# Performance of Different Bioassessment Methods from California: Toward a Uniform Approach

David B. Herbst and Erik L. Silldorff

Sierra Nevada Aquatic Research Laboratory

and Princeton Hydro, Inc

# The most common methods (largest data sets) in use in California employ different methods of sample collection, laboratory processing, identification, and data analysis: what are the differences in the biological assessment results?

- Needs: data uniformity and sharing, geographic cover and program coordination →compare methods side-by-side
- Study Objectives: Use Performance-Based Methods System to contrast methods and evaluate precision, bias, discrimination in separating reference from test streams, and accuracy in assessing impairment (=PBMS)
- Study design: 40 streams, side-by-side field collections and lab processing, same Reference (25) and Test (15) sites, gradient of potential test impacts
- Multimetric & Multivariate analysis tools for each method

# Summary of Differences Between Methods
(all targeted riffle sampling)

| PROTOCOL: | UC-SNARL<br>**Lahontan** | CSBP<br>**Dept. Fish & Game** | R5.USFS-USU<br>**Forest Service** |
|---|---|---|---|
| Net type & mesh | D-frame, 250 µM | D-frame, 500 µM | D-frame, 500 µM |
| Replication | 5 composites of 3 | 3 composites of 3 | 1 composite of 8 |
| Area sampled | 1.39 $m^2$ (1x1) | 1.67 $m^2$ (1x2) | 0.74 $m^2$ (1x1) |
| Subsampling | Drum splitter | Grid Tray | Grid Tray |
| Enumeration | 250-500 count | 300 fixed count | 500 fixed count |
| Taxonomic Resolution | Genus / species (including midges and mites) plus large & rare | Genus / species (midges / mites to subfamily / family) plus large & rare | Genus / species (including midges and mites) plus large & rare |

