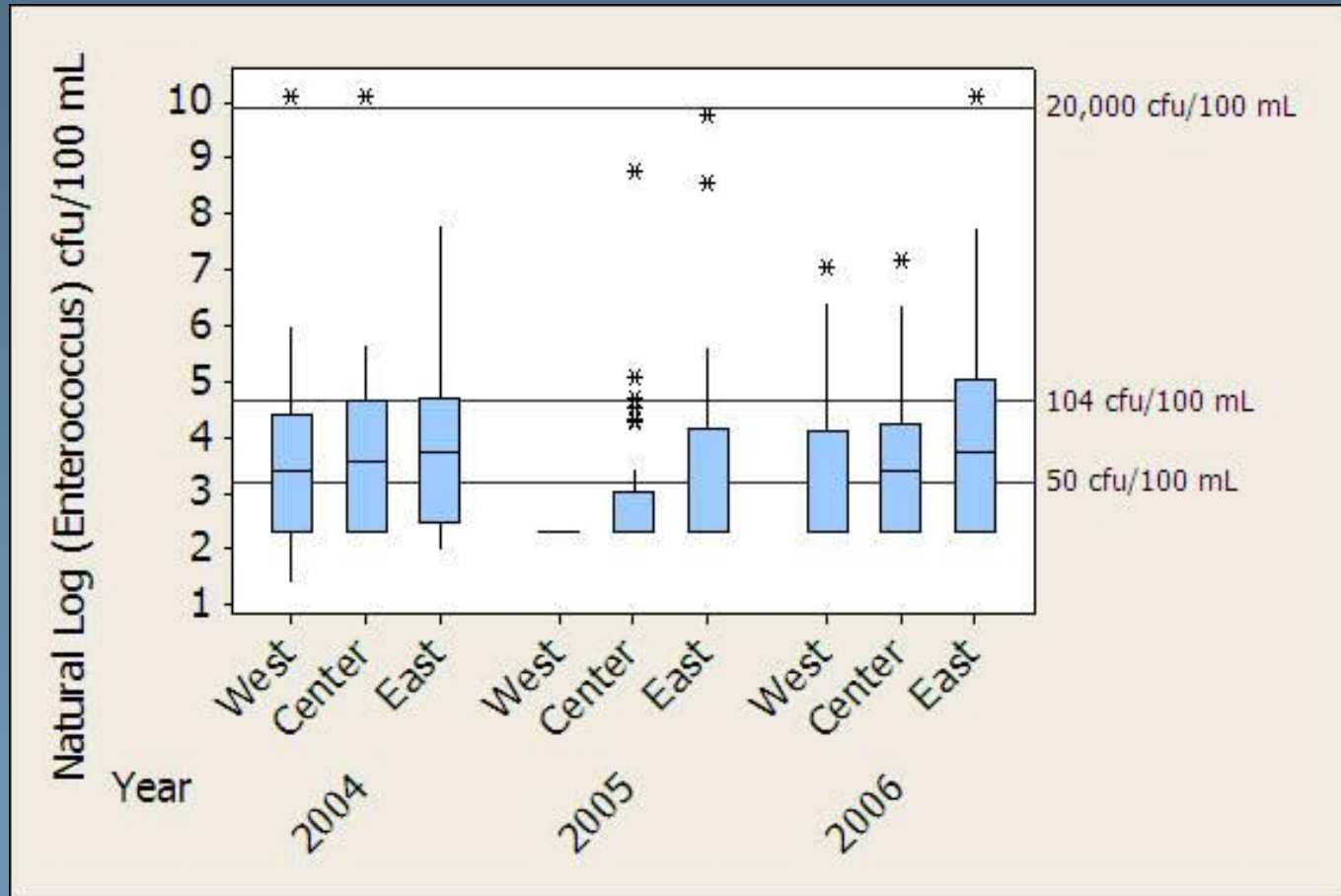
- Boxplots
  - *Bacteria (Newport, RI)*
- Non-linear correlation
  - *Toxicity (Baltimore-Washington Airport)*
- LOWESS
  - *Toxicity (CT Stormwater)*
- Dealing with detection limits
  - *Bacteria (Newport, RI)*



