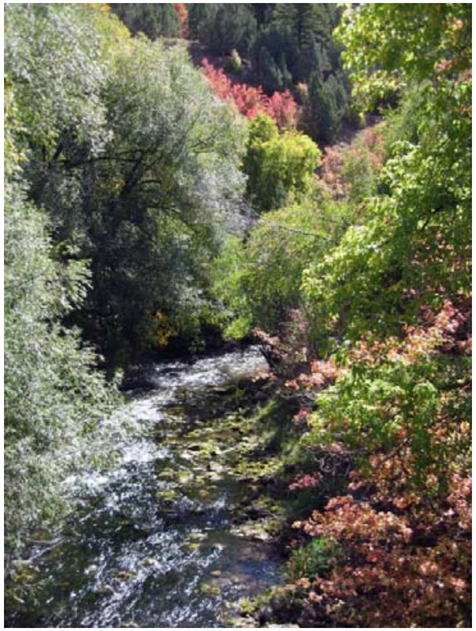
Development and Implementation of a Systematic Biological Monitoring and Assessment Program for Utah's Streams

Jeff Ostermiller, et al.
jostermiller@utah.gov

Division of Water Quality

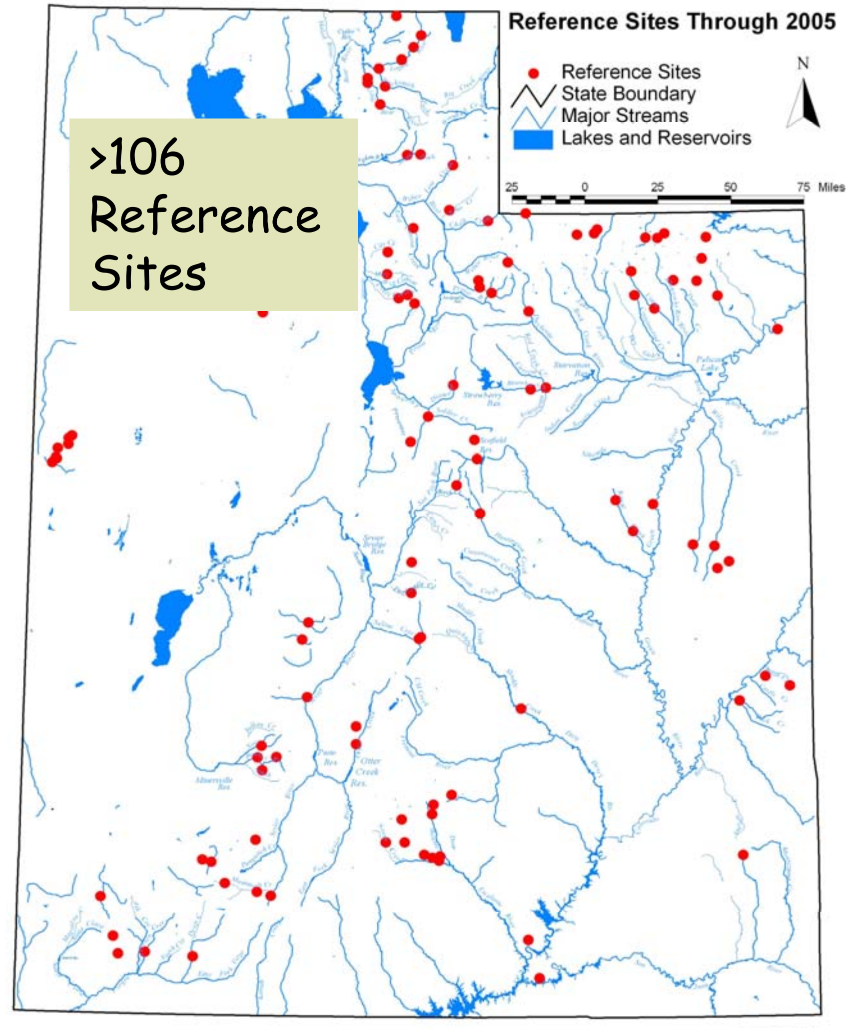
Setting the Context: About Utah

- Utah is the 2nd driest state in the USA.
- Annual precipitation varies from <5" in the deserts to 500" in the mountains.
- Average snowfall in the mountains is >500".
- Utah has ~14,250 miles of perennial streams.
- Many streams are regulated to capture winter snow melt.
- About 65% of Utah is on federal public lands.



Setting the Context: About Utah's Biological Assessment Program

- Until recently, Utah's monitoring and assessment program was very chemical centric.
- About 5 years ago, Utah saw the inevitable and created a position to develop a formal biological monitoring and assessment program.
- We started with the data from the western EMAP and embarked on a significant reference site monitoring plan.
- We now have formal biological assessment tools and assessments were incorporated into the *Integrated Report* this year.



