New Developments involving the Stream Health Index for the Puget Sound Lowland

Grace Chiu Peter Guttorp

NRCSE

Technical Report Series

NRCSE-TRS No. 079

July 16, 2004

Full title:

New Developments involving the Stream Health Index for the Puget Sound Lowland

Short title:

SHIPSL — New Developments

Authors: Grace Chiu*, Peter Guttorp

Affiliation: Department of Statistics, University of Washington

**E-mail: grace@stat.washington.edu*

Tel: 206-616-9262

Fax: 206-685-7419

Address: Box 354322, Seattle, WA 98195, U.S.A.

This research was made possible through a Postdoctoral Fellowship to G. Chiu by the Natural Sciences and Engineering Research Council of Canada.

SUMMARY

Since the introduction of the *stream health index for the Puget Sound Lowland* (SHIPSL) (Chiu and Guttorp, 2004), new issues regarding this and other multimetric indices have arisen. Among them is the comparison between an unbounded continuous metric scale to one that is discrete and/or with bounds. One major issue stems from biases that result from the non-parametric bootstrap in the context of simulating new field samples of benthic organisms. We examine why it is doubtful that such biases are entirely absent in real-life repeated sampling. The discussion of these and other issues leads to two possible variants of SHIPSL to improve practicality and performance. The first involves universal "gold standards" in the metric scoring scheme to improve SHIPSL's tractability over time and space. The second redefines conventional count-valued taxa richness metrics as percentages. Both ideas are possibly applicable to other common multimetric indices, and are shown to perform comparably to what has been demonstrated by Chiu and Guttorp (2004) for SHIPSL and the benthic index of biotic integrity (B-IBI).

KEYWORDS: bioassessment, biomonitoring, index performance, metric scoring, multimetric index, taxa richness

1 INTRODUCTION

The *stream health index for the Puget Sound Lowland* (SHIPSL) (Chiu and Guttorp, 2004) is the sum of scores of the same 10 metrics that define the benthic index of biotic integrity (B-IBI) for the Puget Sound Lowland (PSL) (Karr, 1998). These scores are computed by standardization (across all sites in the study) of metric values from averaging over replicate field samples. The resulting SHIPSL index value has an exactly zero mean across sites, and no lower or upper bound for its range. A similar standardization technique is used to construct the *composite index of leading economic indicators* (Zarnowitz and Boschan, 1975) and similar NBER-BEA composite indices (Zarnowitz, 1992), in whose context time points instead of field sites are compared to each other. However, the SHIPSL scoring scheme was not inspired by any economic index.

Since the introduction of SHIPSL, comments on the index have been provided by general audiences of statisticians, and by ecologists involved in environmental management. We have revisited various issues concerning the performance of this and other multimetric indices of water quality based on those comments. Further developments of SHIPSL have since resulted, and are discussed in this article.

2 A NEED FOR STANDARDIZED SUBJECTIVITY

The attempt by Chiu and Guttorp (2004) to remove subjective input in the construction of a stream health index is largely due to the lack of universal protocols for administering subjectivity. Different definitions of, say, a set of field sites or metrics that effectively reflect the entire spectrum of biotic integrity may result in health indices that are highly incomparable. Furthermore, ideas of what values of certain variables indicate a "healthy" ecosystem could be influenced by local policy preferences (Lackey, 2003), thereby differing across geographical regions.

While protocols which are equally applicable nationwide (continent-wide) are still being developed, existing measures of biotic integrity should perhaps be modified to include protocol-free definitions. Chiu and Guttorp (2004) suggest one such scheme which assigns metric scores to a site relative to all other sites in the study, and not relative to a set of predetermined reference sites. While they do not address subjectivity involved in choosing the metrics, their results provide insight to the performance of a protocol-free metric scoring scheme compared to that of a scheme which (1) requires painstaking input in defining scoring criteria, and (2) is often applicable only to a certain geographical region.

However, SHIPSL index values, which reflect biological conditions relative to other sites being studied concurrently, become ambiguous when one wishes to monitor a single site over time. For this purpose, repeated use of a single historical dataset is suggested by Chiu and Guttorp (2004). A similar approach involves using "gold standard" values for the metric mean and standard deviation (SD) which appear in the scoring formula. Much like the speed of light relative to which the world's fastest traveling objects are gauged, gold standard mean and SD provide pivot points upon which metric values can be weighed. Such gold standards may be computed using, say, observations from a randomly chosen year made on "reference" sites randomly chosen from a census-enumeration-type database. These gold standards may then be reused over time or computed from resampled year-site combinations, neither of which involves subjective definition of reference. The idea of gold standards is not restricted to SHIPSL but possibly to other multimetric indices. Section 5 below examines the performance of SHIPSL defined using gold standards.

3 EQUAL OR UNEQUAL METRIC WEIGHTS?

Many different water quality indices are IBI's modified to suit local management criteria. Currently, we are not aware of any version in use that is an unequally weighted sum of metric scores. In the context of measuring water quality, ideal weights would reflect the relative importance of metrics with respect to their ability in reflecting underlying overall biological conditions.

One way to introduce unequal weights is by conducting a principal component analysis (PCA) on the metric values. However, as the PCA is designed to put more weight on metrics (subject to rotation and translation) with larger variability, the resulting importance ranking of metrics becomes distorted by the different metric scales. Therefore, a PCA on standardized metric scores (such as those of SHIPSL) would provide weights that are more biologically meaningful.

Figure 1 shows the loadings (weights) from the first PCA axis of SHIPSL metric scores for the 1997 and 1998 PSL data (taken from Chiu and Guttorp, 2004). By and large, the weights are close to being identical. In essence, SHIPSL can be regarded as the first PCA of standardized metric scores, which accounts for the most information that can be extracted from the ten metrics. (The respective first PCA axes explain 69% and 68% of the total variability among the scores.)