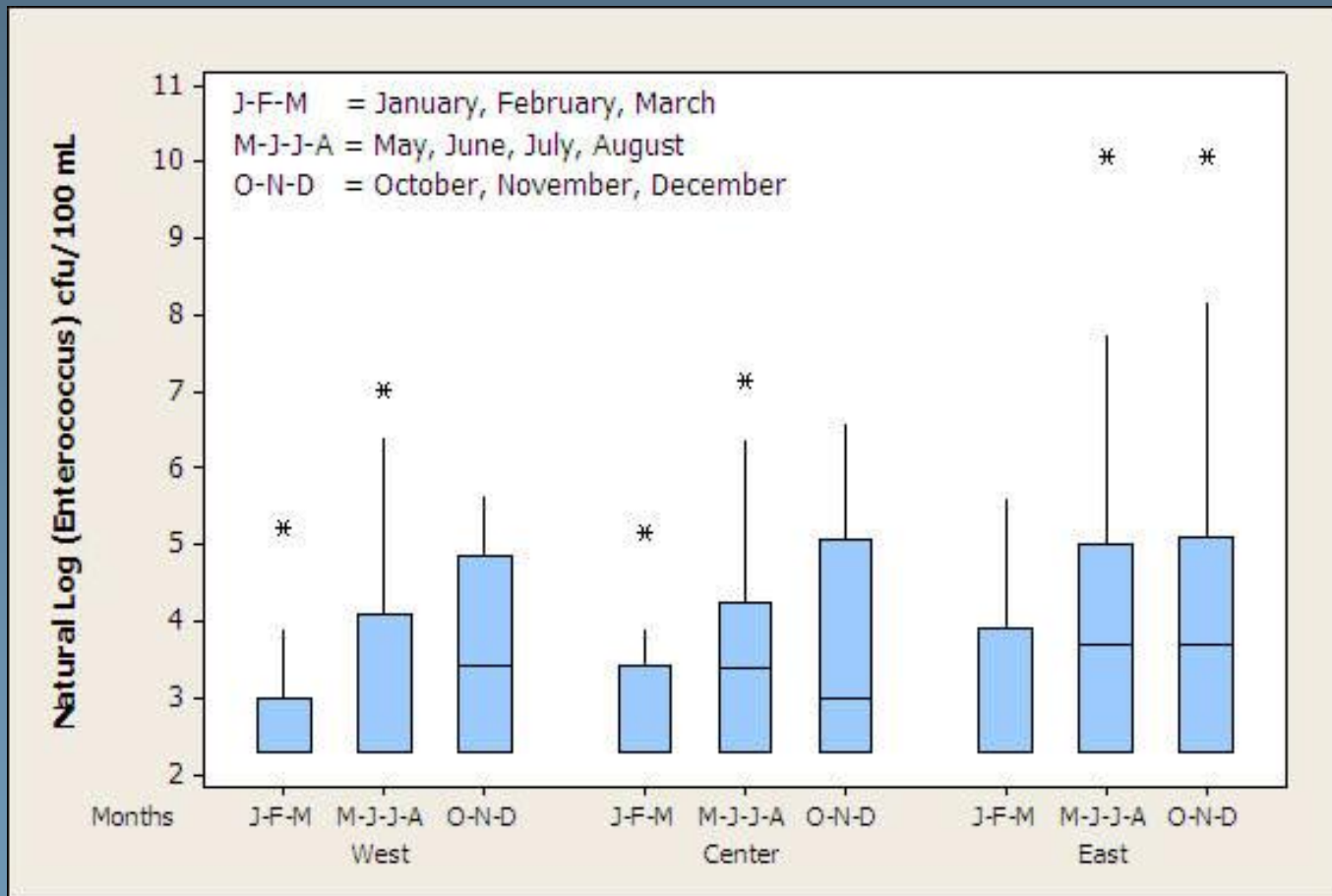
# Anatomy of a Boxplot



# Boxplots: Easton's Beach, Newport

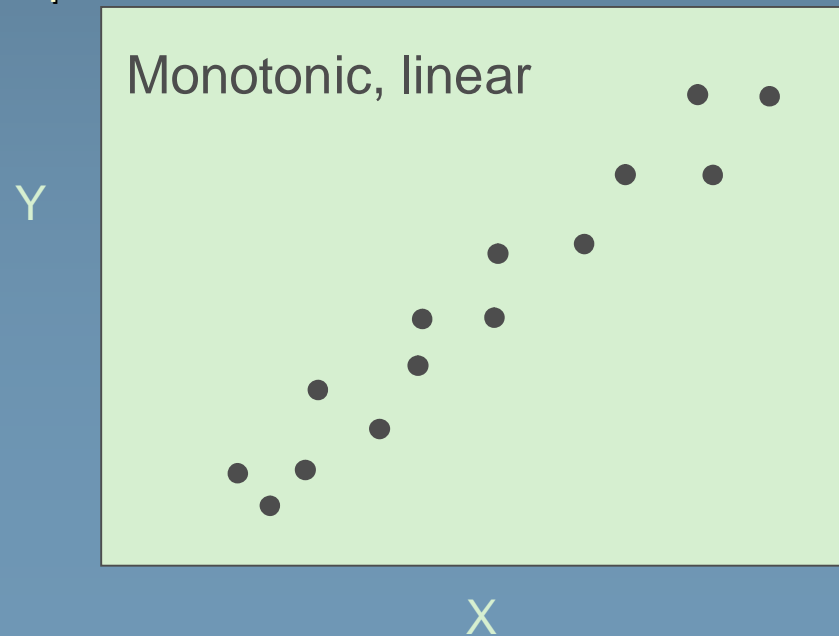


# Boxplots: Easton's Beach, Newport



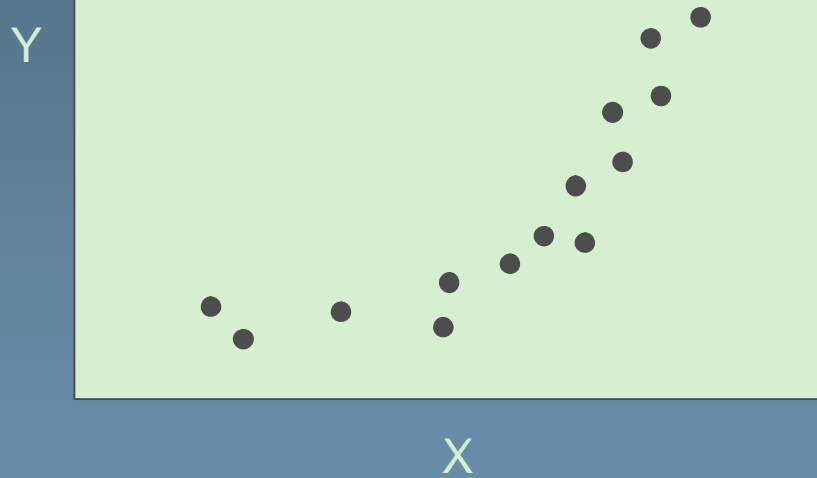
# Non-Linear Correlation

- Correlation measures the strength of association between two variables
- Linear correlation (Pearson's  $r$ )
- $Y$  generally decreases or increases as  $X$  increases
- One type of monotonic correlation
- Visual inspection of data critical – outliers can decrease value of  $r$

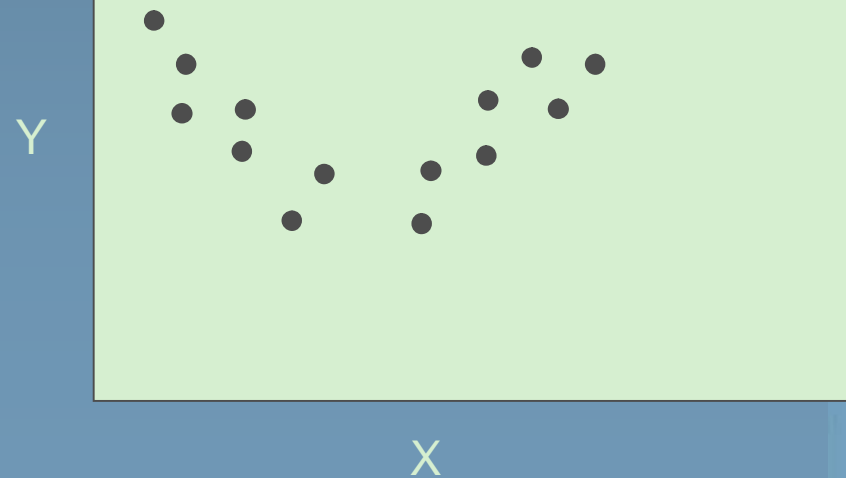


# Non-Linear Correlation

Monotonic, non-linear



Non-monotonic



# Kendall's Tau ( $\tau$ )

- Measure of monotonic correlation based on ranks
- Resistant to outliers
- Can be used when some data is censored
- Measures linear and non-linear monotonic correlations
- Invariant to power transformations
- Can be computed by hand
- “Strong” linear correlations (0.9) indicated by tau  $\sim$  0.7
- Sources:
  - *Helsel and Hirsch (1992)*
  - <http://pubs.usgs.gov/twri/twri4a3/>
  - [http://www.wessa.net/rwasp\\_kendall.wasp](http://www.wessa.net/rwasp_kendall.wasp)



# Spearman's Rho (Rank Correlation)

- Measure of monotonic correlation based on ranks
- Resistant to outliers
- Can be used when some data is censored
- Measures linear and non-linear monotonic correlations
- Different from Kendall's tau because in rho the differences between data ranked values are given more weight – different scales to measure the same correlation
- $\text{Rho} > \text{Tau}$ , but p-values for significance should be similar
- Can be computed by hand
- Sources:
  - *Helsel and Hirsch (1992)*
  - <http://pubs.usgs.gov/twri/twri4a3/>



# Correlation: Baltimore-Washington Airport

- Looking for evidence of a source for toxicity in deicing season runoff
- Landside (roads) or Airside (runway) source?
- Focused on three sampling locations:
  - 003 (*landside and airside*)
  - 306 (*airside*)
  - 007 (*airside*)
- BOD<sub>5</sub>, COD, TOC, TKN, Propylene Glycol, Total Glycol, Specific Conductivity

