Improving Uncertainty Characterization in USEPA's Guidelines for Deriving Aquatic Life Criteria Using Decision Contexts

Doug McLaughlin

Invited Expert Meeting on Revising USEPA's Guidelines for Deriving Aquatic Life Criteria,

September 14-16, 2015

Arlington, VA

Main Message

- The quality and transparency of the science behind EPA's aquatic life criteria can be improved by revising guidelines to include methods for…
  1. …developing quantitative estimates of important statistical uncertainties…
  2. …in ways that can be readily understood by a wide range of stakeholders/decision-makers.

This presentation offers some approaches to consider.

Part I. Introduction

Defining “Decision Context” & the Role of Uncertainty Analysis in WQC Derivation

- For the purpose of this presentation, think of a “decision context” as part of the “so what” of scientific data and information.
  - What decision is the scientific information supporting?
- A numeric criterion makes several types of “Yes/No” decisions quite obvious and necessary. Some examples:
  - is a water quality criterion protective of designated uses?
  - is a water quality criterion being attained?
  - are trends in water quality moving toward a WQC?
- In criteria science, there is (almost) always a “Maybe” because there is (almost) always some degree of scientific uncertainty

Defining "Decision Context" & the Role of Uncertainty Analysis in WQC Derivation, cont'd

- The over-arching WQ management goal is to correctly answer "Yes" or "No"
  - Try not to say "Yes" when the correct (true) answer is "No", and vice versa
  - In practice, this means making WQC-based decisions with confidence, and limiting "false negative" and "false positive" decision errors to acceptable levels

Other Voices on The Importance of Uncertainty Characterization in WQC Science

- SETAC Workshop Publication on WQC Science (Reiley et al. 2003):
  - Numerous benefits to increased use of explicit, quantitative characterization of uncertainty in WQC
  - *''The overall result will be more realistic risk assessments, the inclusion of uncertainty into decision-making, and the appreciation of the potential for over- and under-protection. During implementation, these uncertainty limits could be incorporated into risk assessments for site-specific criteria and recognized in the interpretation of monitoring data.''* (p. 83)
  - *"The statistical uncertainty associated with WQC and species sensitivity curves should be expressed as part of each criterion."* (p. 84)

Other Voices on The Importance of Uncertainty Characterization in WQC Science, cont'd

- Summary Minutes, September, 2005 EPA Science Advisory Board Aquatic Life Criteria Guidelines Meeting
  - *"… important to continue "thinking outside of the box" in order to review and revise water quality criteria using the existing '1985 Guidelines.'"* (p. 4);
  - *"… important for EPA to consider how the Agency would deal with uncertainties in setting thresholds and making decisions."* (p. 9);
  - *"… the Agency should consider how the revisions could decrease uncertainty."* (p. 20)

Other Voices on The Importance of Uncertainty Characterization in WQC Science, cont'd

- EPA 2010 Guidance "Using Stressor-response Relationships to Derive Numeric Nutrient Criteria"

"Before finalizing candidate criteria based on stressor-response relationships, one should systematically evaluate the scientific defensibility of the estimated relationships and the criteria derived from those relationships.

More specifically, one should consider whether estimated relationships accurately represent known relationships between stressors and responses and whether estimated relationships are precise enough to inform decisions." (p. 65)

Part II. Uncertainty Limits Can Be Obtained Using Existing 1985 Guidelines Equations and Monte Carlo Analysis

(see McLaughlin & Jain 2011, for example)

1985 Guidelines Approach for Acute Toxicity: Derive CMC from Toxicity Data

LC50s → SMAV → GMAV → FAV → CMC

- LC50 = Chemical concentration lethal to 50% of a test population, 8 or more families required;
- SMAV = Species Mean Acute Value;
- GMAV = Genus Mean Acute Value;
- FAV = Final Acute Value;
- CMC = Criterion Maximum Concentration=FAV/2

$S = 9.3346$

$L = [4.3331 - (9.3346)(2.04875)]/4 = -3.6978$

$A = (9.3346)(\sqrt{0.05}) - 3.6978 = -1.6105$

$FAV = e^{-1.6105} = 0.1998$

Deriving Uncertainty Limits From Replicate Tests of a Single Test Species

Copper criterion example

Replicate toxicity tests allow for an estimate of the true mean EC50 for this species, and the uncertainty of the estimate.

Table 2. Example SMAV calculations using replicate BLM-normalized EC50 results for an amphipod presented in Table 1 of USEPA (2007)

Amphipod, *Hyalella azteca*		
Test Number	EC50	Ln EC50
1	12.19	2.501
2	9.96	2.299
3	15.77	2.758
4	8.26	2.111
5	8.09	2.091
6	15.49	2.740
7	18.8	2.934
Arithmetic mean		2.490*
Geometric mean (SMAV)	12.07*	
Standard deviation		0.334
n	7	
Standard error ($s/\sqrt{n}$)		0.126

*Note: SMAV = exp(2.490) or 12.07 μg/L; EC50 units are μg/L.

from McLaughlin and Jain (2011)

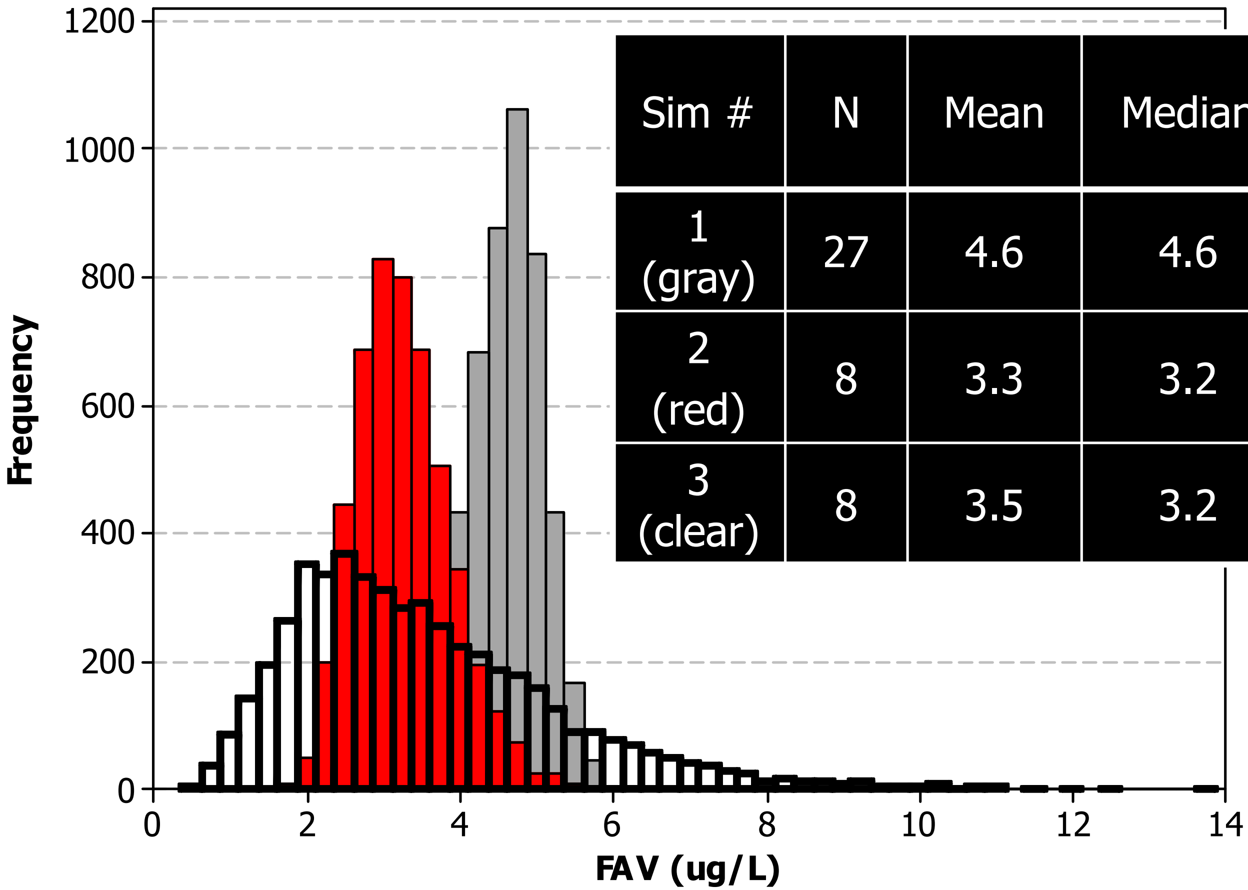
A Monte Carlo Approach:

- Use Monte Carlo computer simulation to generate a new set of SMAVs (1 per species) using the mean and standard error of the acceptable LC50 results for each species
- Derive GMAVs using SMAVs of any tested genus with more than one species;
- Determine the four most sensitive genera;
- Use these GMAVs, their sensitivity rank, and the total number of genera to calculate FAV using 1985 Guidelines equation;
- Repeat (5000 trials in McLaughlin & Jain 2011);
- Select desired FAV confidence limits from the resulting distribution of FAVs (divide each FAV by 2 to get CMC distribution)

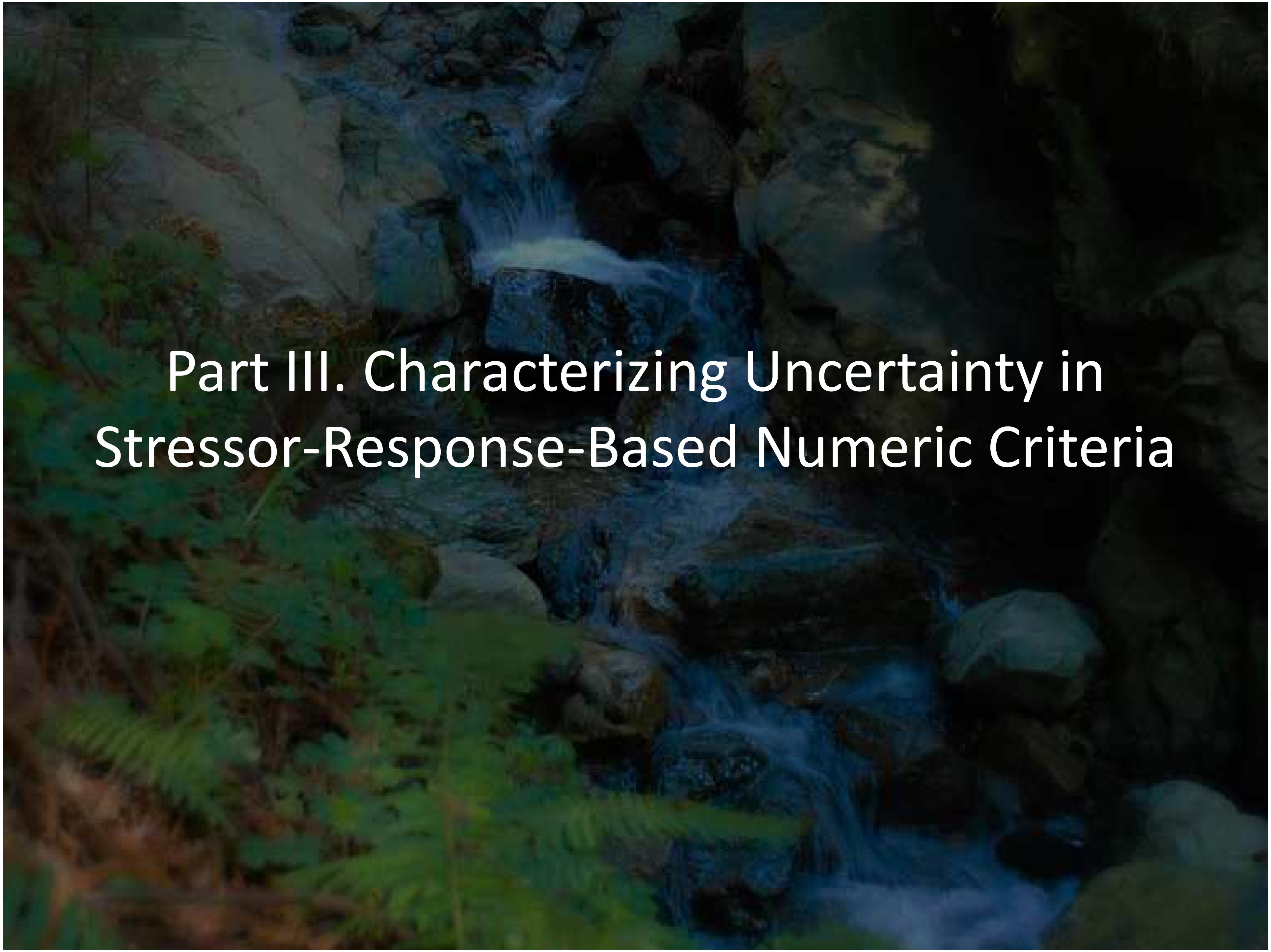
Example: BLM-Adjusted Copper Data, Three FAV Simulations

- 1 - Monte Carlo simulation using the full copper data set;
- 2 - Monte Carlo simulation using a "minimum data set" (8 taxa), with actual number of toxicity tests available for each taxa;
- 3 - Monte Carlo simulation using the same 8 taxa, with the numbers of tests set to 3 for all taxa;

Simulations 1, 2 & 3 Compared



Sim #	N	Mean	Median	5th %ile	95th %ile
1 (gray)	27	4.6	4.6	3.6	5.4
2 (red)	8	3.3	3.2	2.4	4.3
3 (clear)	8	3.5	3.2	1.4	6.8

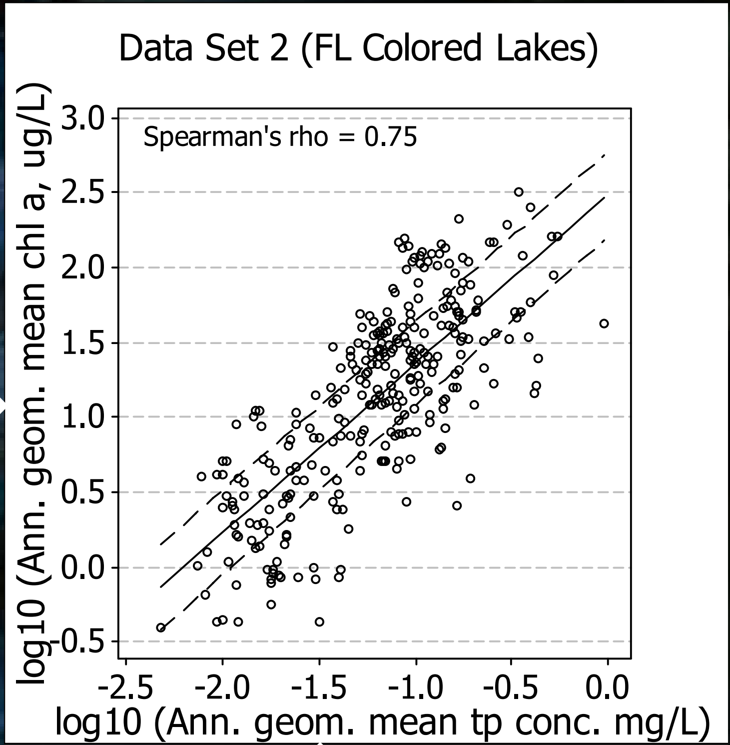


Part III. Characterizing Uncertainty in Stressor-Response-Based Numeric Criteria

Characterizing Uncertainty in Stressor-Response-Based Numeric Criteria

- Stressor-response relationships and numeric thresholds/criteria form a predictive model:
  - What type of predictions?
    1. response levels (a value)
    2. response condition (management implications)