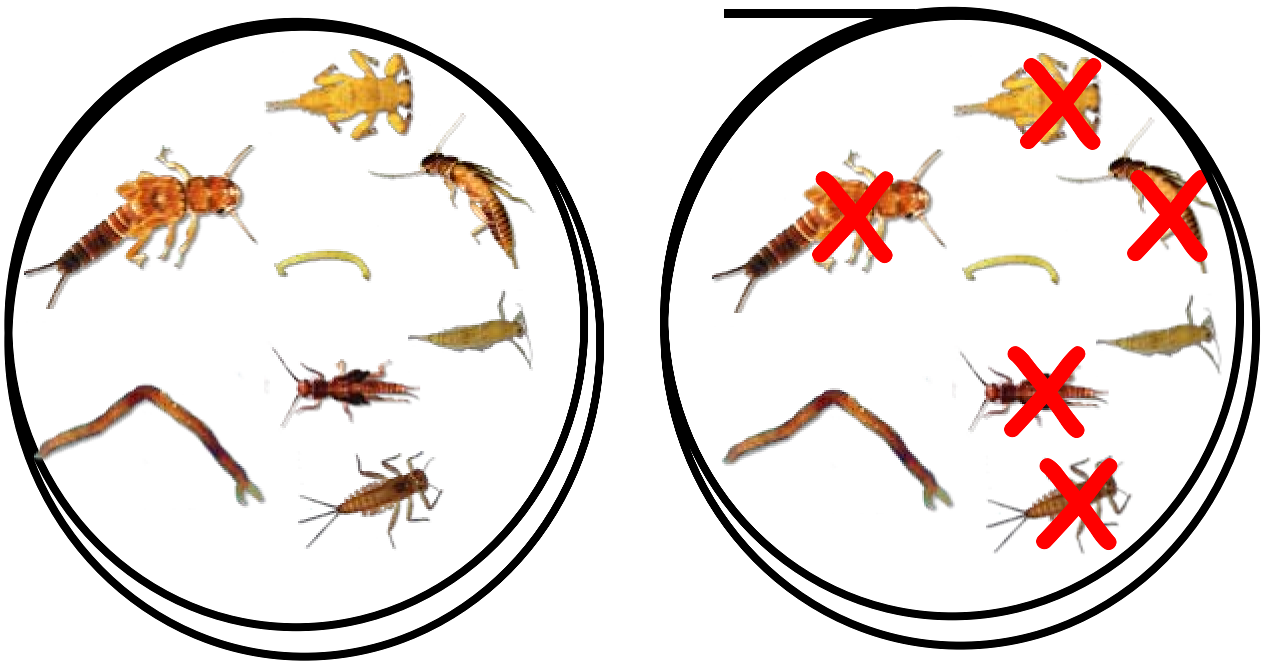
So what did we do with these data?

Step 1 – Use the most rich data source, macroinvertebrates to develop scientifically defensible biological assessment tools.

We tested a number of approaches and settled on a RIVPACS-type model.

What is O/E?

O/E is a measure of the taxonomic completeness of the biological community observed at a site



O/E
0.38

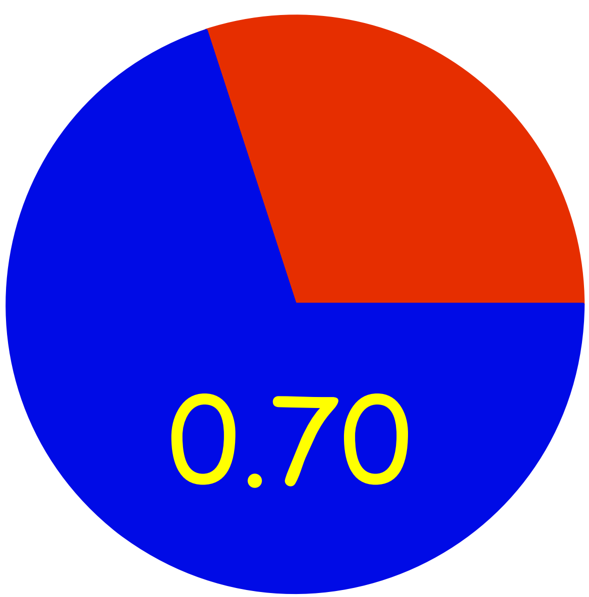
Expected taxa = 8 Observed taxa = 3

O/E standardizes assessments across sites that differ naturally in the number of expected taxa

Desert Site

O = 7

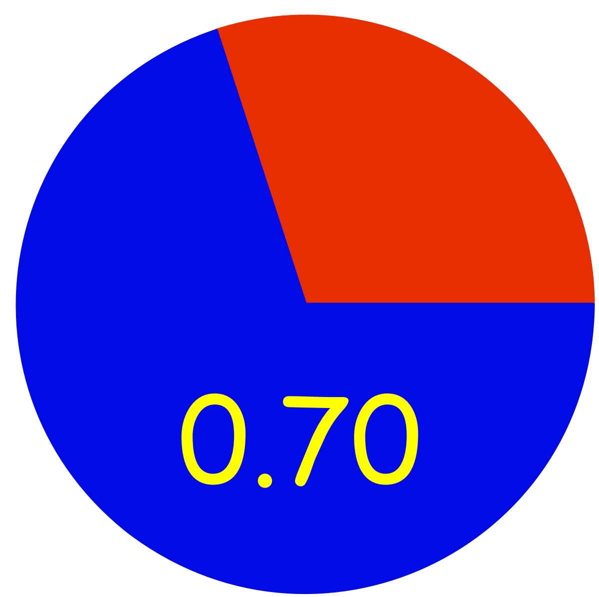
E = 10



Mountain Site

O = 21

E = 30



Predicting E

I'll spare you the details, but…

Step 1: Classify reference sites into biologically similar groups

Step 2: Use a discriminant function procedure to create a model that predicts the probability that new sites belong to each group of reference sites.

Step 3: The probability of capturing each taxon is calculated as the product of the probability of group membership and the frequency of occurrence within each group (% of sites).

Step 4: Calculate E by multiplying the probabilities of capturing all taxa.

E is the number of taxa, on average, expected at a non-degraded site, given that sites specific characteristics.

A Quick Note About Random Forests

Random Forests (RF) are statistical learning procedures (sensu Neural Networks) that yielded substantial improvements in precision and generality over the Multiple Dicriminant Function procedures typically used to generate RIVPACS models.