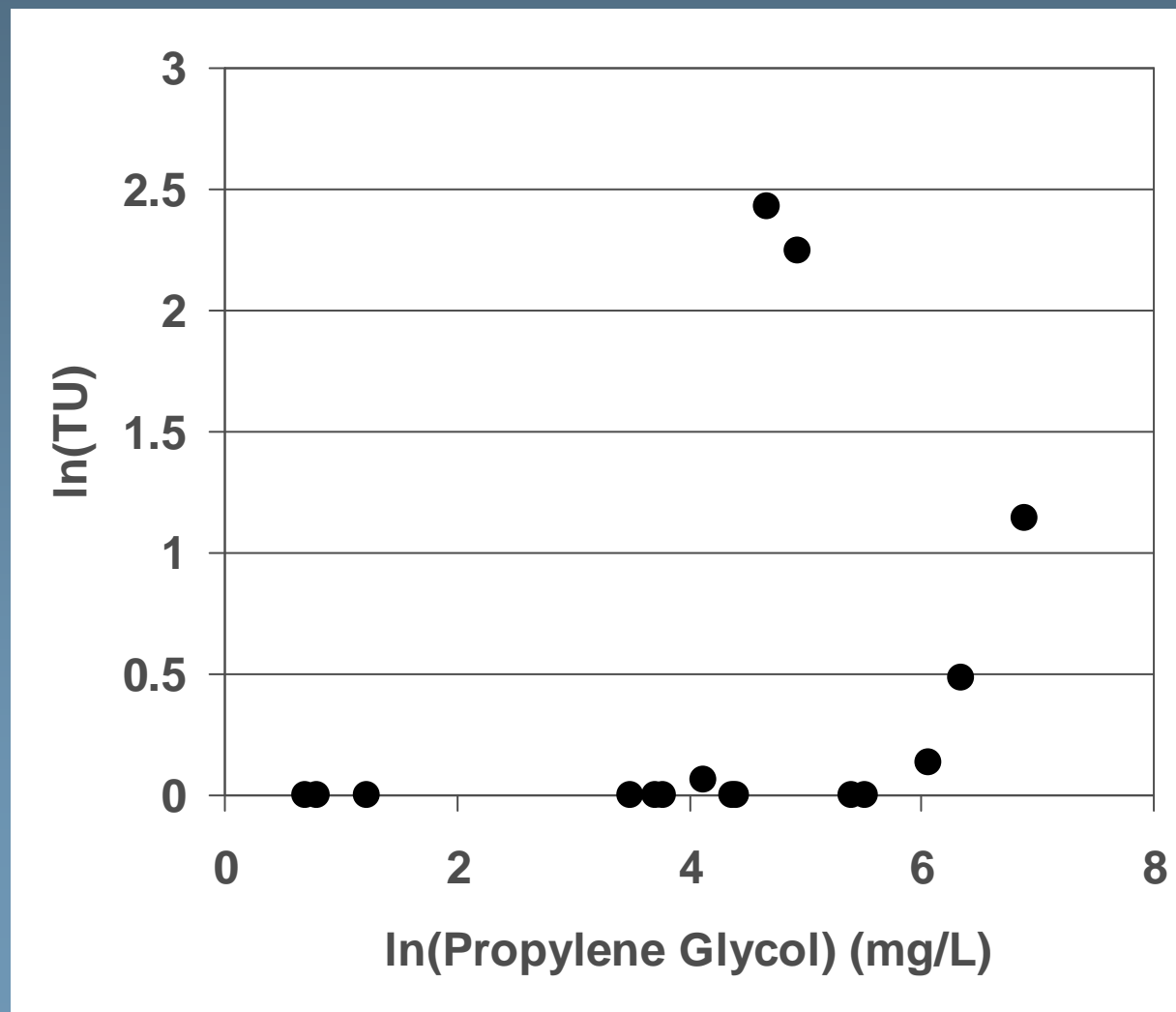


# Correlation: Baltimore-Washington Airport

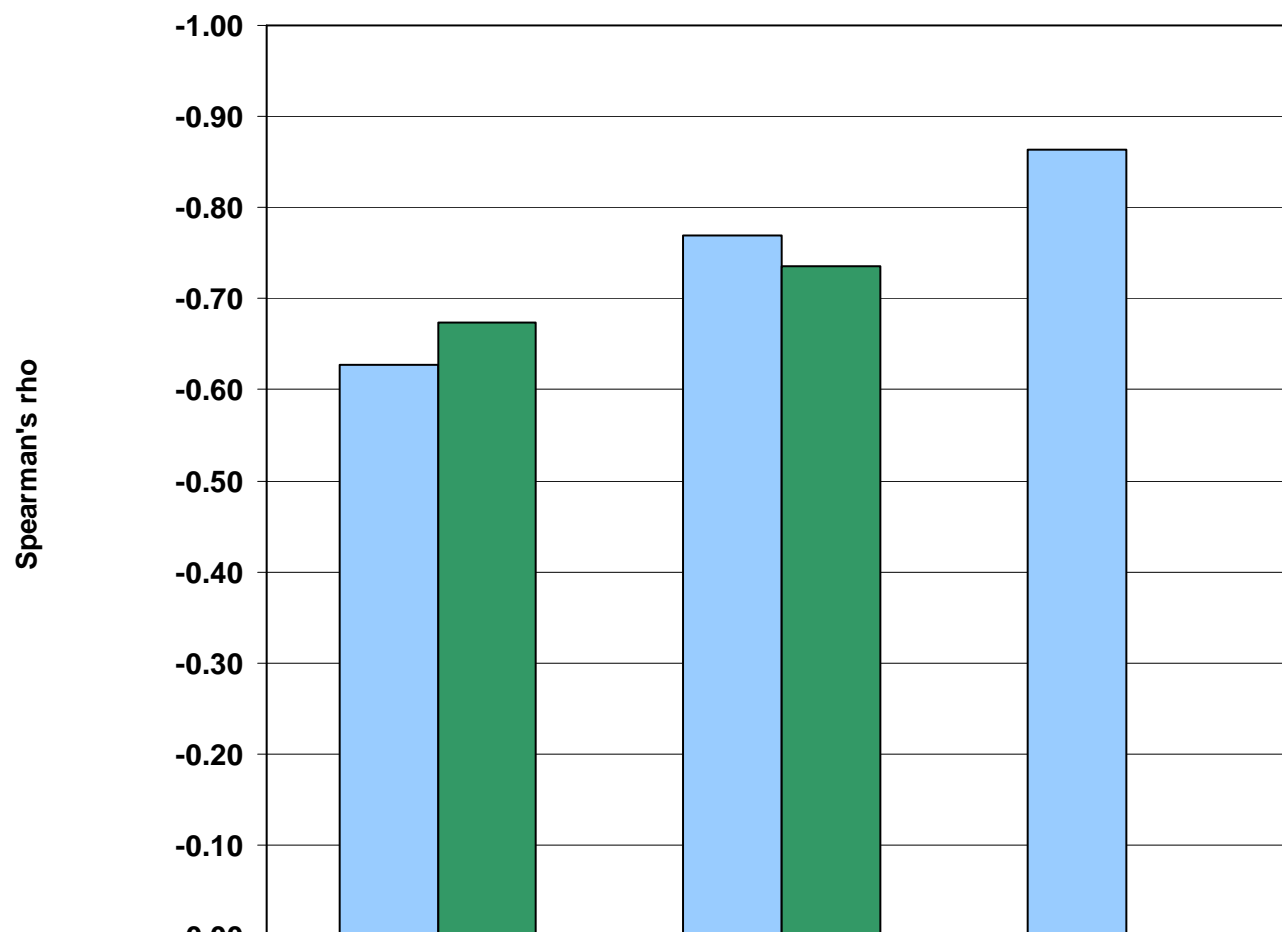
Parameter Station 007	Pearson's r correlation ( <i>p-value</i> )	Spearman's rho correlation ( <i>p-value</i> )	Kendall's tau correlation ( <i>p-value</i> )
BOD <sub>5</sub> (mg/l)	NS	NS	NS
COD (mg/l)	NS	-0.49887 (0.0492)	-0.47434 (0.0190)
TOC (mg/l)	NS	-0.57631 (0.0310)	-0.43819 (0.0461)
TKN (mg/l)	NS	NS	NS
Propylene Glycol (mg/l)	NS	-0.60665 (0.0127)	-0.47434 (0.0190)
Total Glycol (mg/l)	NS	-0.60665 (0.0127)	-0.47434 (0.0190)
Average Conductivity (μmhos)	NS	NS	NS



# Correlation: Baltimore-Washington Airport



# Correlation: Baltimore-Washington Airport



	Total Glycol (mg/l)	TKN (mg/l)	Avg. Conductivity (umhos)
Outfall 003	-0.63	-0.77	-0.86
Monitoring Point 306	-0.67	-0.74	0



# LOWESS

- LOcally WEighted Scatterplot Smoothing
- Smoothing technique
- Not constrained by a prior assumption about the mathematical function of a data relationship (e.g., linear, exponential, etc.)
- Aids data analysis for a large number of samples
- Helps to visualize the shape of the relationship between variables - difficult for the eye to follow the central tendency of a scatterplot



## LOWESS: DEP Stormwater Monitoring Data

- Database of stormwater monitoring reports compiled under the Connecticut General Permit for Discharge of Stormwater Associated with Industrial Activity
- 6402 monitoring reports 1995-2000, >300 industry types
- General Permit aquatic toxicity performance criterion of an  $LC_{50} > 50\%$  effluent for the daphnid, Daphnia pulex
- $LC_{50}$  concentration that results in 50% mortality of test species over the testing time period
- Higher  $LC_{50}$  value means lower acute aquatic toxicity