Response variable



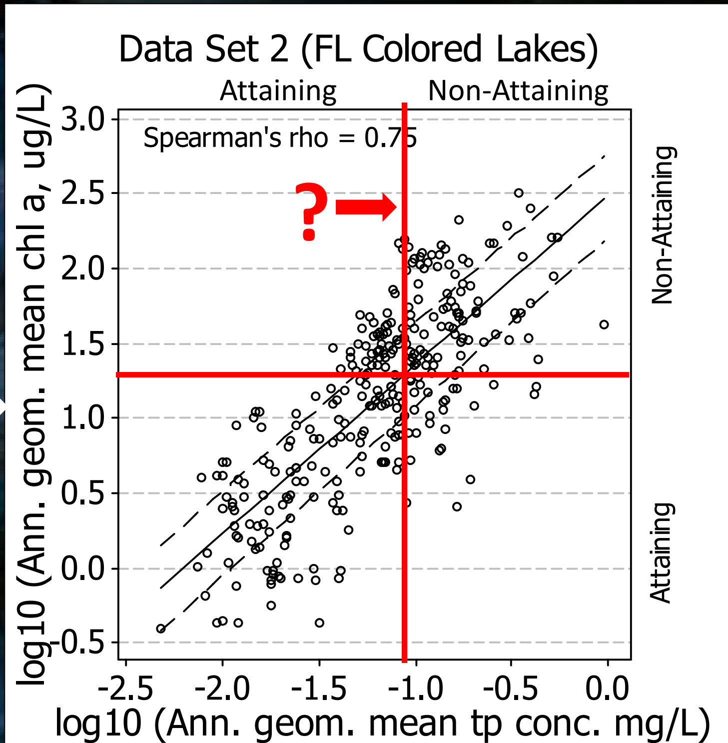
Stressor variable

From McLaughlin (2012b)

Characterizing Uncertainty in Stressor-Response-Based Numeric Criteria

- Stressor-response relationships and numeric thresholds/criteria form a predictive model:
  - What type of predictions?
    1. response levels (a value)
    2. response condition (management implications)

Response variable

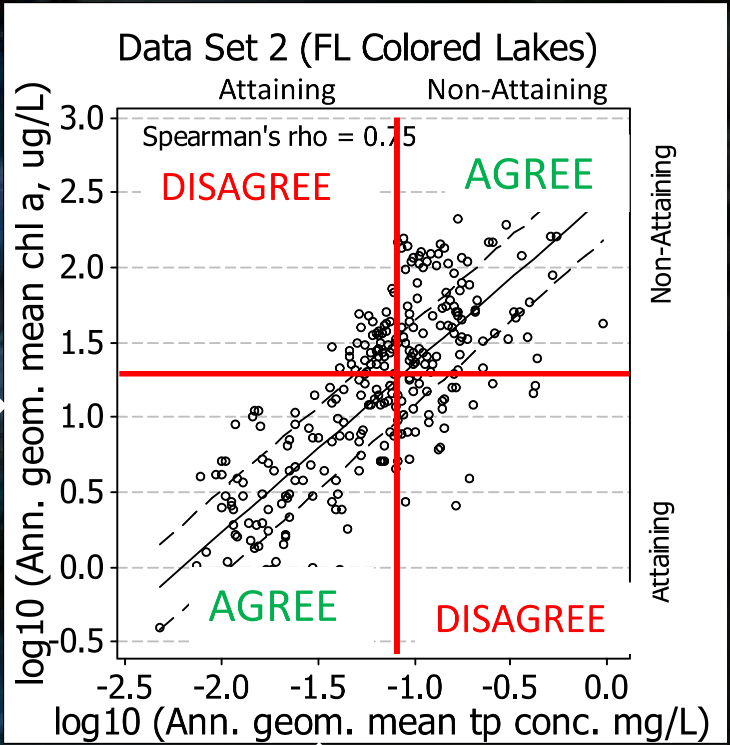


Stressor variable

A Receiver Operating Characteristics (ROC) Approach: The 2x2 Matrix "Overlay"

- Classification method commonly used in medicine and other fields, less in environmental science to date
- Useful for both categorical data and continuous data with numeric thresholds/criteria
- Calculate performance across a range of candidate stressor criteria.
- Can quantify uncertainty in terms of decision error probabilities, to supplement other statistical metrics

Response variable

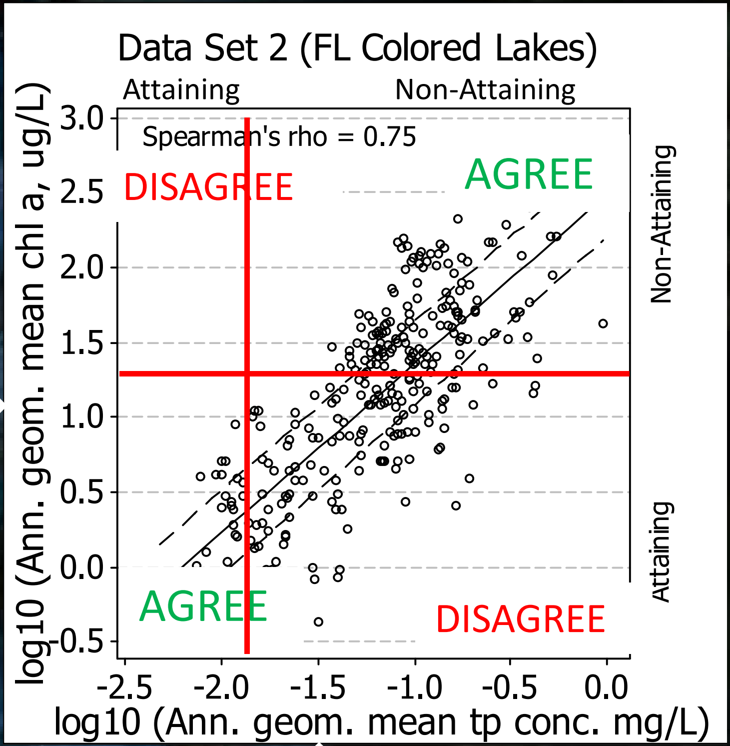


Stressor variable

A Receiver Operating Characteristics (ROC) Approach: The 2x2 Matrix "Overlay"

- Classification method commonly used in medicine and other fields, less in environmental science to date
- Useful for both categorical data and continuous data with numeric thresholds/criteria
- Calculate performance across a range of candidate stressor criteria.
- Can quantify uncertainty in terms of decision error probabilities, to supplement other statistical metrics

Response variable

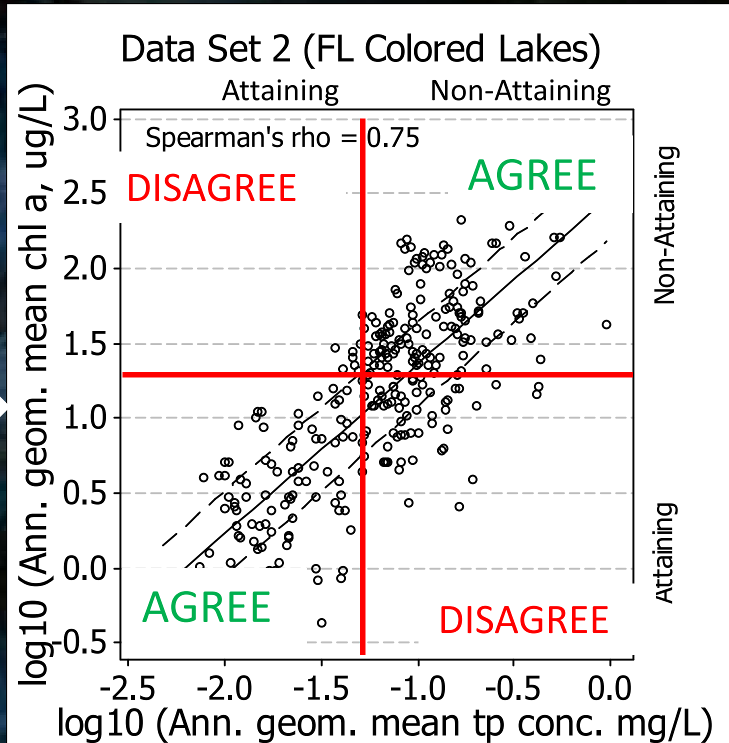


Stressor variable

A Receiver Operating Characteristics (ROC) Approach: The 2x2 Matrix "Overlay"