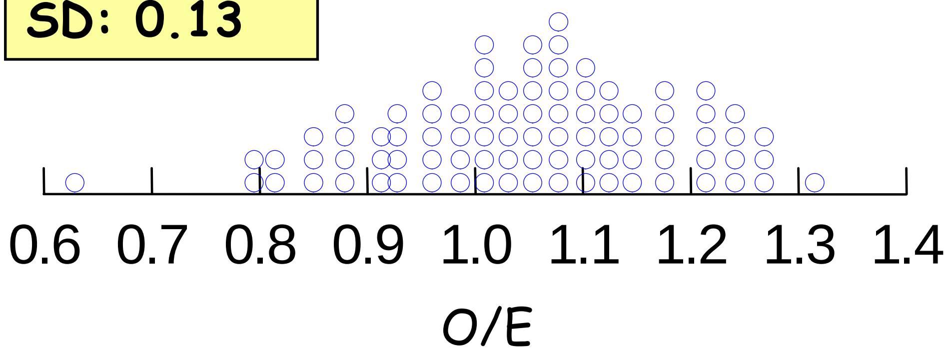
Measuring O

O is simply the number of species that we observe at a site that models predicted to occur.

Fall RF Model

Overall the model performed quite nicely!

Mean: 1.05
SD: 0.13

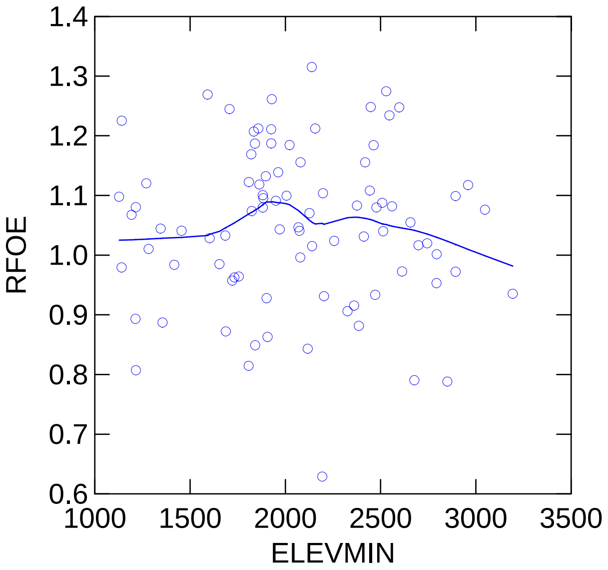


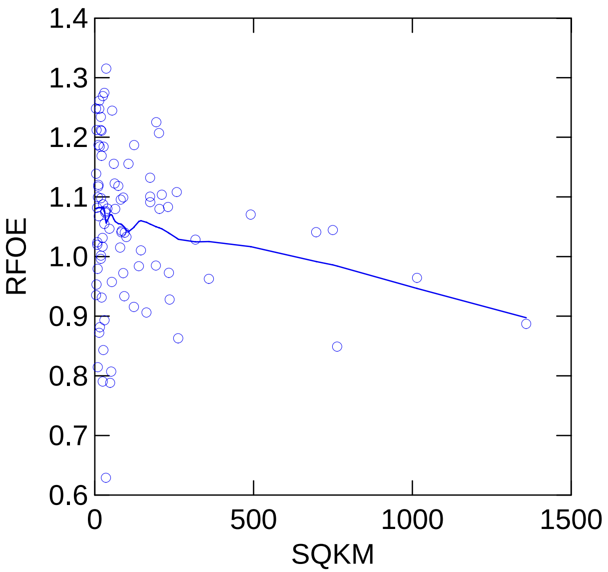
15 Variables Used

- MINWD.WS
- BDH.AVE
- G.PH.STD
- AWCH.AVE
- GPT.VOLC
- ELEV.MAX
- FST32AVE
- MEANP.PT
- SQ.KM
- TMEAN.WS
- MINP.PT
- ELEV.WS
- SLOPE.GIS
- LST32AVE
- TMEANNET