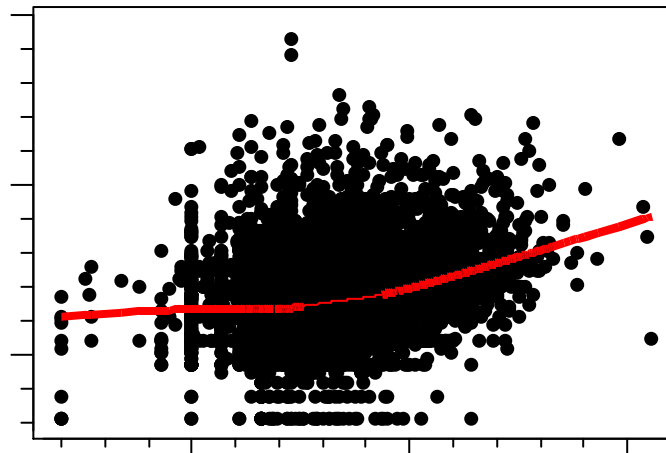


# LOWESS: DEP Stormwater Monitoring Data

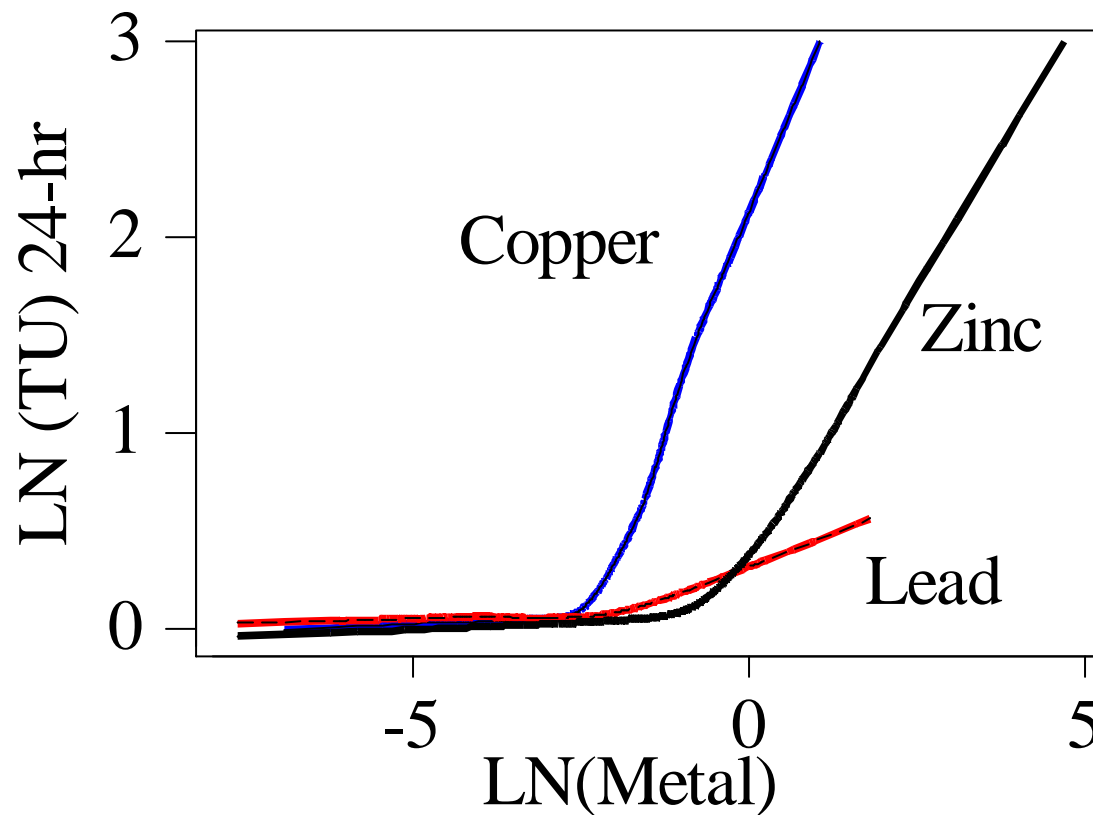
- Expression of toxicity in “toxicity units”
- $TU = (100\%/LC_{50})$  - Higher values means greater toxicity
- Examined relationship between toxicity and
  - *Total Copper*
  - *Total Lead*
  - *Total Zinc*
  - *Oil and Grease*
  - *Chemical Oxygen Demand*
  - *Total Suspended Solids*
  - *Hardness*
- Analyzed for various industry types



# LOWESS: DEP Stormwater Monitoring Data



# LOWESS: DEP Stormwater Monitoring Data



## LOWESS: DEP Stormwater Monitoring Data

- Suggests both a persistent relationship between metals concentration and aquatic toxicity and that metals toxicity is not necessarily a function of industrial process
- Linear regression may not be suitable mathematical model to describe data relationship
- No apparent relationship between toxicity and other water quality parameters - O&G, COD, TSS
- The data also shows a relationship between increasing metals concentration and total suspended solids



# LOWESS: DEP Stormwater Monitoring Data

Parameter	EPA Acute Water Quality Criteria	This Study
Copper	9.2 $\mu\text{g/l}$	~80 $\mu\text{g/l}$
Lead	34 $\mu\text{g/l}$	~80 $\mu\text{g/l}$
Zinc	65 $\mu\text{g/l}$	~370 $\mu\text{g/l}$

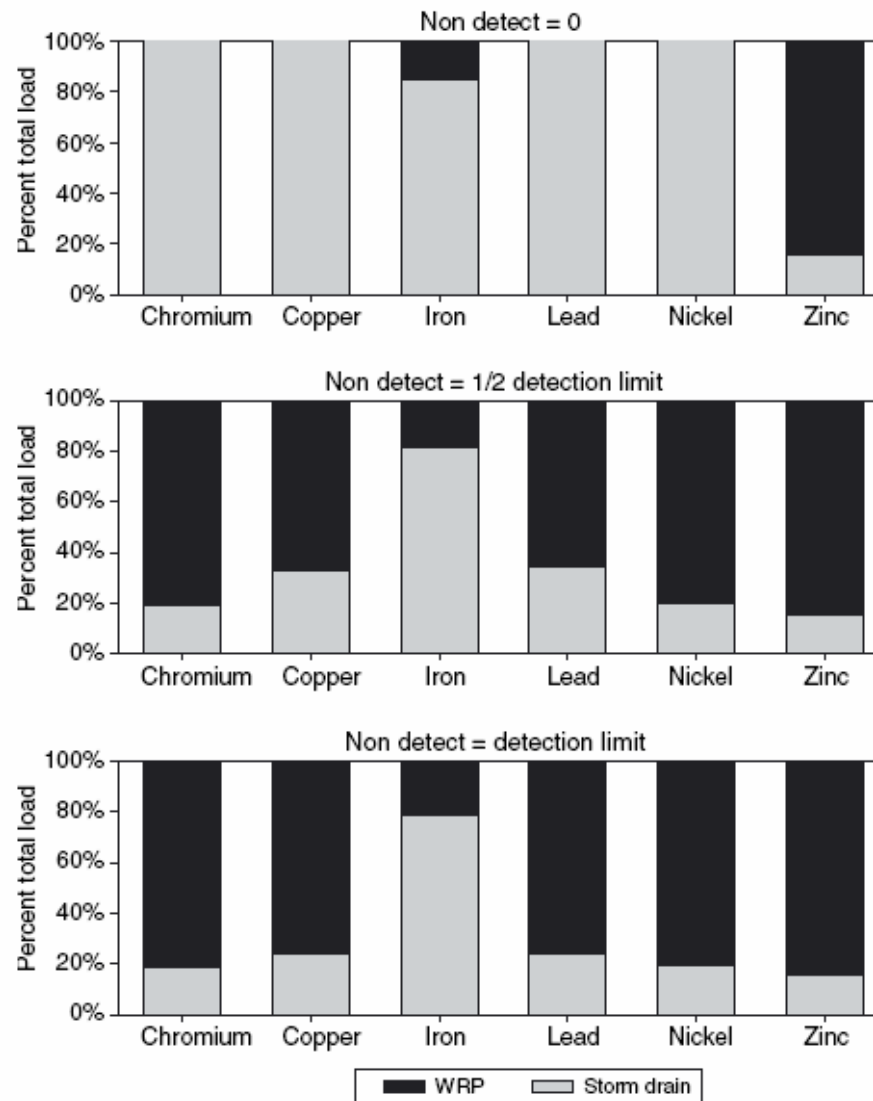


# Detection Limits

- Data that is “censored,” less than or greater than values
- Simple substitution (detection limit or zero), using half of the detection limit
- How can you take advantage of all the data?
- Correlation
  - *Use rank correlation (Kendall’s tau or Spearman’s rho)*
- Use cumulative distribution curves
  - *Allow for at least some comparison of the data*
- More information
  - *Helsel (2005), Nondetects and Data Analysis*