- Classification method commonly used in medicine and other fields, less in environmental science to date
- Useful for both categorical data and continuous data with numeric thresholds/criteria
- Calculate performance across a range of candidate stressor criteria.
- Can quantify uncertainty in terms of decision error probabilities, to supplement other statistical metrics

Response variable

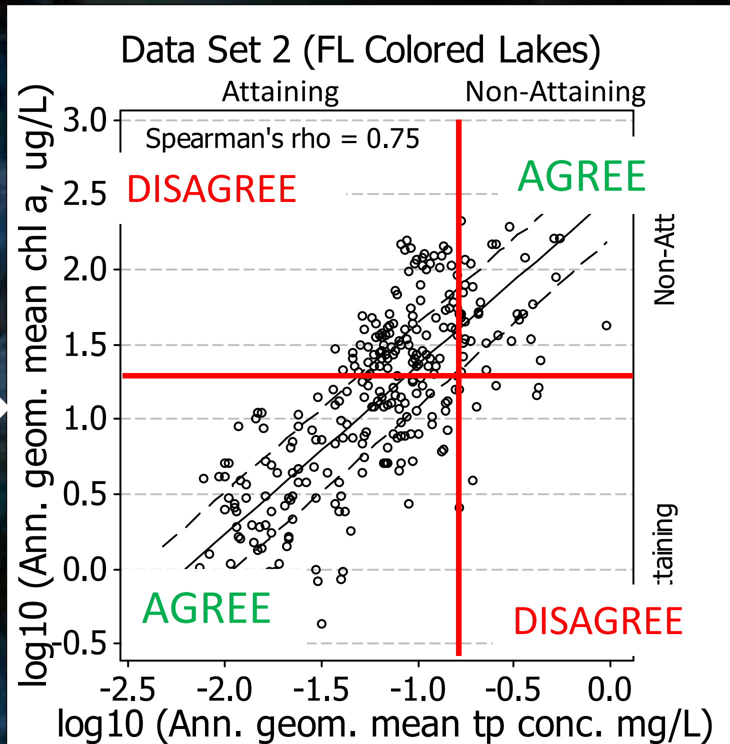


Stressor variable

A Receiver Operating Characteristics (ROC) Approach: The 2x2 Matrix "Overlay"

- Classification method commonly used in medicine and other fields, less in environmental science to date
- Useful for both categorical data and continuous data with numeric thresholds/criteria
- Calculate performance across a range of candidate stressor criteria.
- Can quantify uncertainty in terms of decision error probabilities, to supplement other statistical metrics

Response variable

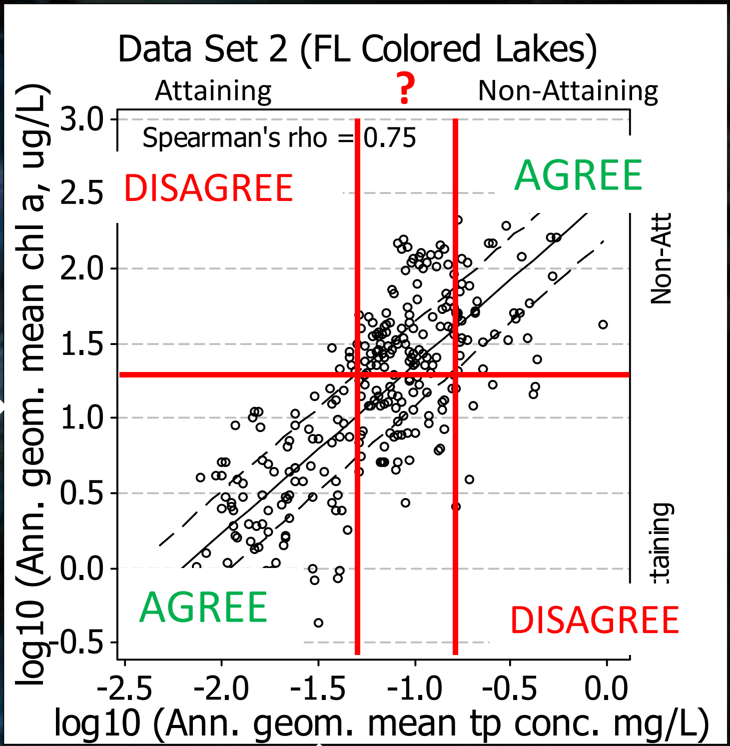


Stressor variable

A Receiver Operating Characteristics (ROC) Approach: The 2x2 Matrix "Overlay"

- Classification method commonly used in medicine and other fields, less in environmental science to date
- Useful for both categorical data and continuous data with numeric thresholds/criteria
- Calculate performance across a range of candidate stressor criteria.
- Can quantify uncertainty in terms of decision error probabilities, to supplement other statistical metrics

Response variable



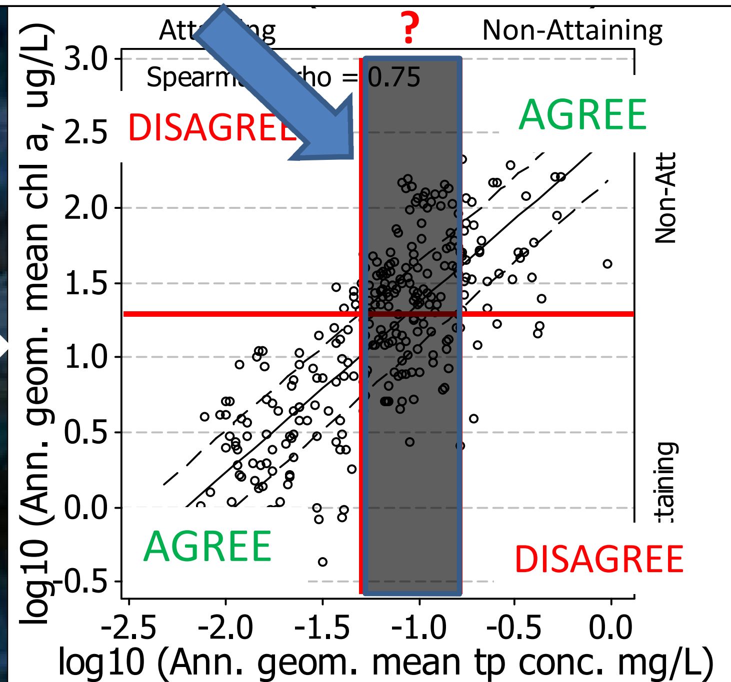
Stressor variable

A Receiver Operating Characteristics (ROC) Approach: The 2x2 Matrix "Overlay"

- Classification method commonly used in medicine and other fields, less in environmental science to date
- Useful for both categorical data and continuous data with numeric thresholds/criteria
- Calculate performance across a range of candidate stressor criteria.
- Can quantify uncertainty in terms of decision error probabilities, to supplement other statistical metrics

"Gray Region" – a role for a criteria range and "combined criteria", aka "bio-confirmation"