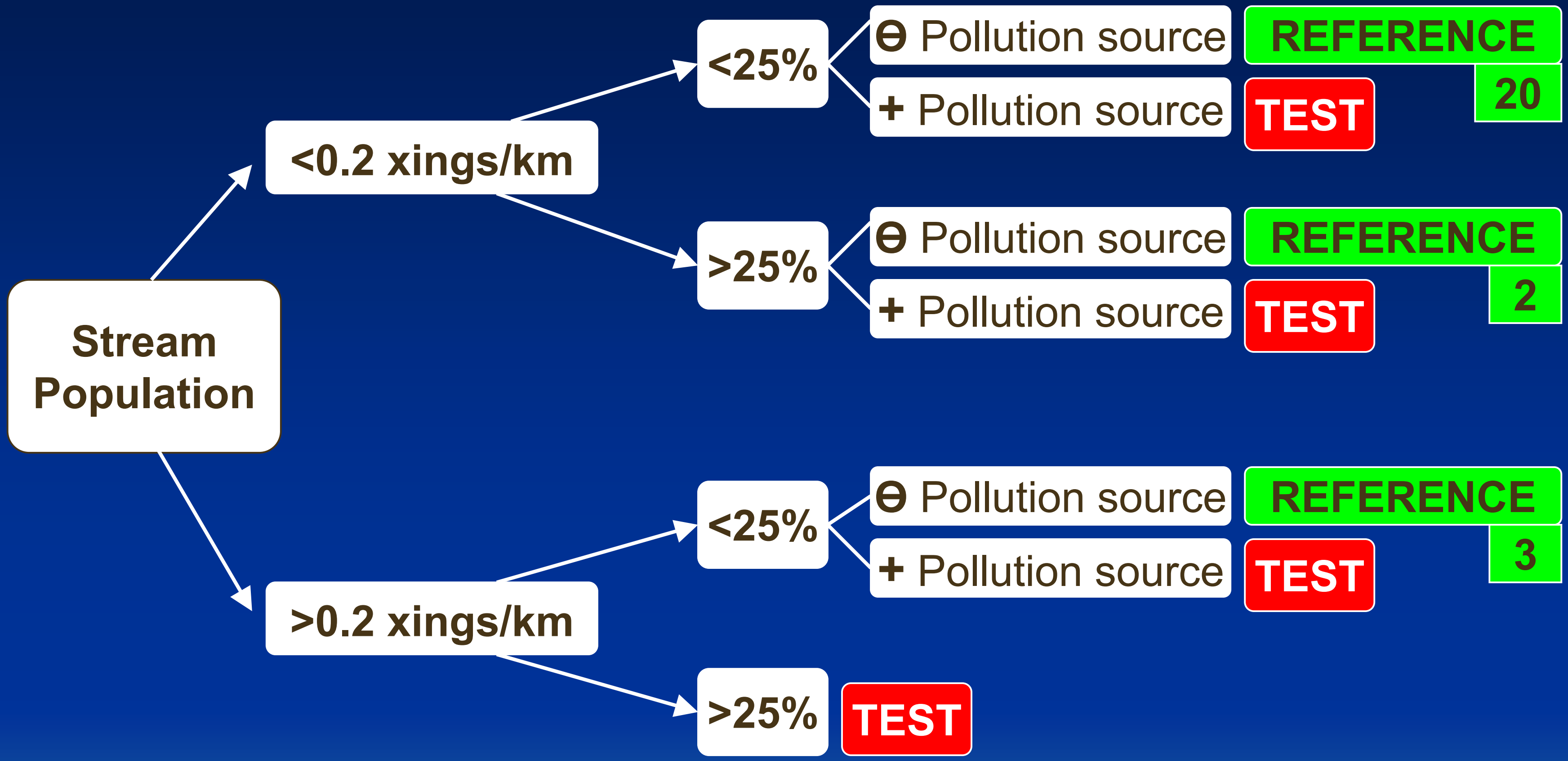
# PBMS Criteria for Comparisons

- Precision - coefficient of variation (CV) for metrics, and for IBI and O/E values for reference streams
- Bias – applicability to different stream classes: do CVs for the same measure differ between habitat types?
- Discrimination – separation of test and reference means
- Accuracy – trade-offs between type I and type II error rates in obtaining best assessment certainty

Plus other considerations:

- Correlations between methods for IBI and O/E scores (co-plots of stream scores, regressions among methods)
- Conversion options for standardizing data sets
- Relation of IBI ranks to environmental stress gradients

CRITERIA: Upstream road crossings /km | Local bank erosion | Known pollution or degradation