# Detection Limits



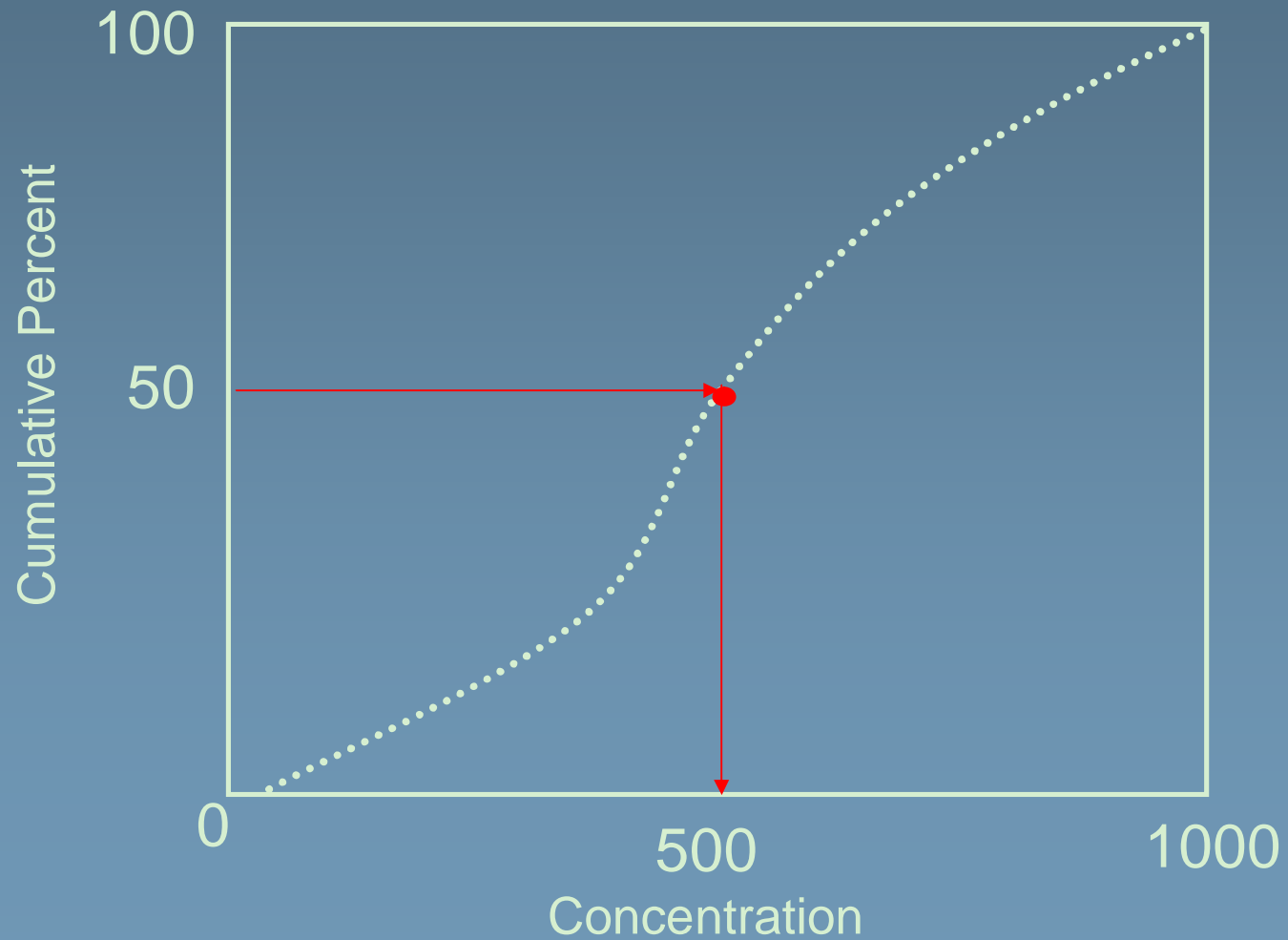
Stein and Ackerman, 2007

# Detection Limits

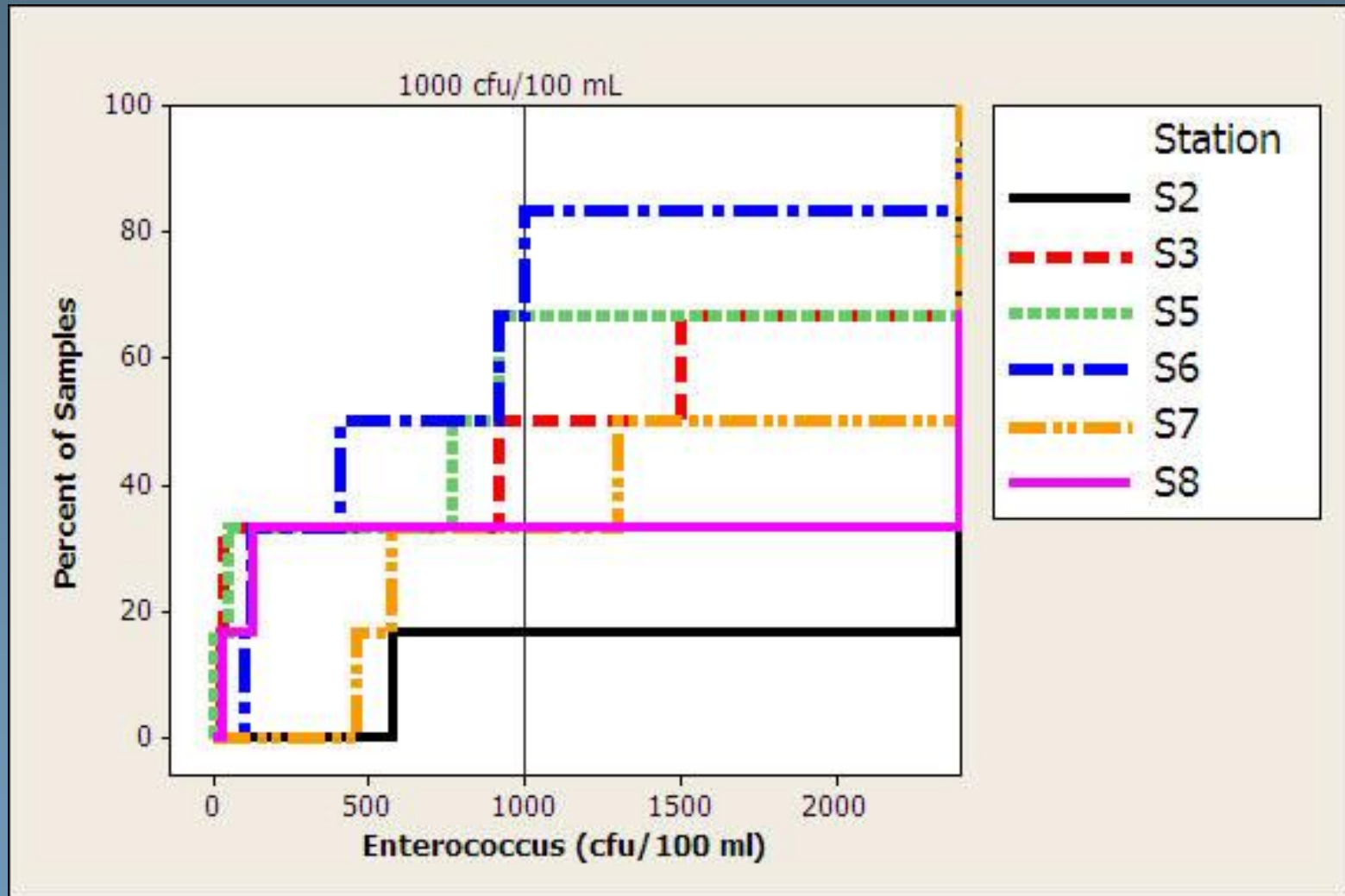
- Data that is “censored,” less than or greater than values
- Simple substitution (detection limit or zero), using half of the detection limit
- How can you take advantage of all the data?
- Correlation
  - *Use rank correlation (Kendall’s tau or Spearman’s rho)*
- Use cumulative distribution curves
  - *Allow for at least some comparison of the data*
- More information
  - *Helsel (2005), Nondetects and Data Analysis*