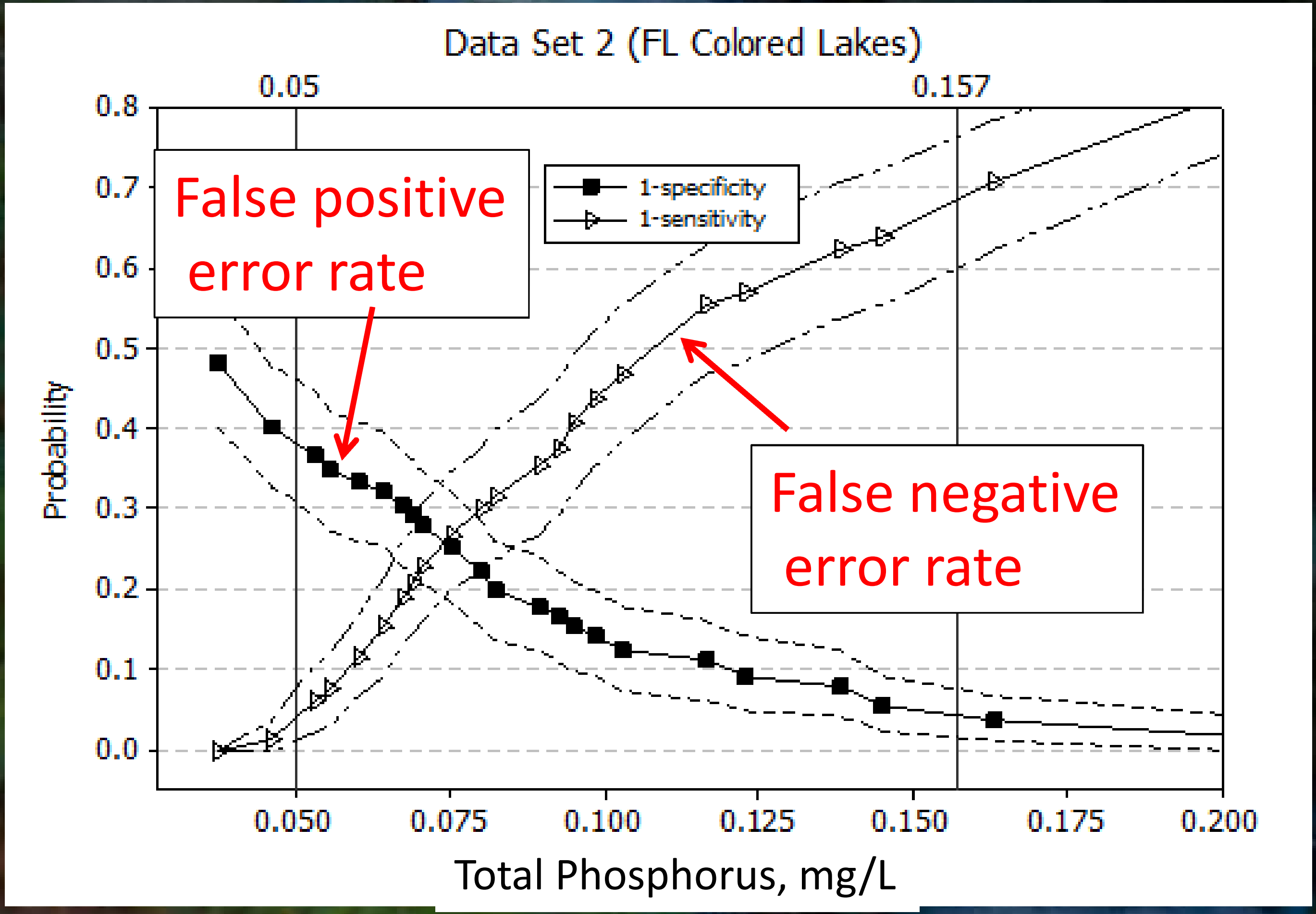
Response variable



Stressor variable

ROC Provides Information on Trade-Offs Among Decision Error Types

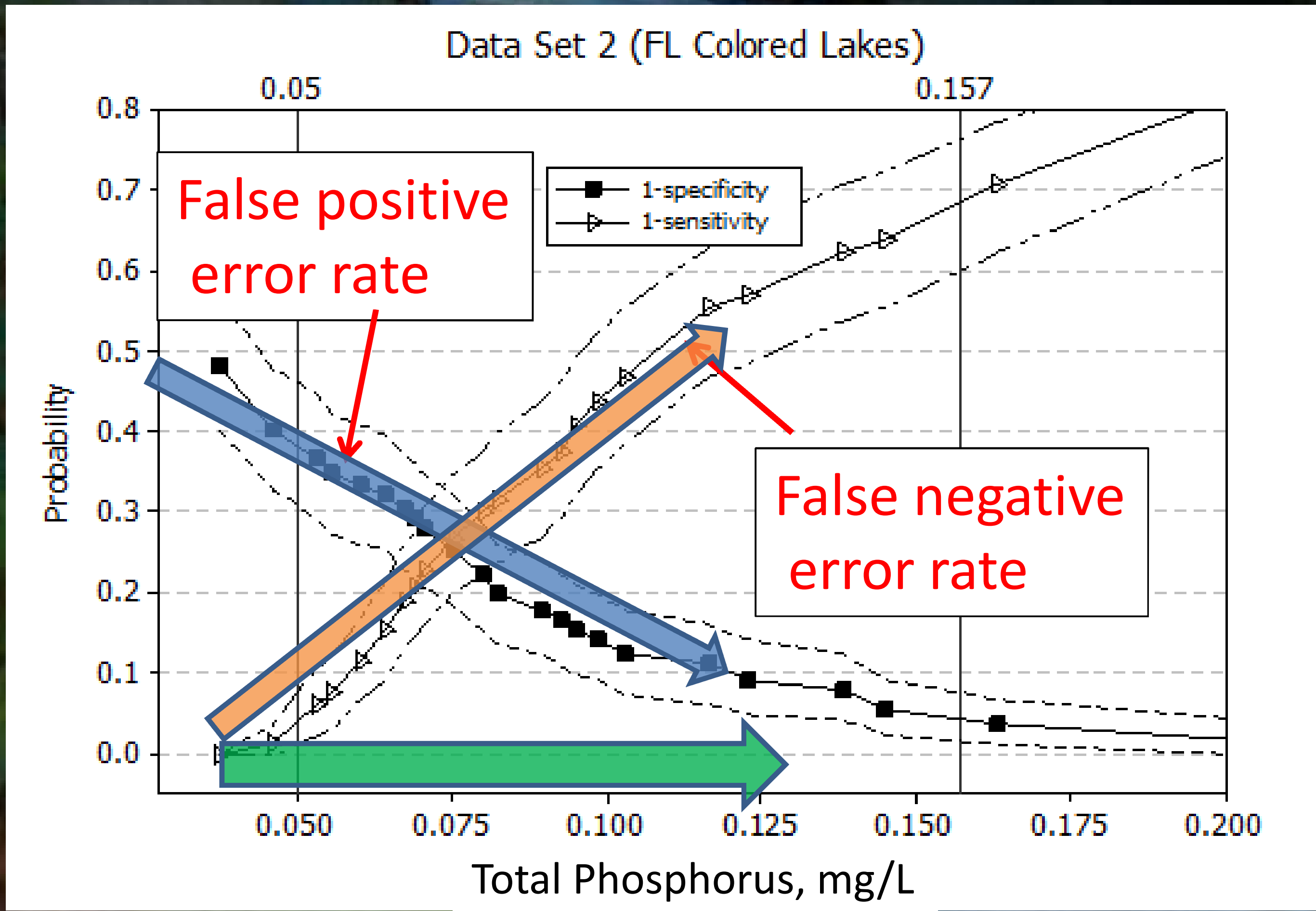
Error rates as a function of candidate TP criterion



From McLaughlin (2012b)