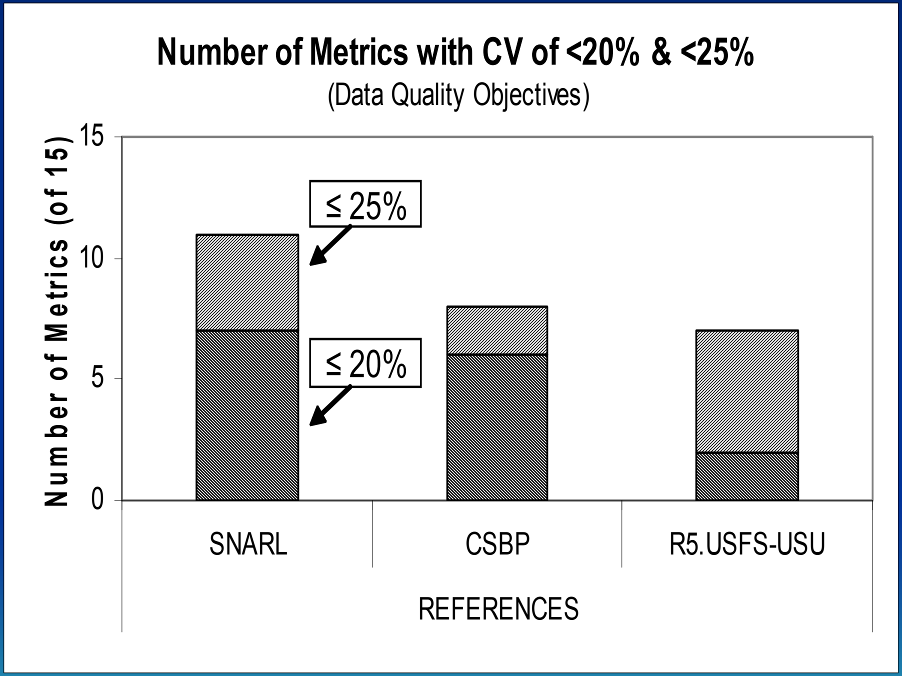


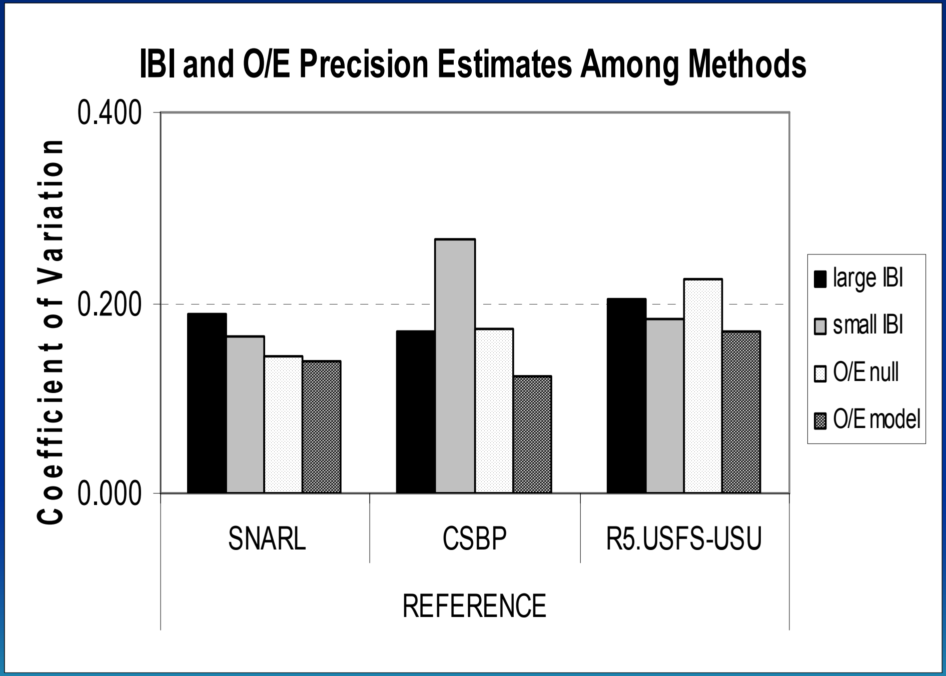
- Stream Population
  - <0.2 xings/km
    - <25%
      - ⊖ Pollution source → REFERENCE 20
      - + Pollution source → TEST
    - >25%
      - ⊖ Pollution source → REFERENCE 2
      - + Pollution source → TEST
  - >0.2 xings/km
    - <25%
      - ⊖ Pollution source → REFERENCE 3
      - + Pollution source → TEST
    - >25% → TEST

## First Step: Screening and Selection of Reference Stream Sites

# Precision Differences:

- Though the SNARL method exhibits slightly better metric performance at DQOs of 20-25%, IBIs and O/Es for all are near or below this DQO

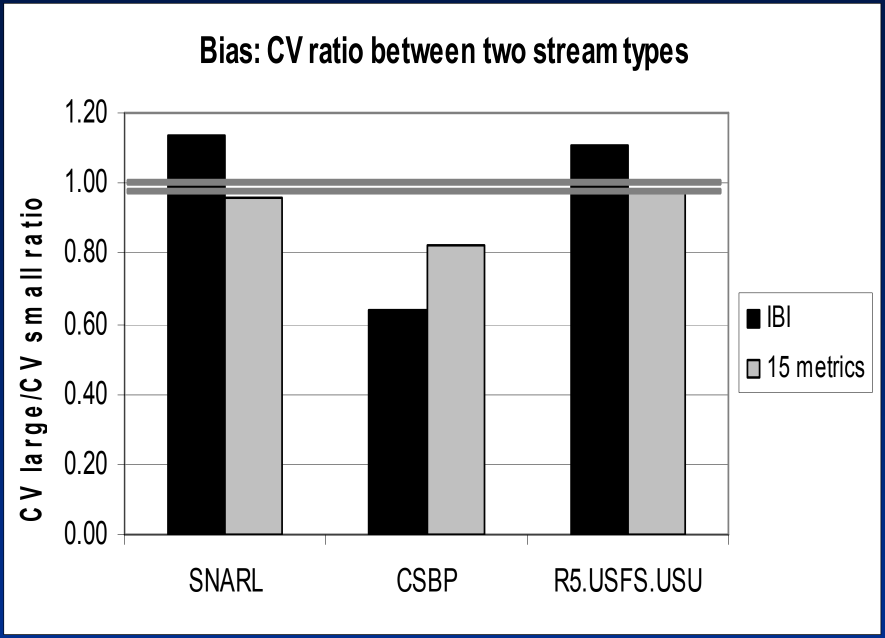




# Bias:

Do metrics have the same performance in precision for different types of streams or for different regions?