weather, soils, geology, slope, elevation, and watershed size

Testing Model Generality

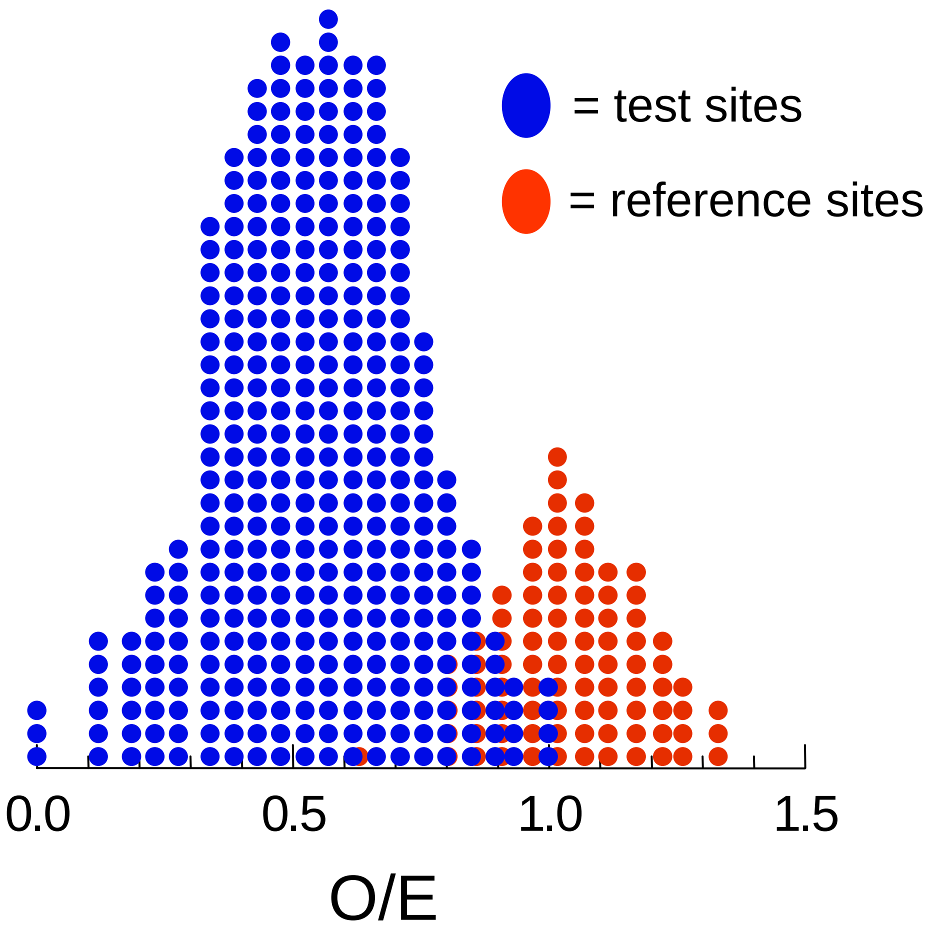




Models were evaluated against a number of environmental gradients, and we generally found that they performed about the same everywhere.

However, we do lack reference information at large rivers. Who would have guessed?

What did we find when we applied the models to other sites?



- A total of 444 samples from 234 sites were evaluated with the model, including the 89 reference sites initially used in model construction.

- The bimodal distribution of these data is likely a reflection of using a targeted sampling design.

- Test sites were selected to address water quality concerns, not overall conditions throughout Utah.

So what does this mean?

Determining Impairment Thresholds

Biological Methods

- directly coupled to beneficial uses
- more amenable to multiple thresholds

but,

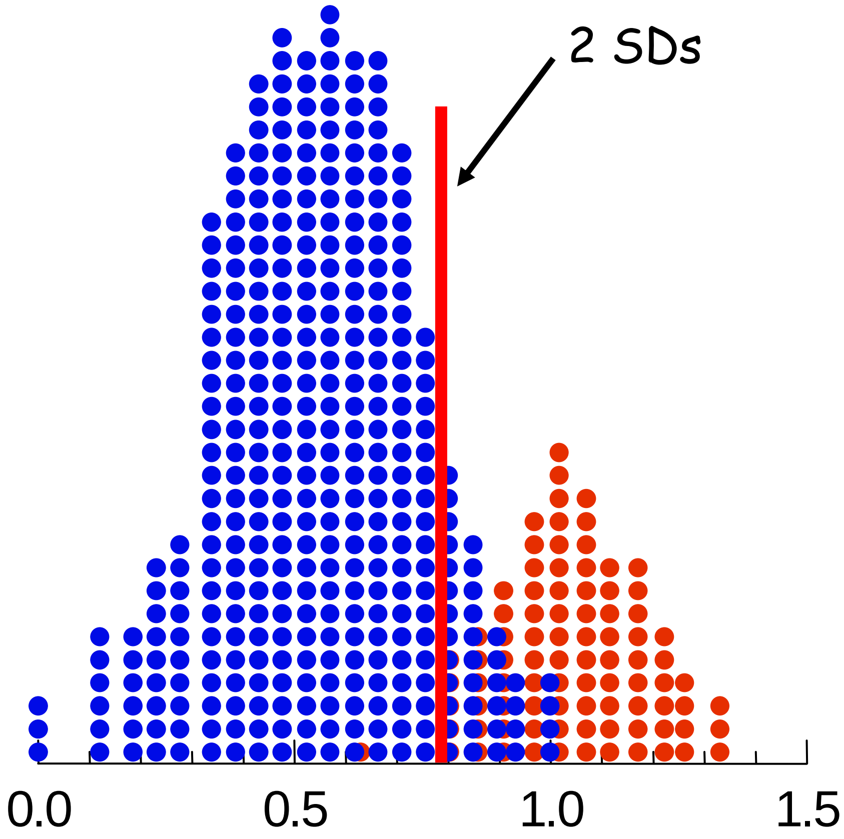
- somewhat arbitrary
- 'acceptable' losses may differ greatly among stakeholders

Statistical Methods

- More easily defensible
- Thresholds are tied to known model error

(also modeling, but we just are not there yet)

Statistical Thresholds



- We decided to use 2 standard deviations from an O/E of 1 as out first cut.
- Most reference sites are within this range.

Rule 1 – A site fails to meet its biological beneficial uses if O/E < 0.74

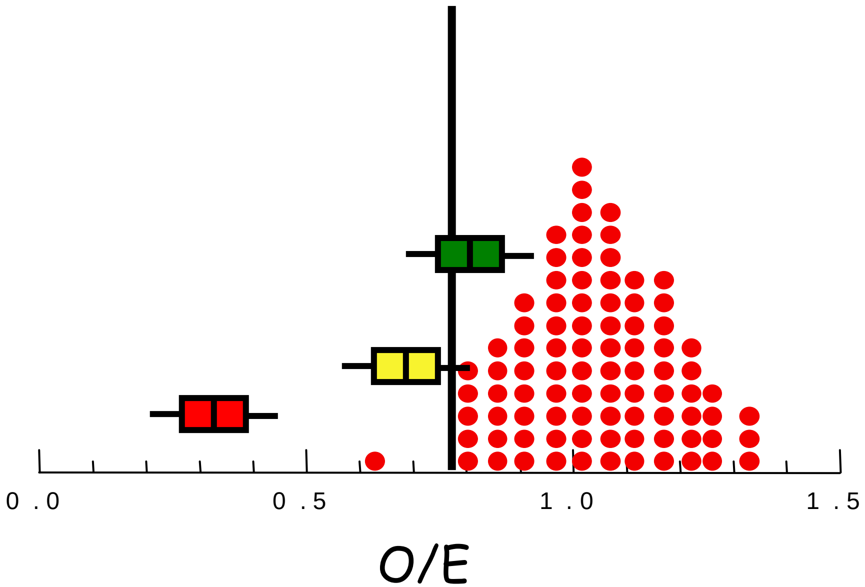
Is a single finding of impairment sufficient to list a site?