ROC Provides Information on Trade-Offs Among Decision Error Types

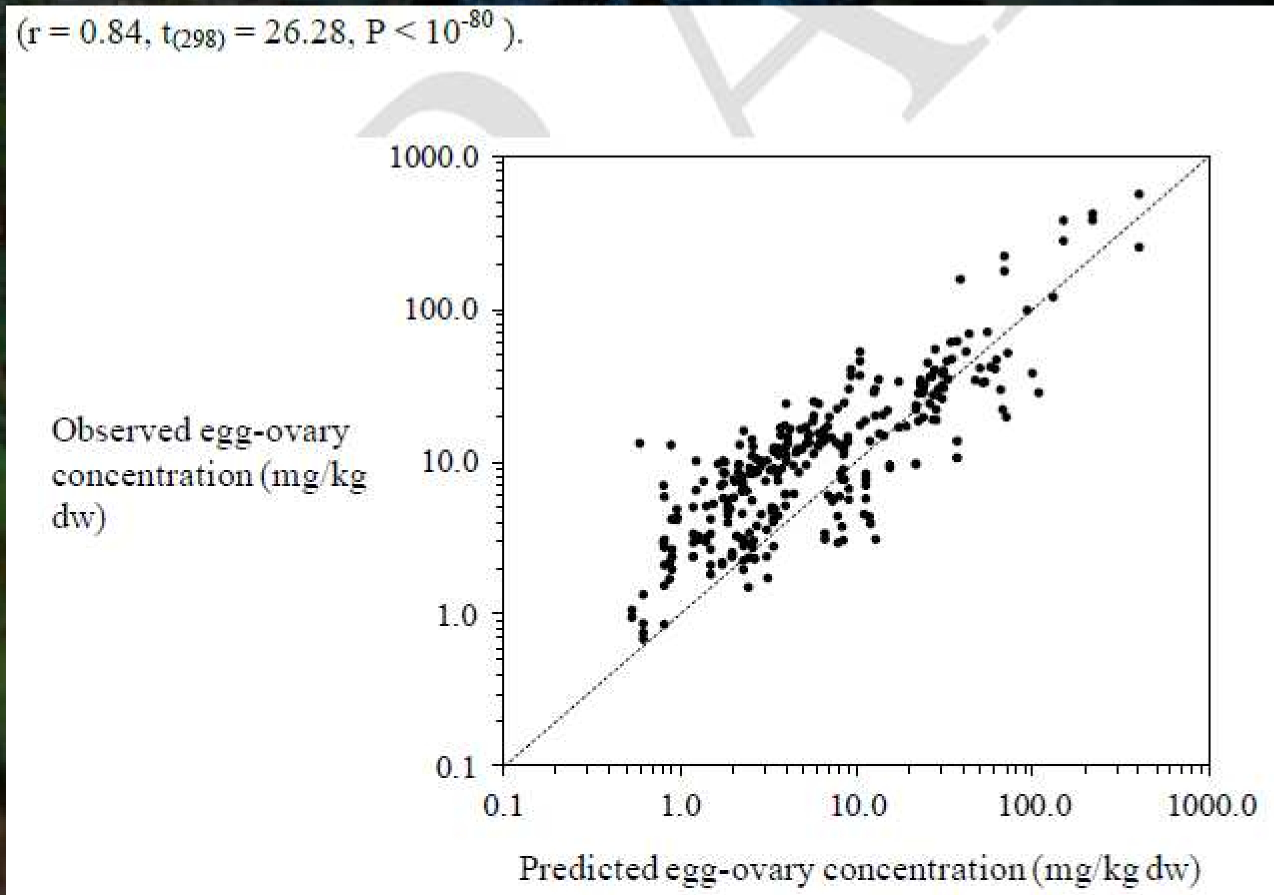
Error rates as a function of candidate TP criterion



From McLaughlin (2012b)

Binary Classification Method in Currently in USEPA Draft WQC Guidance

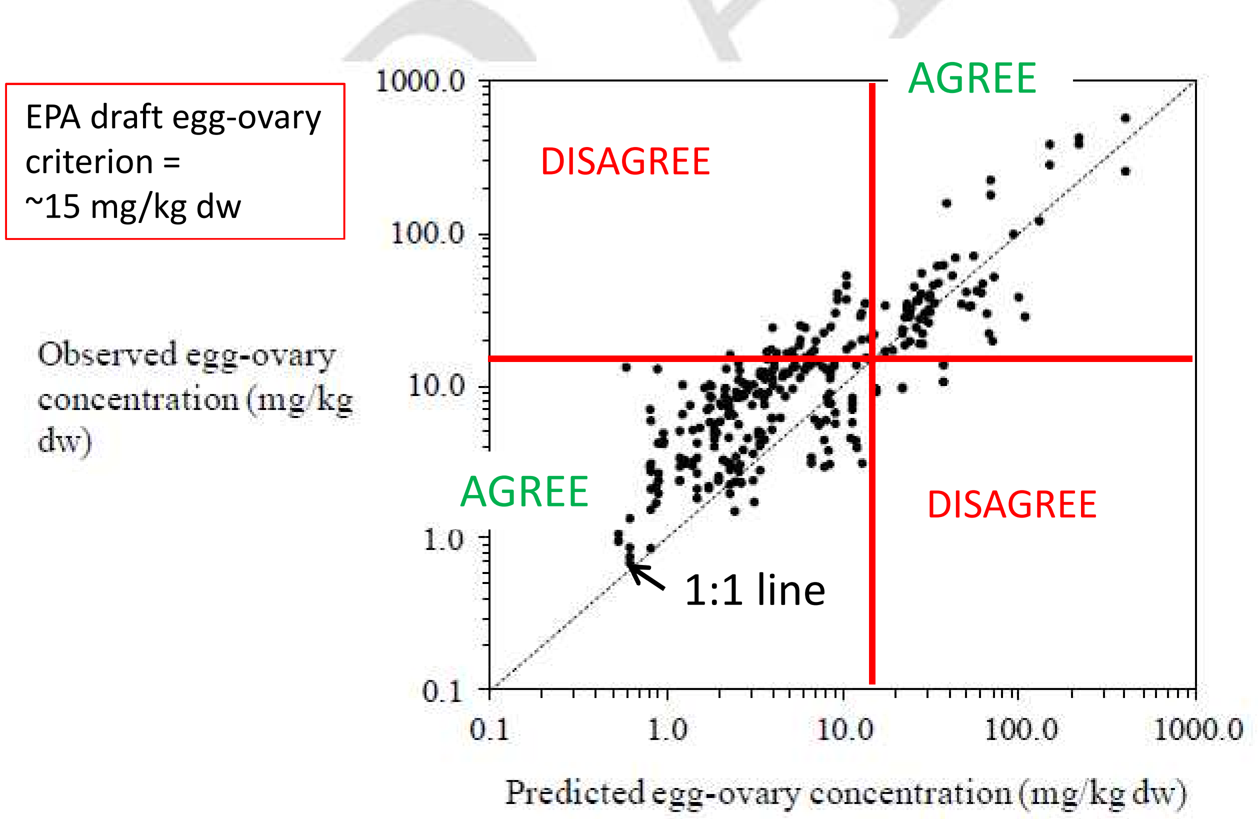
USEPA Draft Selenium Criterion:
Observed vs. Predicted Egg-Ovary Concentrations



Binary Classification Method Currently in USEPA Draft WQC Guidance

USEPA Draft Selenium Criterion:
Observed vs. Predicted Egg-Ovary Concentrations

$(r = 0.84, t_{(298)} = 26.28, P < 10^{-80})$.



See also Tables 18 & 19 in draft Se guidance document

A Biotic Ligand Model Example: Toxic Units as "Decision Context" for Evaluating the Fit of BLM Validation Data (McLaughlin 2015)