$CV_{large}/CV_{small}$ ratio = 1.0 if unbiased

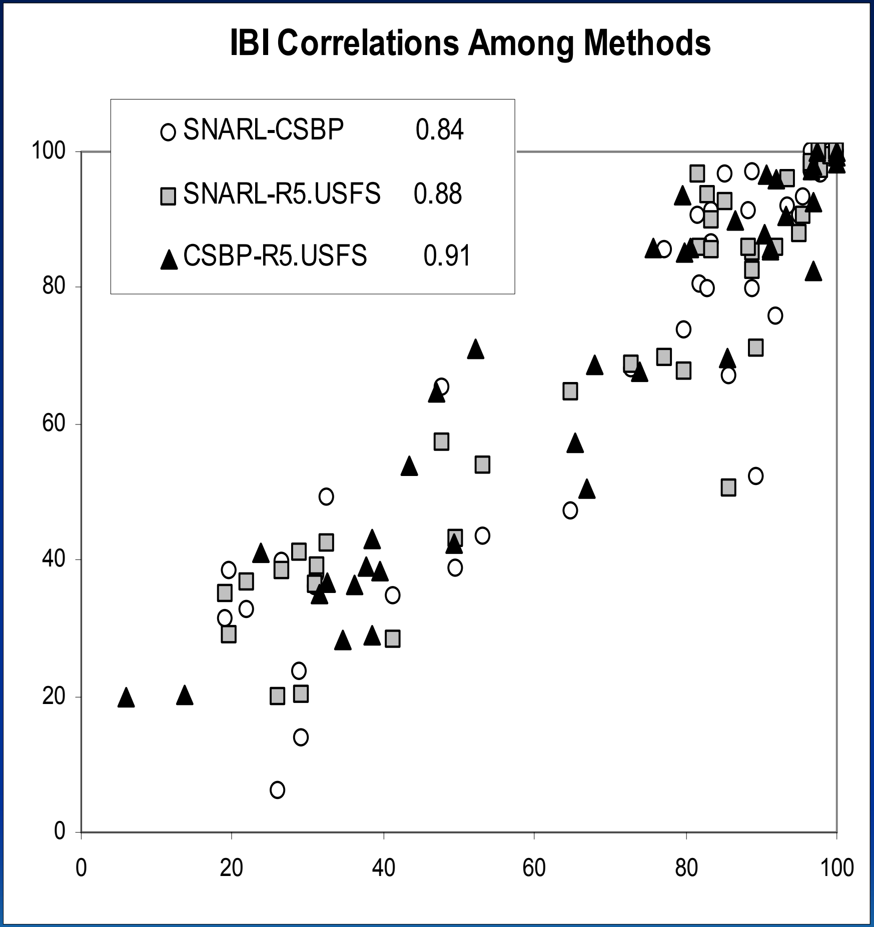


- Using either metric average or composite IBI, CSBP appears to show a bias as less precision for small streams
- SNARL & R5.USFS.USU show unbiased measures between stream types
- But does this matter in terms of method comparability or to discerning impairment?

# Correlations of IBI Scores Among Methods

- Scores highly correlated among methods
- Correlations indicate good agreement in assessments among methods (IBI $R^2 \geq 0.84$)

  >confidence that results are interchangeable (similar result for O/Es)