# Cumulative Distribution Curve



# Detection Limits



# Conclusions

- Visual and statistical analysis helps to extract the most information possible from data
- All data, even censored values, can and should be used in most cases
- Boxplots and cumulative distribution curves can provide concise, but rich, visual summaries
- Non-linear correlation can detect associations that might be missed with linear methods
- LOWESS is useful for large data sets; to detect relationships between variables



# Modeling

- Beaches, rivers and lakes are the top vacation destinations for Americans.
- In 2005, beach closings and advisories exceeded 20,000.
- Monitoring indicator organisms is the method used to determine water quality.
- Predicting bacterially-induced closings - a new paradigm?
- New methods: Regression models, Artificial Neural Networks



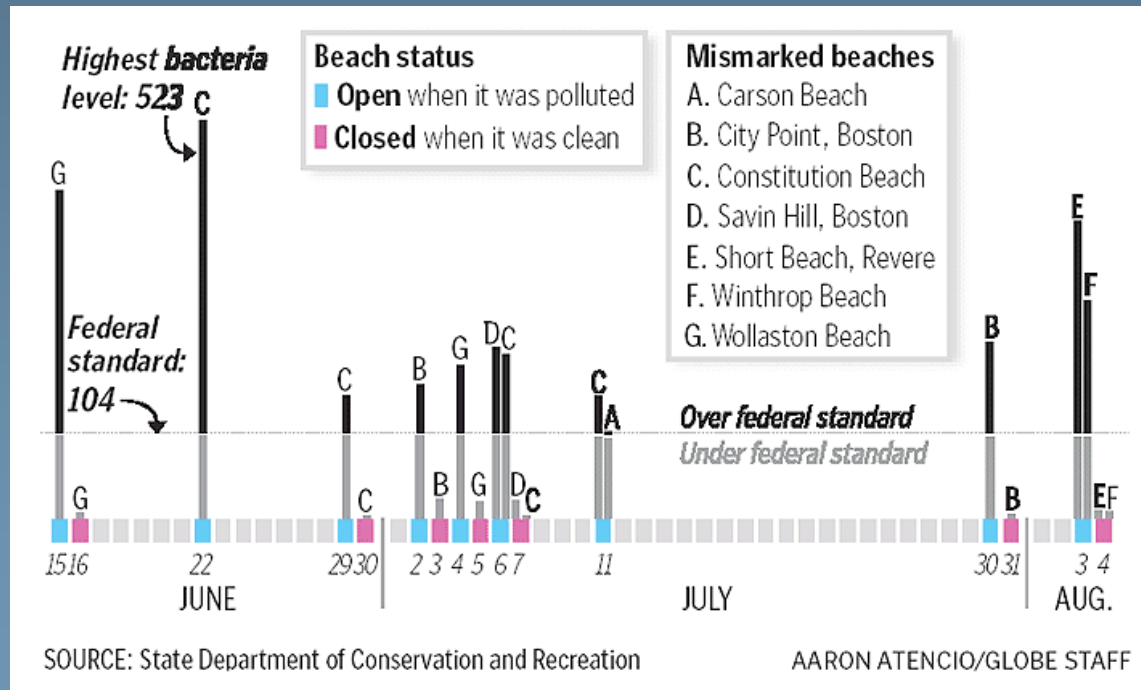
# Modeling Pathogens/Indicator Organisms

- Pathogens (bacteria) are a leading cause of water quality impairment in rivers and coastal waters in the United States (EPA, 2002)
- Implications for public health - drinking water and recreational waters - and aquatic health
- Ability to predict water quality conditions
  - *Understand processes*
  - *Manage resources, make decisions*



# Modeling Pathogens/Indicator Organisms

- How to address the problem of the time lag associated with indicator organism sampling and testing?



# Modeling Pathogens/Indicator Organisms

- Regression models used at Great Lakes Beaches for “nowcasting”
- Use readily measured variables to predict same day exceedance of water quality standards



In Cooperation With the Cuyahoga County Board of Health, Northeast Ohio Regional Sewer District, Ohio Water Development Authority, and Ohio Lake Erie Office

## Procedures for Developing Models To Predict Exceedances of Recreational Water-Quality Standards at Coastal Beaches



Techniques and Methods 6–B5

U.S. Department of the Interior  
U.S. Geological Survey



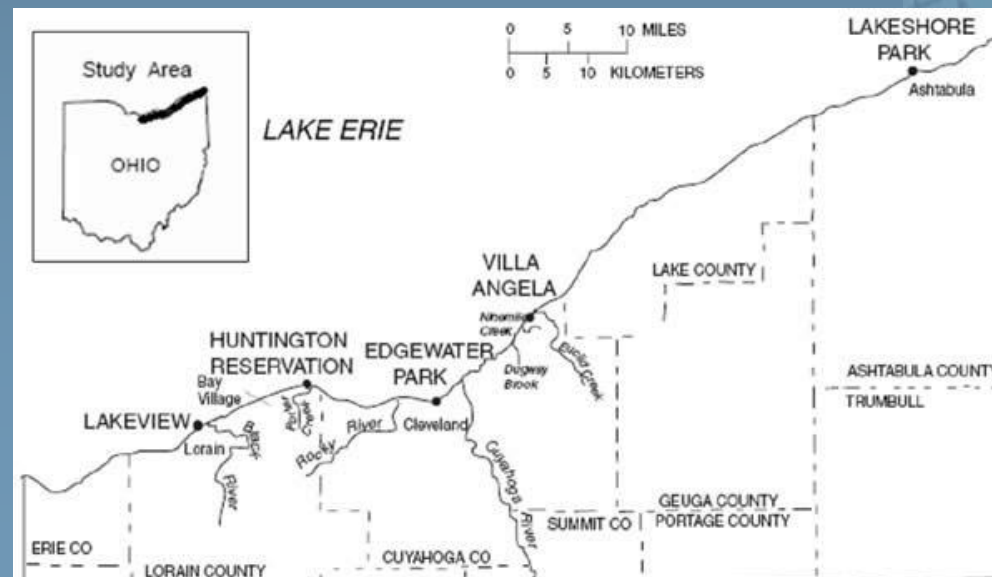
# Modeling Pathogens/Indicator Organisms

Study	Independent Variable/Hydrologic Setting	R <sup>2</sup>	Correct Classification of Violations
Christensen et al., (2000)	fecal coliform/ freshwater river	55.6%, 62%	n/a
Eleria (2002)	fecal coliform/ freshwater river	46.11- 60.40%	44-66%
Francy et al. (2003)	<i>E. coli</i> / freshwater beach	17- 58%	73.2-90.9%
Olyphant and Whitman (2003)	<i>E. coli</i> / freshwater beach	71%	88%
Olyphant (2005)	<i>E. coli</i> / freshwater beach	42 - 58%	~ 90%
Mas and Ahlfeld (in press)	fecal coliform/ freshwater stream	n/a	46-81%