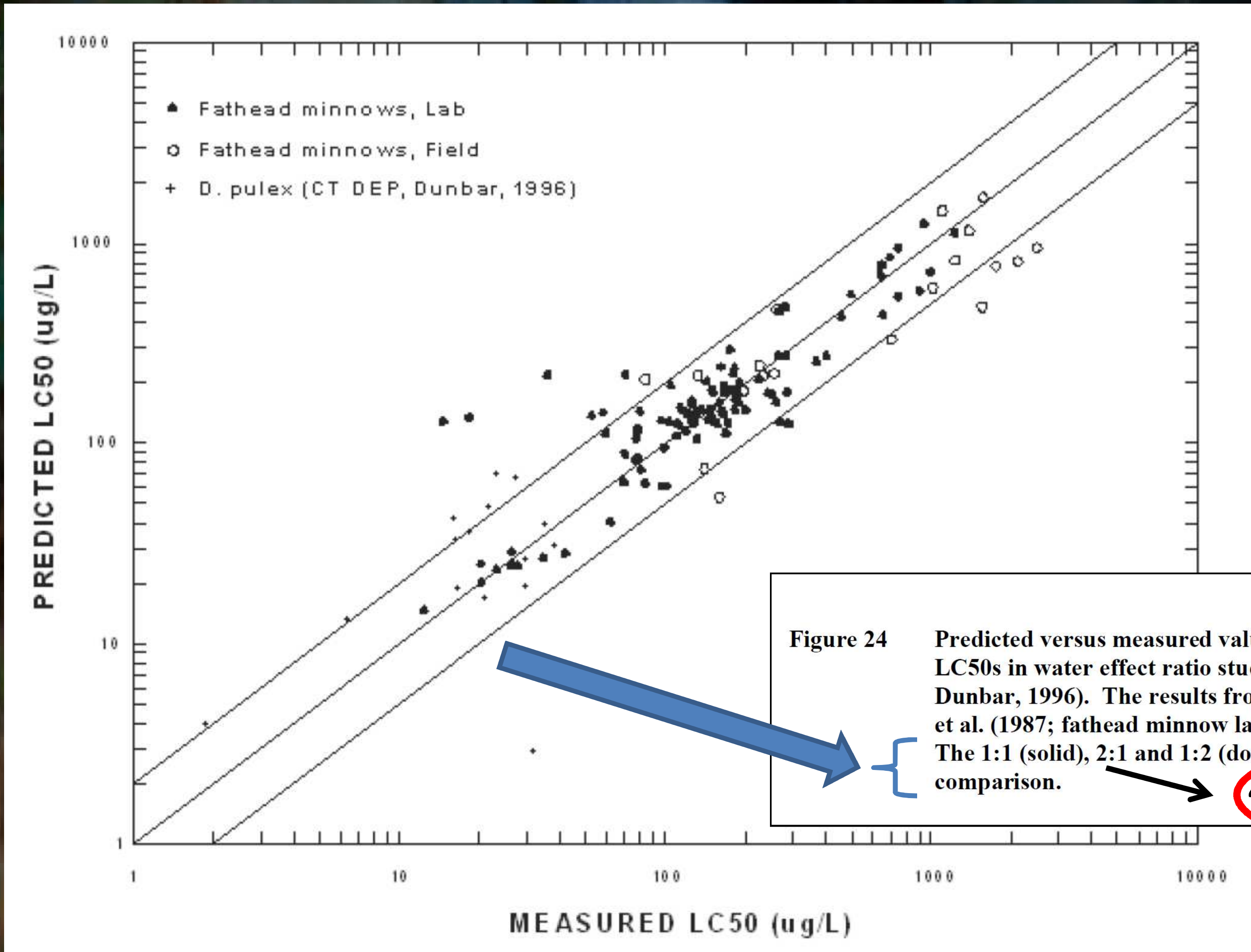
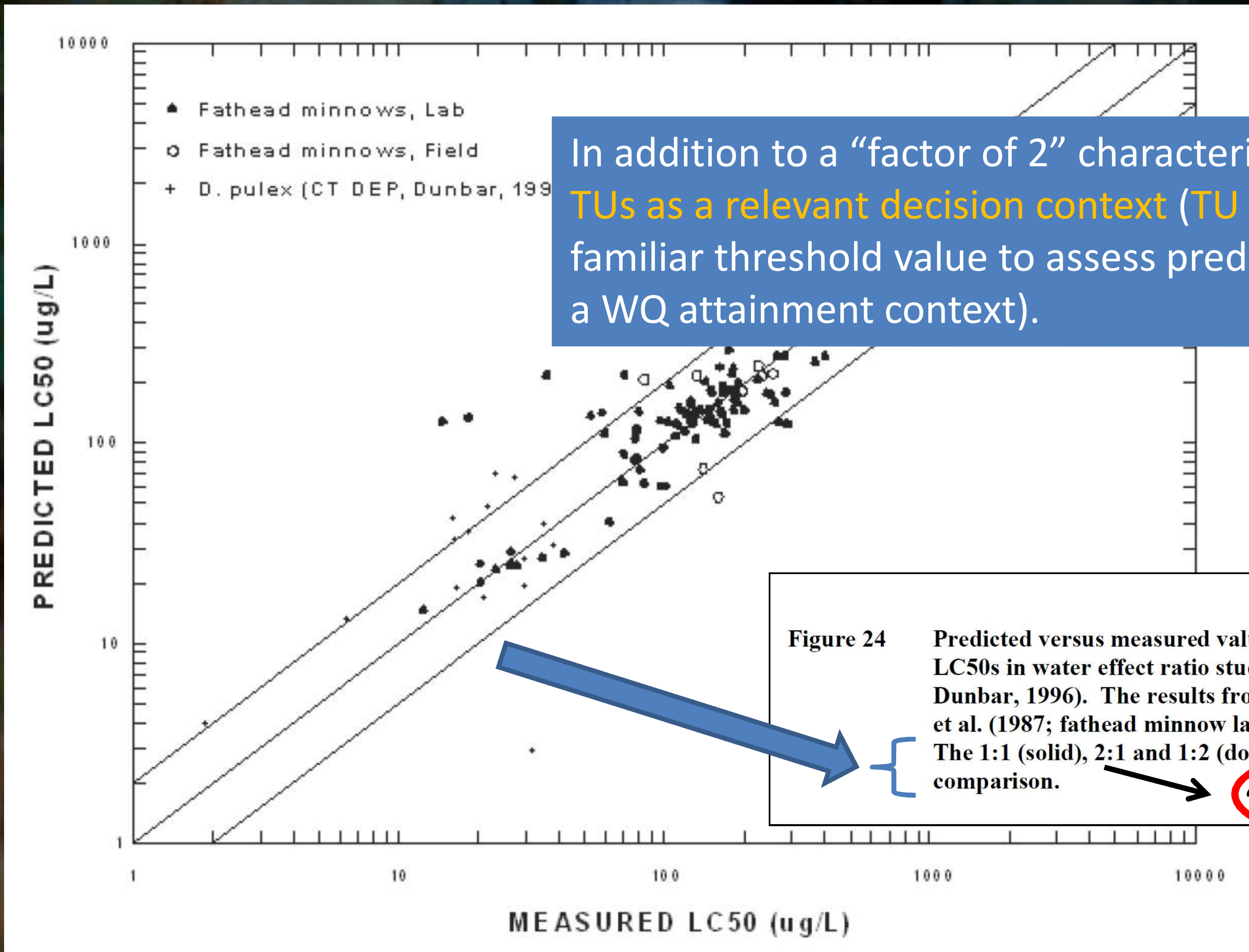


Figure 24 Predicted versus measured values for fathead minnow copper LC50s in water effect ratio studies (Diamond et al., 1997; Dunbar, 1996). The results from static exposures from Erickson et al. (1987; fathead minnow lab) are included for comparison. The 1:1 (solid), 2:1 and 1:2 (dotted) reference lines are drawn for comparison.

"…within a factor of 2"

A Biotic Ligand Model Example: Toxic Units as "Decision Context" for Evaluating the Fit of BLM Validation Data (McLaughlin 2015)



In addition to a "factor of 2" characterization, evaluate using TUs as a relevant decision context (TU = 1 provides a useful, familiar threshold value to assess predictive performance in a WQ attainment context).

Figure 24 Predicted versus measured values for fathead minnow copper LC50s in water effect ratio studies (Diamond et al., 1997; Dunbar, 1996). The results from static exposures from Erickson et al. (1987; fathead minnow lab) are included for comparison. The 1:1 (solid), 2:1 and 1:2 (dotted) reference lines are drawn for comparison.

"…within a factor of 2"

Use "What If" Metal Concentration Scenarios to Characterize Scatter in BLM Validation Data as TUs

- Unified Zn BLM, $ZnOH^+$ binding constant -2.4 (DeForest and Van Genderen 2012)
- Three simulations (~55%, ~80%, ~95% within "factor of 2")
  - values generated using loglinear regression equation, adjusting variance around "perfect fit" line
- Each evaluated at hypothetical ("what if") metal concentration equal to low (10th %ile), medium (50th %ile), and high (90th %ile) of the "observed" EC_x values
- Evaluated using both ROC and linear regression prediction limit approaches.
- Uses laboratory EC_x data to represent "true" toxicity, to be compared with BLM-derived EC_x predictions to characterize BLM performance.