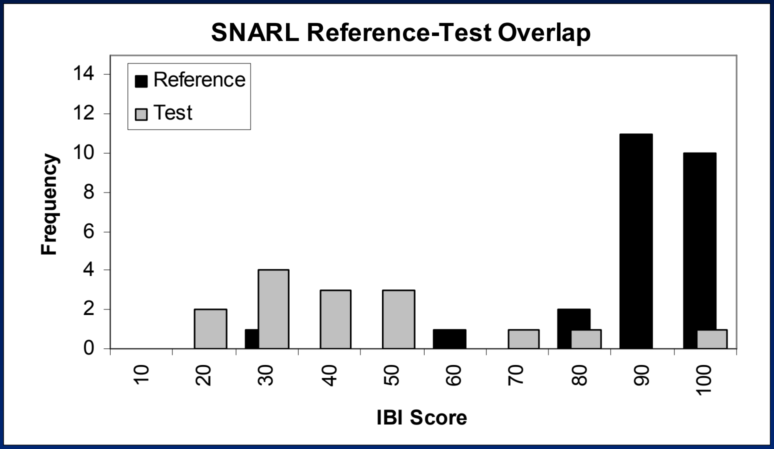


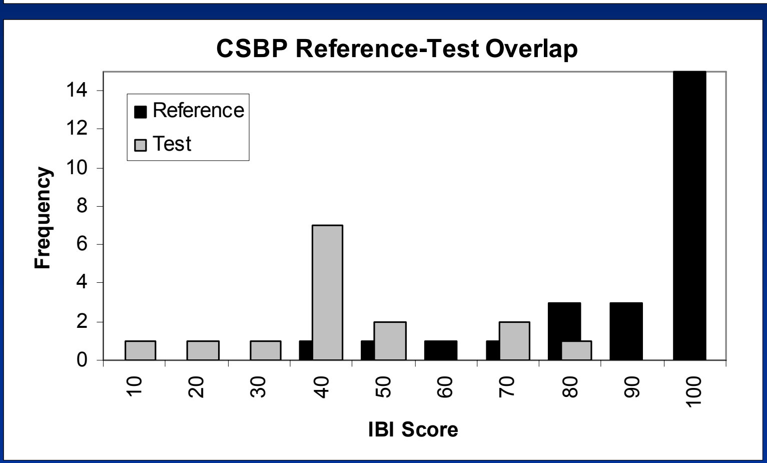
## Methods Exhibit Similar Discrimination of Impairment: Overlap of Reference and Test IBIs

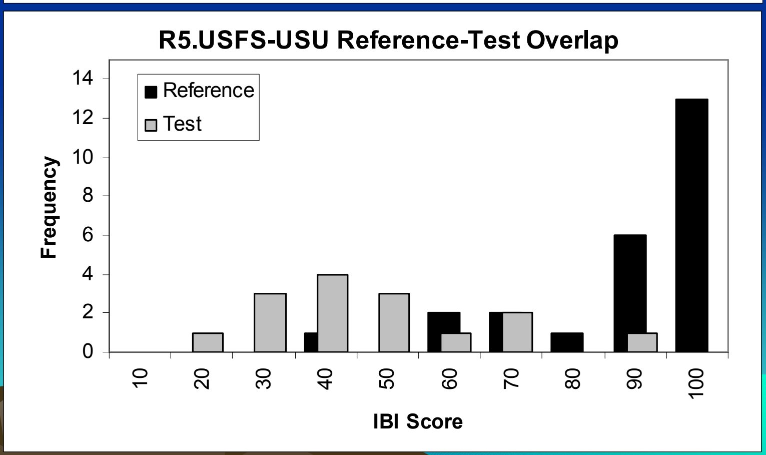
We want to discriminate reference from test: with the objective of minimizing type II error (=not detecting impaired sites)

so cut the tail of the reference range so that fewer test sites will overlap the reference, but not so much that unimpaired sites are detected as impaired (=type I error)

When error rates are optimized, all methods show about 80-90% accuracy in identifying truly impaired sites (type II error minimized), [=resource protection]
while at the same time only about 10-20% of low range reference sites are eliminated (type I error minimized,
few unimpaired sites are misjudged)
[=reasonable standards]