- In general, replication and an associated idea of variability is obtained before making environmental decisions.
- However, exceptions are sometimes made for biological assessments because the decision is based on overall condition (many species) and the sample is considered integrative.
- The danger of not replicating is increasing the number of false positives and false negatives, which would increase trepidation of a new program.

How much of an issue is this?

- In our case, anywhere from 1 to 8 samples were analyzed from each site.
- For those sites with >1 samples, average max-min O/E = 0.20

We suggest that, where possible, at least 3 samples be used for assessments.

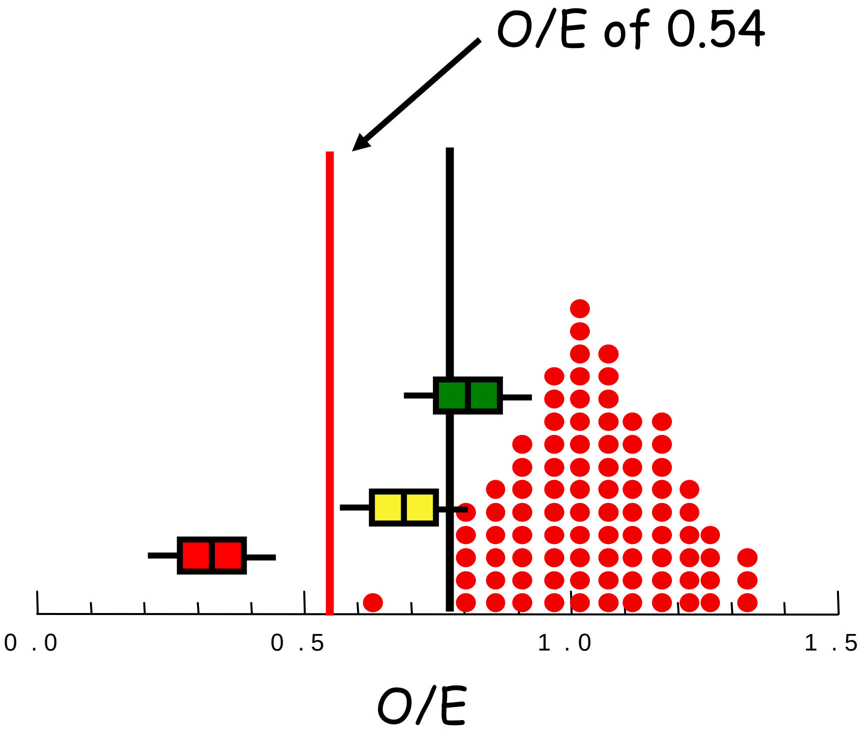


But if we follow this rule strictly, we could ignore problems for >3 years!

Moreover, such an approach is not really consistent with the 'weight of evidence' criterion.

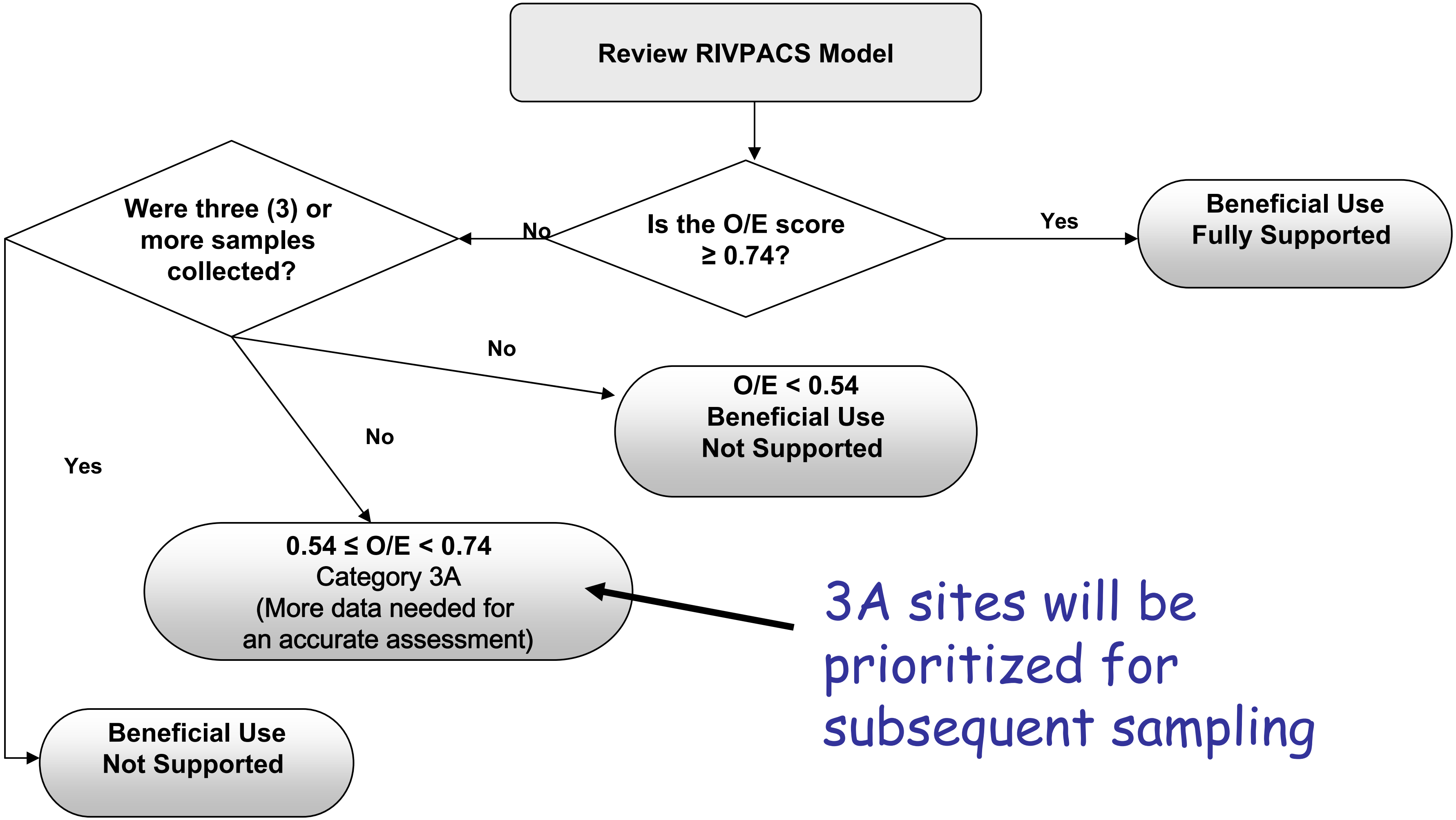
We need to consider the magnitude of impairment!