Convert EC_x to Toxic Units: "What If" Metal Concentration Scenarios

$$TU_{obs,i} = \frac{M_{diss}}{ECx_{obs,i}}$$

$$TU_{pred,i} = \frac{M_{diss}}{ECx_{pred,i}}$$

To make the conversion to TU for each "what if" scenario, choose a dissolved metal concentration, M_{diss}, equal to a percentile of interest (e.g., 10th, 50th, 90th etc.) of the observed EC_x data

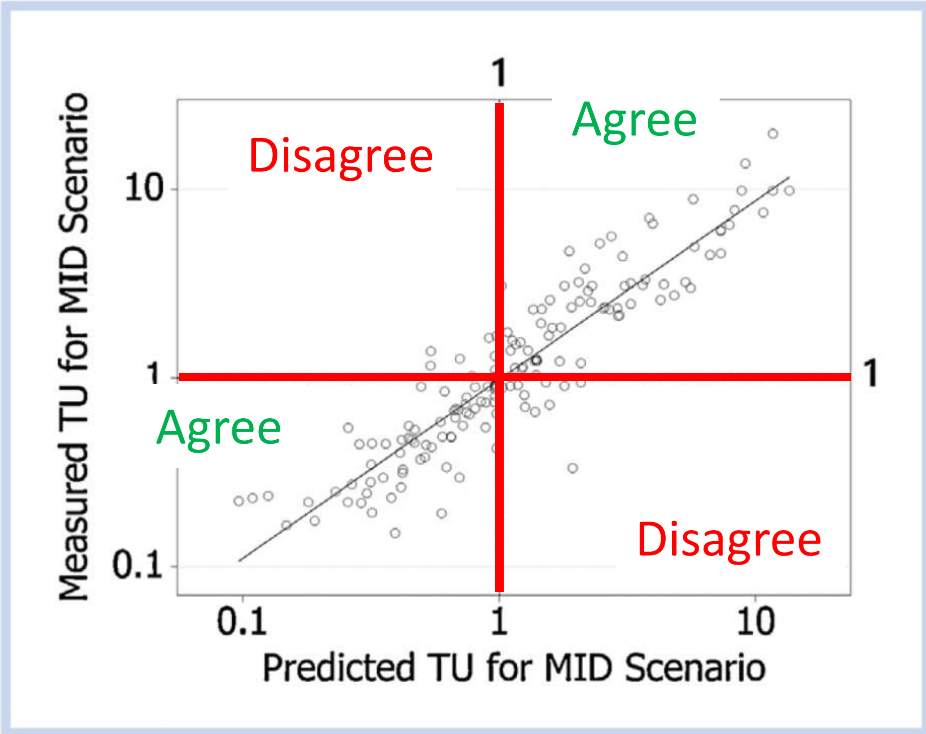


Figure 2. ECx validation data from DeForest and Van Genderen (2012) plotted as toxic units after conversion assuming a Zn concentration equal to the 50th percentile of the measured ECx data (the MID scenario).

(from McLaughlin 2015)

Can Evaluate "Gray Region" Using ROC Error Rates and/or a Regression Limit Prediction Interval Approach

- Can use probability plots of TU data to assess the size of the "gray region"
- Outside gray region, TU predictions have greater than the minimum specified level of confidence
- Stronger relationships lead to gray regions that cover a smaller fraction of the validation data

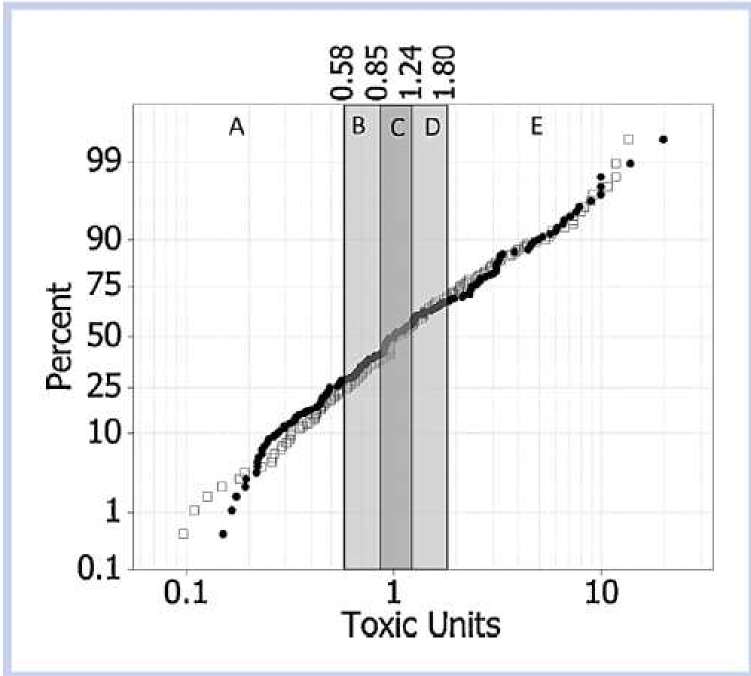


Figure 3. Cumulative probability plot of measured and predicted toxic units for the MID scenario using the unified Zn BLM validation data from DeForest and Van Genderen (2012). Open symbols: TU_{pred}; filled symbols: TU_{meas}. Shaded and labeled regions of the plot show gray region boundaries from prediction limit analysis (Table 4). A = probability of Type II error < 10% if TU < 0.58; B = probability of Type II error < 33% if TU_{pred} < 0.85; C = region where the probability of either Type I or Type II error exceeds 33%; D = probability of Type I error < 33% if TU > 1.24; E = probability of Type I error < 10% if TU > 1.80.

How Could This Uncertainty Information Be Used To Describe/Set Goals for Predictive Performance of WQC?

Some Ideas

- For a candidate criterion (e.g., FAV, CMC, CCC):
  - confidence limits
- For predictions based on a candidate criterion:
  - accuracy (ROC) > **X**%
  - false negative error rate < **Y1**%;
  - false positive error rate < **Y2**%
  - “gray region” covers less than **Z**% of the measured/response data

Some Concluding Thoughts

- "Splitting" continuous data into categories reduces the amount of information in the original data, so ROC or other classification methods should supplement, not replace, traditional statistical methods

- Explicit quantitative uncertainty analysis in water quality criteria derivation can:
  - Improve scientific defensibility and transparency
  - Promote the consideration of multiple types of decision errors
  - Help drive improvements to criteria-based predictions and management decisions

Questions?

Contact Information:

Doug McLaughlin, Ph.D.
Principal Research Scientist
National Council for Air & Stream Improvement, Inc.
Northern Regional Center
Kalamazoo, MI, USA

Phone: 269-276-3545
dmclaughlin@ncasi.org or
douglas.mclaughlin@wmich.edu

Citations/Further Reading

- DeForest DK, Van Genderen EJ. 2012. Application of US EPA guidelines in a bioavailability-based assessment of ambient water quality criteria for zinc in freshwater. Environ Toxicol Chem 31:1264–1272.

- McLaughlin, DB. 2015. Assessing the fit of biotic ligand model validation data in a risk management decision context. Integrated Environmental Assessment and Management, DOI: 10.1002/ieam.1634

- McLaughlin, DB. 2014. Maximizing the accuracy of field - derived numeric nutrient criteria in water quality regulations. Integrated Environmental Assessment and Management 10 (1): 133-137.

- McLaughlin, DB. 2012a. Estimating the designated use attainment decision error rates of USEPA's proposed numeric total phosphorus criteria for Florida colored lakes. Integrated Environmental Assessment and Management 8(1):167-174.

- McLaughlin, DB. 2012b. Assessing the predictive performance of risk-based water quality criteria using decision error estimates from ROC analysis. Integrated Environmental Assessment and Management. 8(4): 674-684.

- McLaughlin, DB and V Jain. 2011. Using Monte Carlo Analysis to characterize the uncertainty in final acute values derived from aquatic toxicity data. Integrated Environmental Assessment and Management 7(2): 269-279.

- Reiley MC, Stubblefield WA, Adams WJ, Di Toro DM, Hodson PV, Erickson RJ, Keating FJ Jr, editors. 2003. Reevaluation of the state of the science for water quality criteria development. Pensacola (FL): Society of Environmental Toxicology and Chemistry. 197 p.

- USEPA. 2010. Using Stressor-response Relationships to Derive Numeric Nutrient Criteria. EPA-820-S-10-001. November.

- USEPA. 2005. Summary Minutes of the U.S. Environmental Protection Agency (EPA) Science Advisory Board (SAB) Aquatic Life Criteria Guidelines Consultative Panel Meeting, September 21, 2005 Washington, D.C.