# Ohio Nowcasting Beach Advisories

- Huntington Beach, Ohio – Lake Erie
- Multiple Linear Regression (MLR) model using wave height, turbidity, day of year, 48-hr rainfall (more weight to past 24-hrs),  $R^2=42\%$
- Developed using 2000-2005 swimming season data
- Implement by predicting a probability of exceeding the threshold water quality standard of 235 cfu/100 mL *E. coli*
- In 2005 – Model predicted 50% of closures correctly; prior day *E. coli* concentrations predicted none correctly
- Used operationally in 2006
- <http://www.ohionowcast.info>

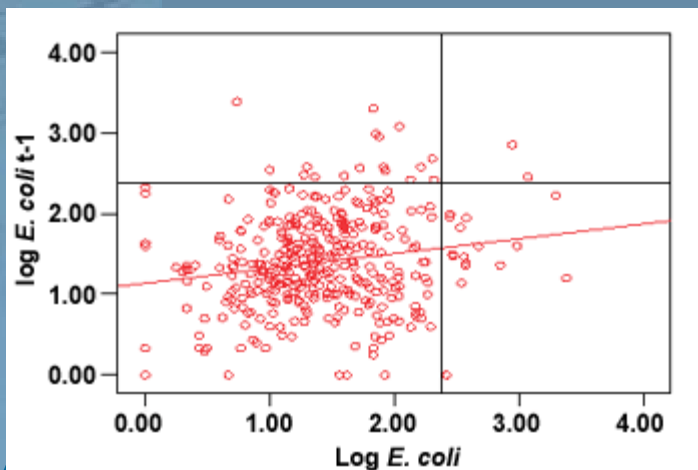


# Project S.A.F.E

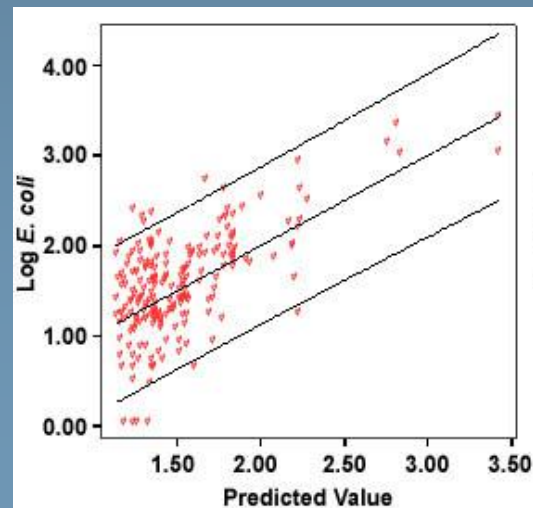
- Swimming Advisory Forecast Estimate - Lake Michigan
- USGS Great Lakes Science Center



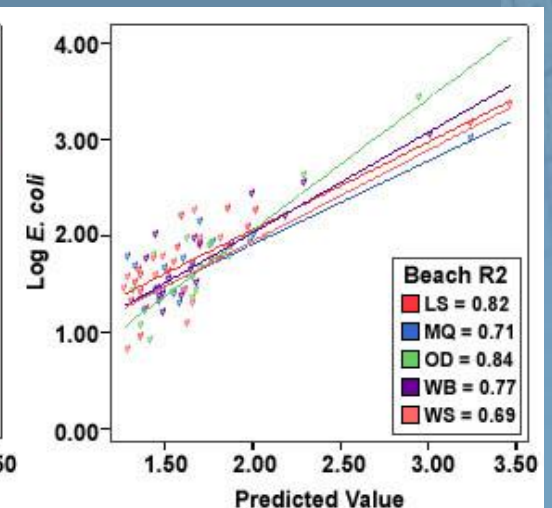
- Multiple Linear Regression (MLR) model using 24-hr rain, wind speed and direction, air and water temperature, wave height, UV index, gage height at stream discharging into Lake
- <http://www.glsc.usgs.gov/ProjectSAFE.php>



Current Monitoring Model:  
Day 2 *E. coli* (Advisory Day) = Day 1 *E. coli* (Sampling Day)



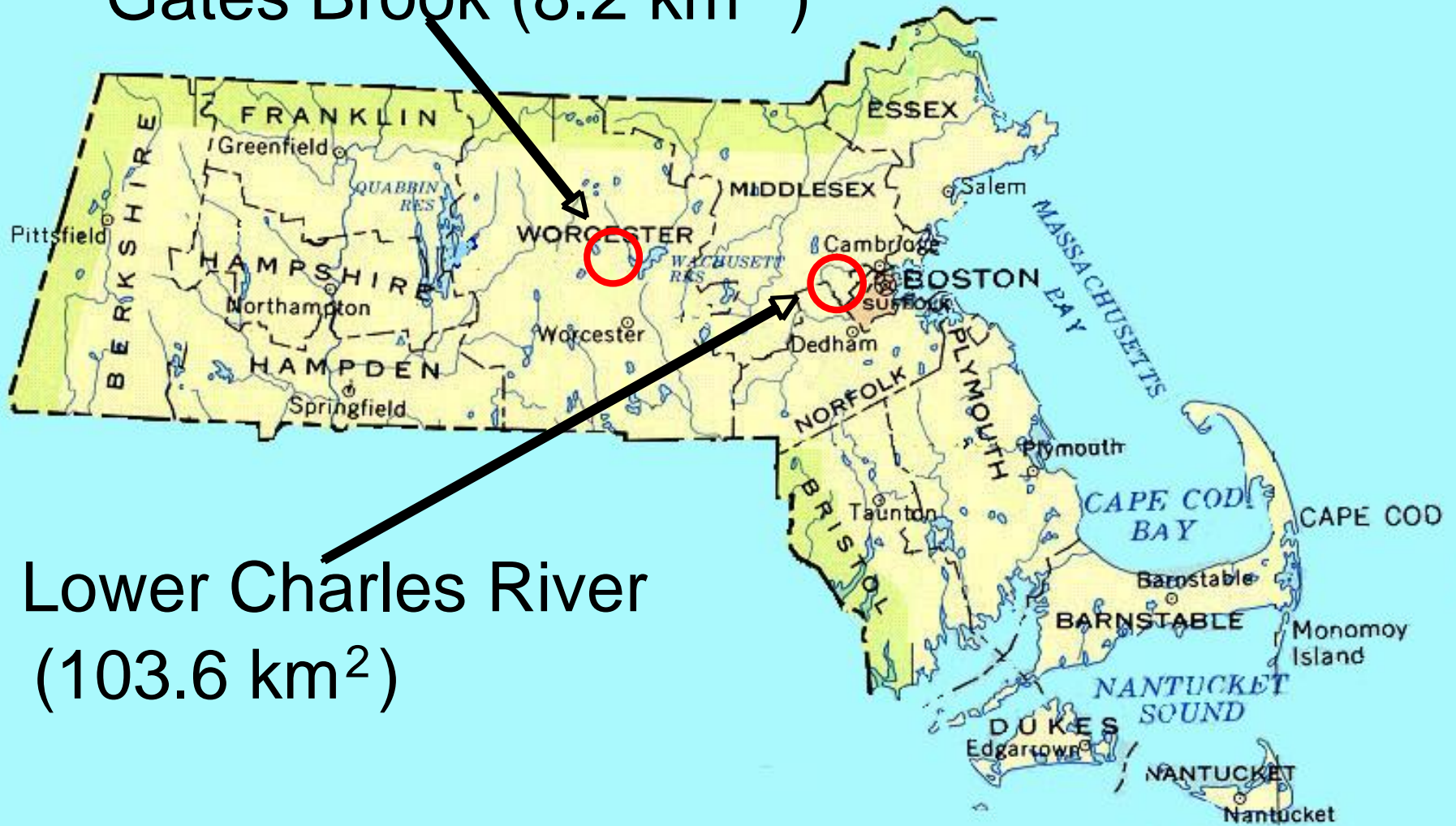
Predicted vs. actual *E. coli*, with 95% confidence interval for all beaches and conditions.