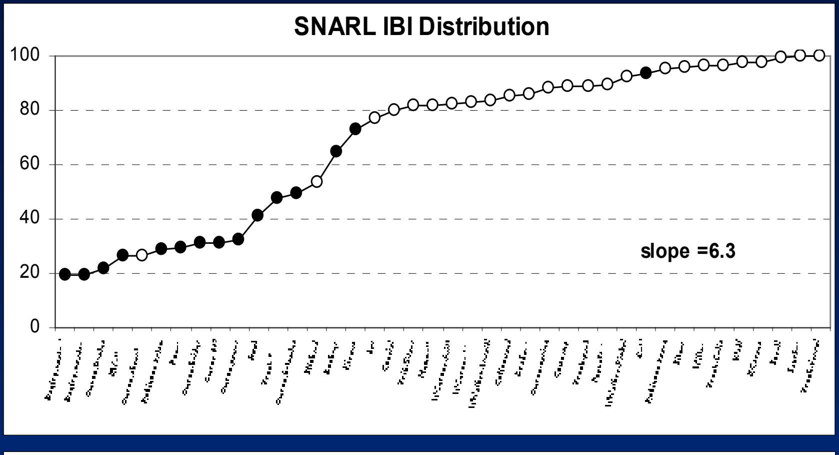

SNARL Reference-Test Overlap

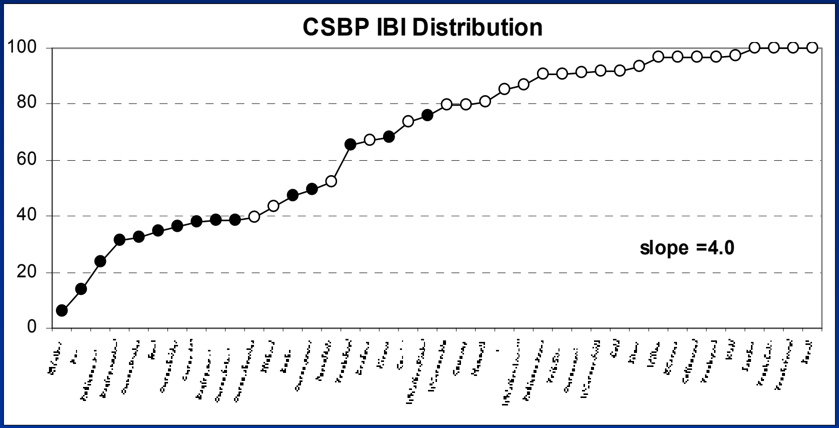

CSBP Reference-Test Overlap

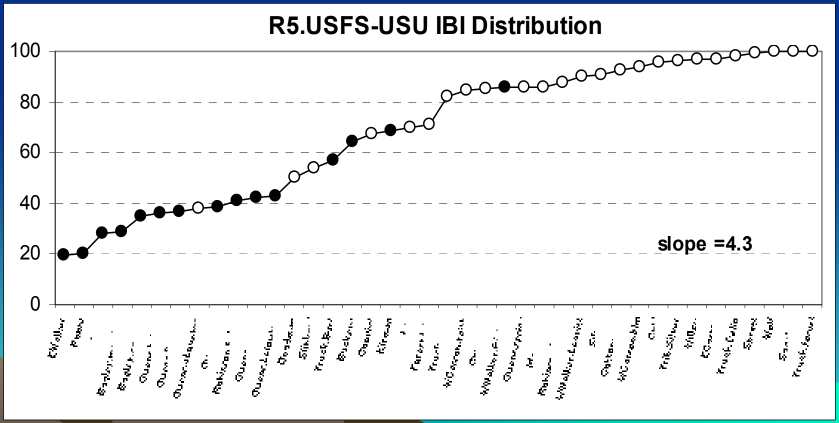

R5.USFS-USU Reference-Test Overlap

# Another way of looking at this: ranked distributions

- Note that the distributions follow approximate sigmoid forms, as in dose-response curves – with thresholds (slope and inflection points) more clearly defined by the SNARL method, and fewer references falling below the upper threshold
- How do these IBIs relate to environmental stress gradients?


SNARL IBI Distribution


CSBP IBI Distribution


R5.USFS-USU IBI Distribution

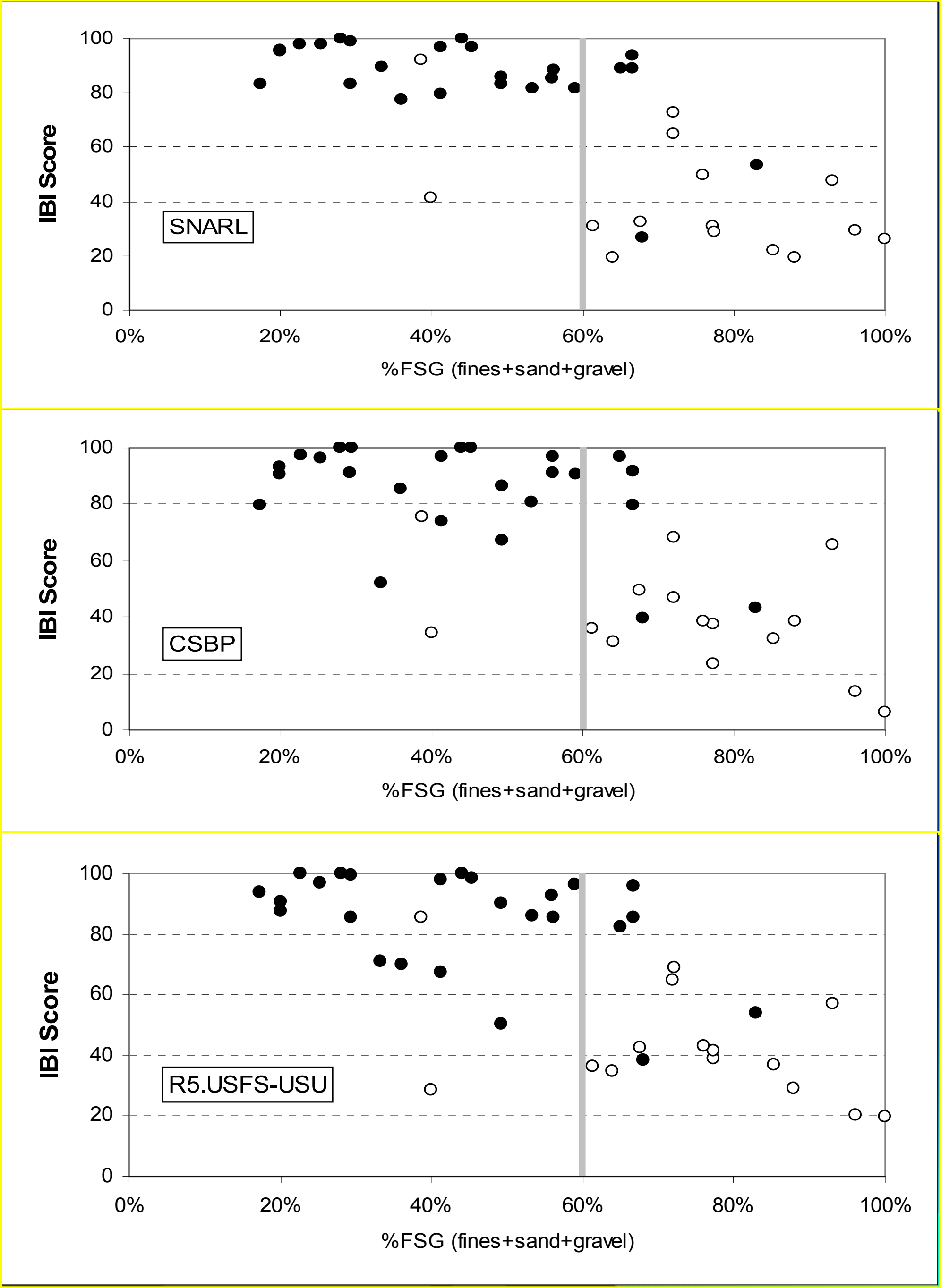
# Primary environmental stress gradient related to sedimentation (grazing and channel geomorphic alteration)