We evaluated a second threshold for sites where <3 samples are available to make assessments



If we subtract 0.20 from our initial threshold, how would the scenario play out?

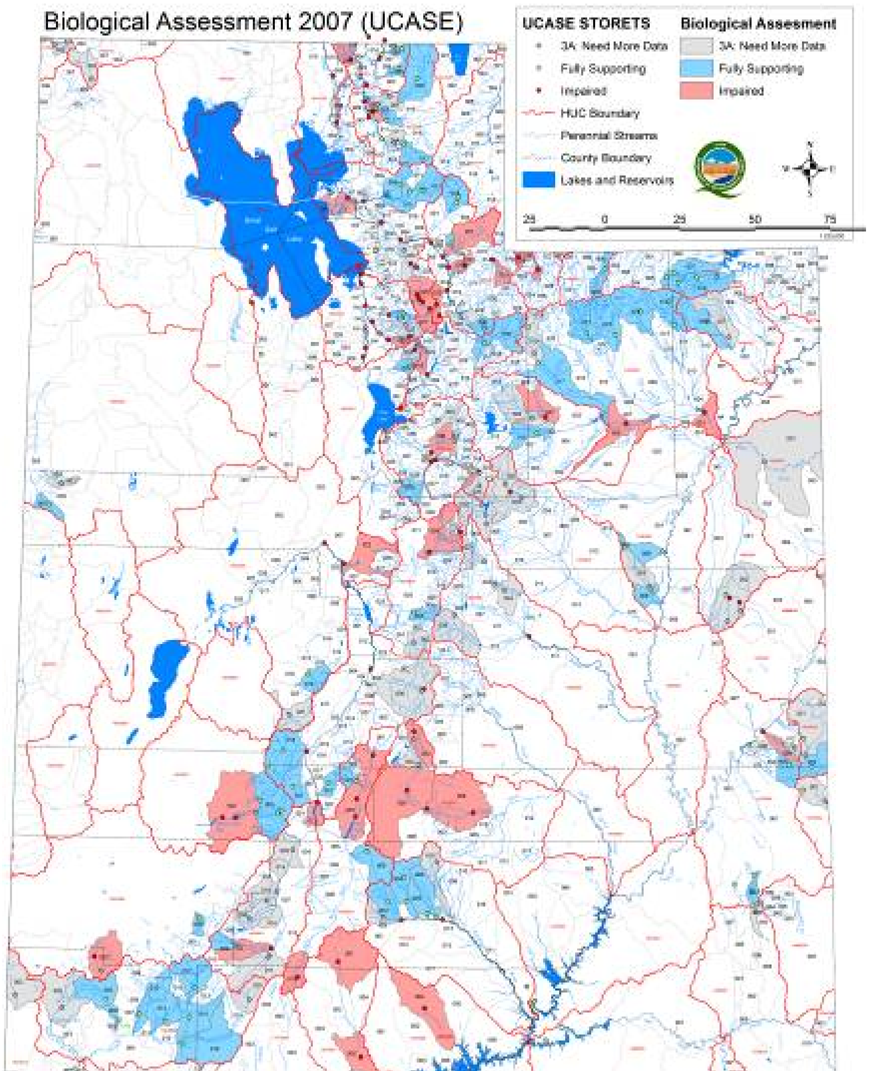
For our test run…
85% correct
7% incorrect impairment decision
9% incorrect full support decision



3A sites will be prioritized for subsequent sampling

Seems simple enough…

Our first question: What about >1 site in an Assessment Unit (AU) that differ in assessment decisions?



There were 18 AUs with >1 sample and disagreeing assessments.

In such cases, we can either:

1) average the scores
2) split the AU
3) conduct additional research to determine the best option

How to determine the 'ecological setting'?

- **Hydrology.** Any major changes in flow (i.e., diversions, reservoirs)?
- **Vegetation Changes.** Similar upland and riparian vegetation?
- **Land Ownership.** Shift from public to private lands?
- **Change in land use.** Especially if alterations to in-stream habitat were observed

Our final determination was based on a weight of evidence of all factors

When multiple samples were collected in an AU, were the samples from a clearly different ***ecological setting***?

Yes

- Split AUs if possible boundaries were immediately clear.
- If potential boundaries were not immediately obvious, then list as 3A and study appropriate boundaries for the future.