Modeling Results for Individual Beaches

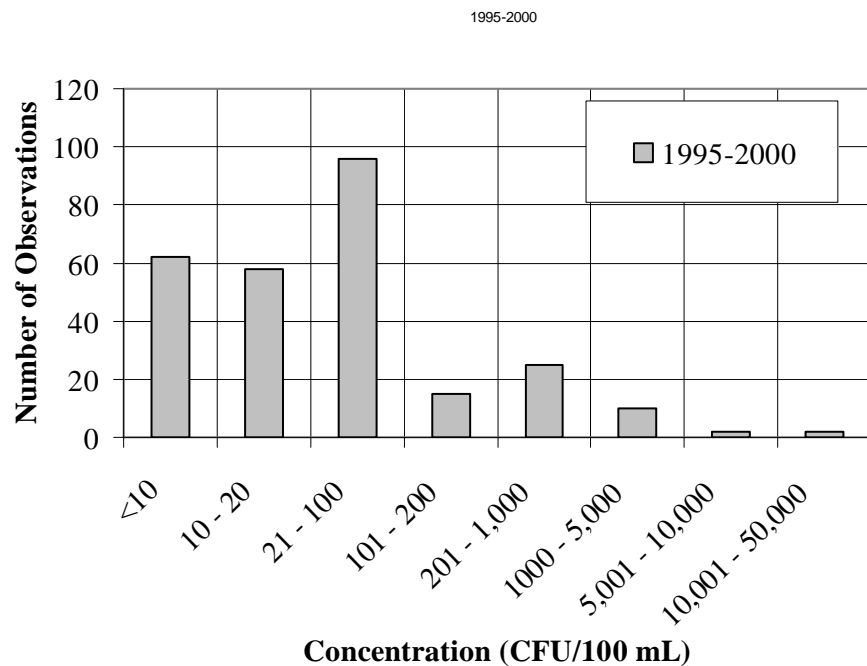
# Gates Brook, Central Massachusetts

Gates Brook (8.2 km<sup>2</sup>)



Lower Charles River  
(103.6 km<sup>2</sup>)

# Gates Brook, Central Massachusetts



1 sampling location

Weekly or biweekly

1989-2003

602 input/output pairs

Conductivity, water temp.,  
fecal coliform  
air temperature,  
precipitation

Streamflow (1995-2003)

Potential sources - failing  
septic, runoff



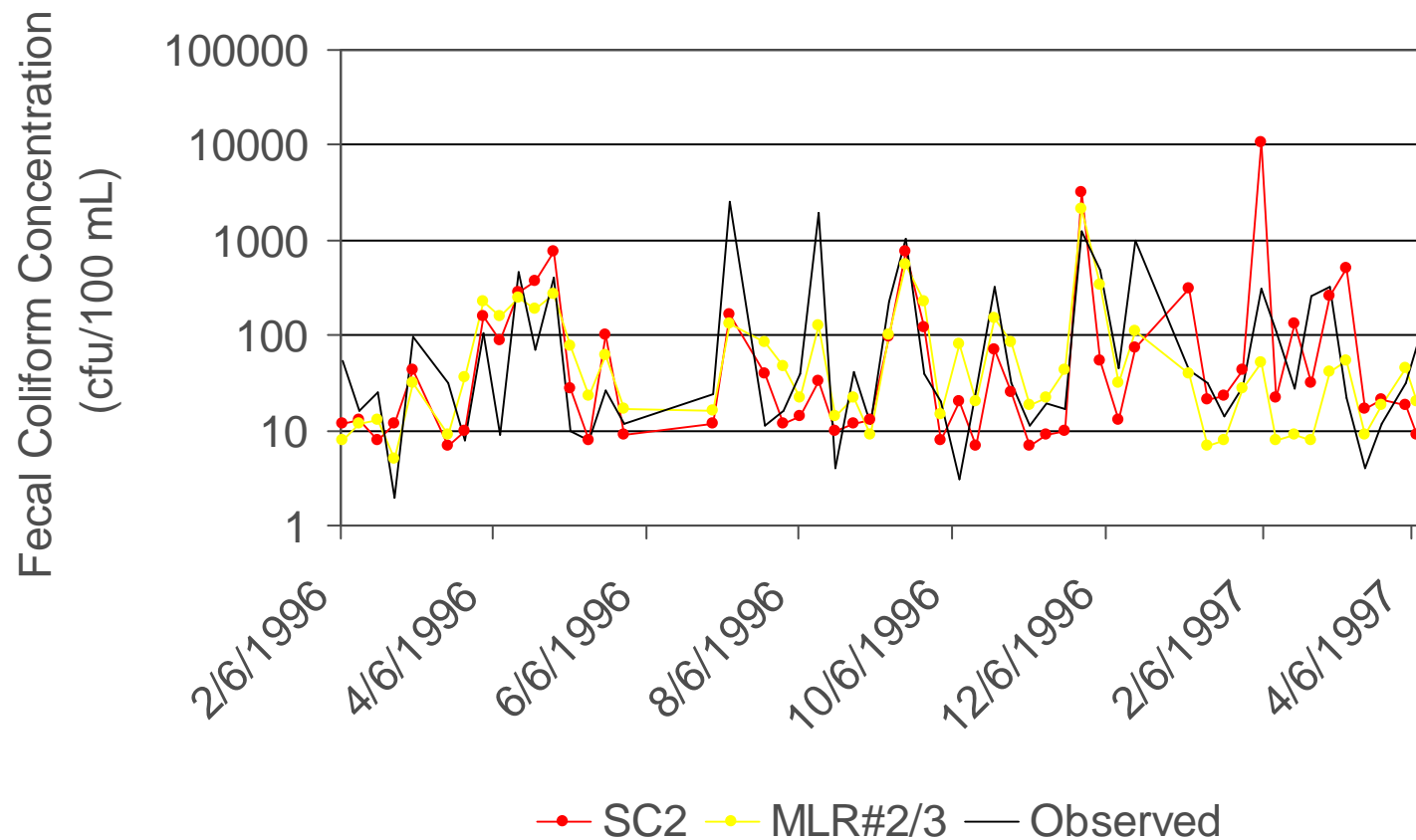
# Gates Brook, Central Massachusetts

- Do Artificial Neural Network (ANN) models provide an alternative to regression in complex settings?
- ANN and regression models developed using 270 data values from 1995-2000
- Evaluated performance based on visual inspection, correct identification of violations of water quality standard, and percentage of false positives and negatives
- Mas and Ahlfeld, *Hydrological Sciences Journal*, in press (2007)



# Gates Brook, Central Massachusetts

- SC2:  $\ln(\text{cond})$ ,  $\ln(\text{strm})$ ,  $\ln(\text{Tmean})$ ,  $\ln(\text{pcpn})$ ,  $\ln(\text{pcpn}-1)$
- MLR #2/#3:  $\ln(\text{strm})$ ,  $\ln(\text{cond})$ ,  $\ln(\text{pcpn})$ ,  $\ln(\text{pcpn}-1)$ ,  $\text{WTemp}$ ,  $\text{Tmax}$



# Gates Brook, Central Massachusetts

- ANN models better at classification, especially relative to Class B water quality standards
- ANNs generally have lower percentage of false negatives

Model	Class A (20 cfu/100 mL)		Class B (200 cfu/100 mL)	
	Violations	FP/FN (%)	Violations	FP/FN (%)
ANN SC1	61%	13/26	62%	6/9
ANN SC2	69%	11/20	46%	6/13
ANN SC3	81%	15/17	62%	6/9
MLR#1	58%	15/28	38%	4/15
MLR#2/#3	75%	17/20	38%	4/15



# Conclusions

- Culture-based methods lack the timelines desired for beach management
- Regression methods have demonstrated effectiveness at Great Lakes beaches
- Other non-linear classification methods, including ANNs, may provide useful predictive models for some systems.
- Combination of monitoring and modeling likely to emerge as the new paradigm for recreational water quality management