- SNARL data set is clearest in defining a sediment threshold for impaired integrity

  (as %fines+sand+gravel)

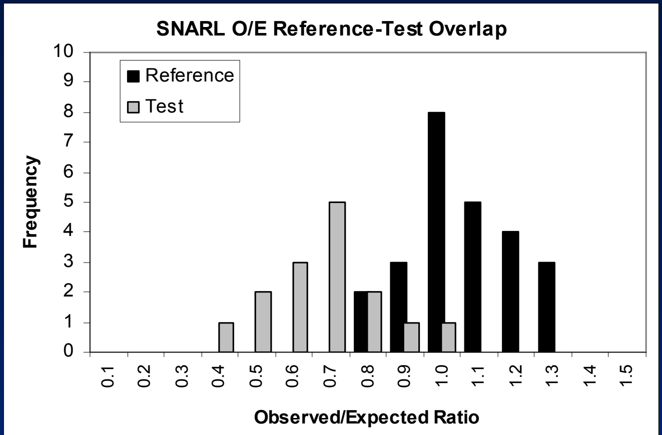
- Similar thresholds were found for riparian cover (below 30%)

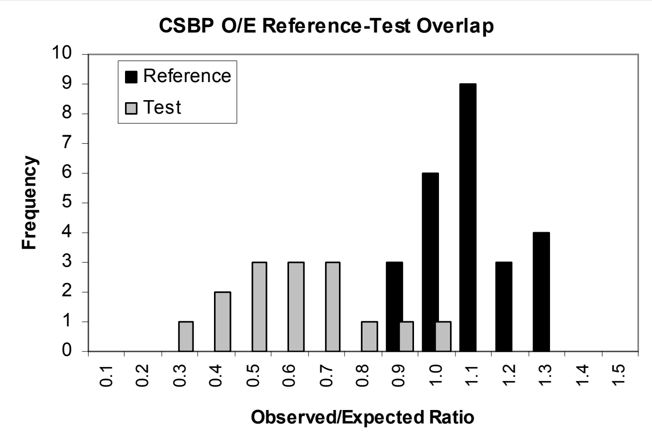
  and for conductivity (above 200 µS)