No

- Average assessments across samples within the AU

Final Results when Dealing with Multiple Sites within an AU with Differing Assessments

Total Number of Sites = 18

List as 3A and Collection Additional Data = 7

Split the AU = 10 (>50% of cases)

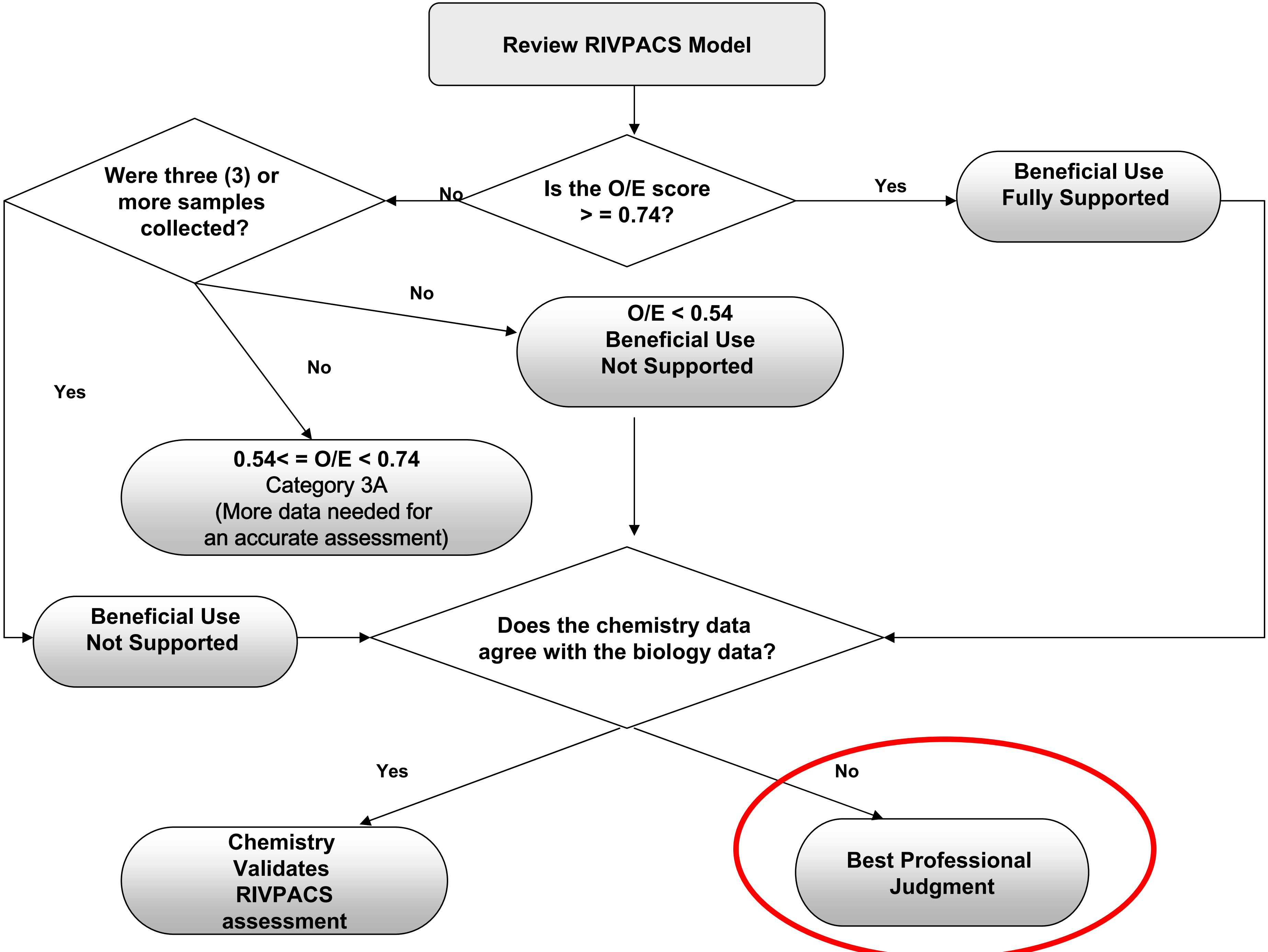
List as 4A (clear indication that the difference was habitat related) = 1

Take home message - We need to think hard about the 'pour point' assumption of site selection, especially with biological assessments.

Our next question: What to do when the chemical and biological assessments differ?

Interestingly, of 168 AUs that we assessed, there were only 18 cases where the chemistry and biology disagreed (not counting AU splits)!

Given our long history of chemical assessments, we thought that it was important to examine each of these disagreements individually.