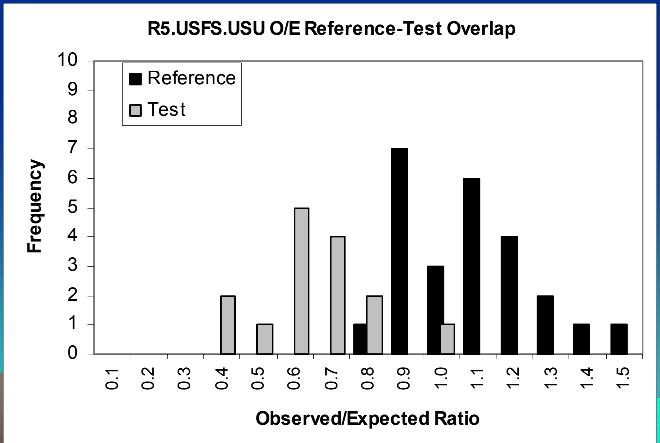


# Overlap of Reference and Test O/Es

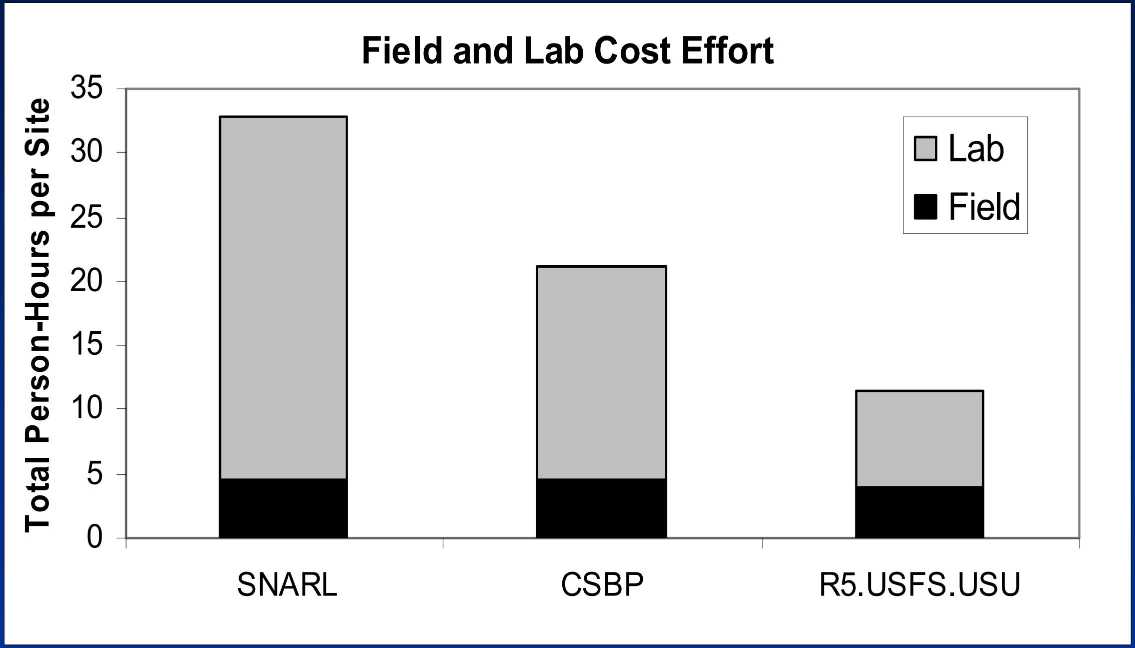
Again the methods are comparable in terms of the degree of overlap between the reference and test distributions – 15-25% of test range extends into reference


SNARL O/E Reference-Test Overlap


CSBP O/E Reference-Test Overlap


R5.USFS.USU O/E Reference-Test Overlap

# How do methods compare in terms of cost?



- SNARL method with 5 replicate riffle samples taken per site is about 1.5X the cost-effort of CSBP and 3X that of the single R5.USFS.USU targeted riffle composite sample

# Data Set Conversion: SNARL into 500 fixed-count

- Different calibration options:
  - 1. simple randomized re-sampling to 500 count
  - 2. re-sampling to 500 including large+rare taxa
  - 3. re-sampling to 500 with reduced proportion of small taxa collected by 250 micron mesh net
- IBI correlations before $R^2=0.88$ and after $R^2=0.85-0.88$
- Bray-Curtis distance measure of community similarity [0=identical communities, 1=no taxa shared]

  before and after:
  - within-SNARL original data = 0.32 (target similarity)
  - SNARL to R5.USFS-USU before re-sampling = 0.38
  - versus after re-sampling = 0.33 > nearly same as target

# Conclusions

- Different methods show similar performance characteristics and assessment scores
- Methods had high correlation, were independent of multimetric or multivariate analysis, and showed similar accuracy in discriminating reference from test
- Methods are easily calibrated and converted from previous data sets (SNARL to 500 fixed-count)
- Alternative sampling approaches may offer additional information, flexibility, or more resolution for the purposes of stressor identification

- Promising potential for data sharing, conversion, and adoption of a uniform standard approach: the targeted riffle composite method