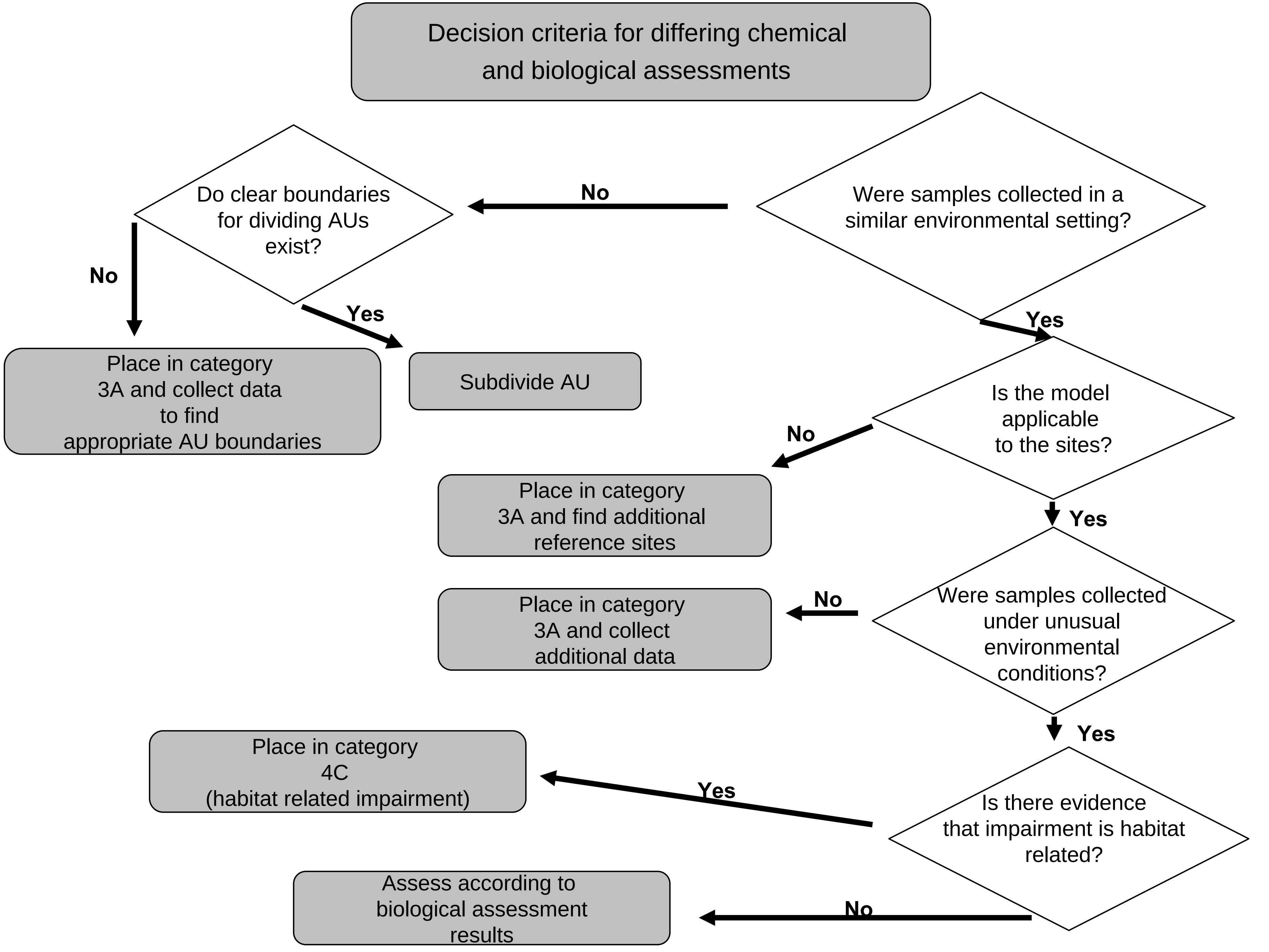

Review RIVPACS Model
Is the O/E score > = 0.74?
Yes
Beneficial Use Fully Supported
No
Were three (3) or more samples collected?
No
O/E < 0.54 Beneficial Use Not Supported
No
Yes
0.54< = O/E < 0.74 Category 3A (More data needed for an accurate assessment)
Beneficial Use Not Supported
Does the chemistry data agree with the biology data?
Yes
No
Chemistry Validates RIVPACS assessment
Best Professional Judgment

Our next question: What to do when the chemical and biological assessments differ?

Initially, these discussions were a little heated, but through the conflict we developed an objective, rational, and defensible process.

<u>Thanks to our professionals and their sound judgment:</u>

Tom T.
Mark S.
Ying Ying M.
Harry J.
Jeff O.
Dorrie P.


Decision criteria for differing chemical and biological assessments
Were samples collected in a similar environmental setting?
No
Do clear boundaries for dividing AUs exist?
No
Place in category 3A and collect data to find appropriate AU boundaries
Yes
Subdivide AU
Yes
Is the model applicable to the sites?
No
Place in category 3A and find additional reference sites
Yes
Were samples collected under unusual environmental conditions?
No
Place in category 3A and collect additional data
Yes
Is there evidence that impairment is habitat related?
Yes
Place in category 4C (habitat related impairment)
No
Assess according to biological assessment results

Final Results for chemistry and biology conflicts

Chemistry indicates supporting, but biology doesn't (N=11)

Why?

Strong indication of habitat-related issues = 4
Potential problems with the model or suspect samples = 5
From different environmental settings = 2

Result

List as Non-supporting = 3
List as 3A (additional data needed) = 4
List as 4C (impairment is clearly habitat-related) = 4

Biology indicates supporting, but chemistry doesn't (N=7)

Why?

Data collected under unusual, naturally-occurring environ. cond. = 4
From different environmental settings = 3

Result

List as fully supporting (remove from list) = 6
List as 3A (additional data needed) = 1

Our 'final' decisions (if we maintain this process)

Total number of AUs Assessed = 168

Fully Supporting = 80

Non-Supporting = 45

3A (more data needed) = 39

4C (clear indication impairment is habitat related) = 4

Summary and Next Steps

Our approaches were systematic and transparent, which helped greatly in communicating this new program to stakeholders.

Our process would benefit from improving objectivity in as many of the assessment processes as possible.

We are currently in the process of developing assessment tools for diatoms, which will add another layer of ambiguity when these results do not agree with macroinvertebrates.

Additional research is needed into understanding what exactly it means when different approaches to assessing the same beneficial use do not agree.

Next Steps and Caveats

Consider using more robust statistical techniques to develop thresholds. 2 SD works for now, but still somewhat arbitrary.

We need to start working on similar tools for other assemblages.

We need to take these data and develop diagnostic tools.

These analyses represent the first step in our journey toward TALU.

Outreach always continues…

My son's first Halloween costume, a caddisfly!

Questions?