The issue of weighting was previously addressed by Auerbach (1982) in the context of a leading economic index. His findings indicate that equal weighting tends to smooth out fluctuations of the relationships between similarly standardized index components and underlying economic conditions. Perhaps the same justification applies to an equal weighting of biological metric scores used by such indices as SHIPSL and existing versions of the IBI.

4 OTHER METRIC SCORING METHODS

A thorough comparison of performance between six schemes of scoring metrics appears in Blocksum, 2003. There, all schemes are applied to the seven metrics of

the macroinvertebrate biotic integrity index (MBII). Three of these schemes lead to continuous scales for metric scores, while the remaining three lead to discrete scales. Both types of scales are calibrated using observations made either on predetermined reference sites or on all sites that have yielded field samples. All six schemes involve the definition of an upper and/or lower bound for the metric score, made based on predetermined non-trivial (i.e. neither 0th or 100th) percentiles of the distribution of raw metric values across calibration sites. For easy comparison, the index corresponding to each scheme is multiplied by 100 and divided by its largest possible value so that the resulting scaled indices have a range within 0 and 100. (However, this scaling may be somewhat unfair, as the minimum score value is 1 for two of the discrete scales, and 0 for the other four scales. Thus, scaled indices that correspond to the former two scales have minimum possible values of $7 \times 100 \div 35$ instead of 0.) Performance of the scaled indices is judged in a similar fashion to that of Fore *et al.*, 1994. In particular, bootstrap simulations are used to assess the indices' precision and related properties, such as power for detecting a pairwise difference.

Blocksum's results indicate that, with one exception, indices with a discrete scoring scheme have larger variability than those with a continuous scheme. The two schemes (continuous) which involve all sites in its calibration yield indices that are more precise than those of the remaining schemes which are calibrated using reference sites. One of these two schemes compresses only the upper tail of the metric value distribution (via a non-trivial percentile) for defining the metric score's range, and is observed to produce an index with least variability and most power, or, "an index that is closest to ideal." Blocksum argues that the high performance of this scoring scheme can be attributed to (1) better depiction of data by avoiding gaps in metric scores associated with discrete scoring schemes, and (2) less distortion of data due to compression of the tails of the metric value distribution. Altogether, Blocksum's results appear to provide justification for

the entirely bound-free continuous scaling of SHIPSL metrics, and the merit of basing on sites other than reference sites when calibrating the scoring scheme.

However, the issue of bias is not addressed by Blocksum, 2003. In particular, the conclusions for variability and power may not be entirely valid if the indices' SD's have different biases. In this case, a measure of "intrinsic variability" such as that of Chiu and Guttorp, 2004, is required for a fair comparison. Section 5 below revisits this comparison between SHIPSL and the PSL B-IBI. Included in the comparison is the version of SHIPSL based on gold standard values of metric means and SD's.

5 PERFORMANCE OF SHIPSL WHEN GOLD STANDARDS ARE USED

Let $\overline{y}_{ij}$ denote the value of metric j for site i that has been averaged over its replicate samples, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, 10$. Then, the statistically standardized metric score is

$$z_{ij} = \frac{\overline{y}_{ij} - \overline{\overline{y}}_j}{s_j} \tag{1}$$

where $\overline{\overline{y}}_j$ and s_j are the mean and sample SD, respectively, of $\overline{y}_{1j}, \overline{y}_{2j}, \ldots, \overline{y}_{nj}$.

When the aforementioned gold-standard approach is desired, one may replace $\overline{\overline{y}}_j$ and s_j of (1) by their respective predetermined gold standards, denoted by μ_j and σ_j. The resulting metric score is

$$z_{ij}^{(g)} = \frac{\overline{y}_{ij} - \mu_j}{\sigma_j}\,. \tag{2}$$

"Gold-standard" SHIPSL, or GS-SHIPSL, for site i is then $w_i^{(g)} = \sum_j z_{ij}^{(g)}$.

In practice, the major difference between (1) and (2) is that μ_j and σ_j of the latter may be reused in a new study (possibly involving different sites), whereas $\overline{\overline{y}}_j$ and s_j must be recomputed based on $\overline{y}_{ij}$'s which vary due to sampling variability or a different set of studied sites. Ideally, μ_j and σ_j would be national

or universal standards, so that GS-SHIPSL values from any geographical region would represent local stream health relative to a universal standard.

Here, as in Chiu and Guttorp, 2004, we study the performance of $w^{(g)}$ via its bootstrap distribution using the 1997 PSL data. As universal values are currently rare in practice, we treat $\{\overline{\overline{y}}_j, s_j\}$ of the original field sample from 1997 as $\{\mu_j, \sigma_j\}$ in the bootstrap version of (2). These are then reused throughout the 10,000 bootstrap samples. That is, for each of the 10,000 resamples,

$$z_{ij}^{(g)*} = \frac{\overline{y}_{ij}^{*} - \overline{\overline{y}}_j}{s_j} \tag{3}$$

is the (i, j)-th bootstrap GS-SHIPSL metric score, where "$*$" denotes a bootstrap value (as opposed to an observed value from the field, denoted without "$*$").

5.1 *Bias due to limitations of the bootstrap*

For their bootstrap study based on the same field data, Chiu and Guttorp (2004) obtain the following results. (1) For B-IBI, both sample mean and sample SD significantly underestimate the underlying center and variability across sites. (2) For SHIPSL, the sample mean is unbiased by definition, and the sample SD has no significant bias. (3) Correlation between site-wise bias and stream health is negative, and is significant if the B-IBI is used, but not so when SHIPSL is used.

How, then, does GS-SHIPSL perform with respect to these types of biases? The "BEFORE" column of Table 1 and the "before" curves of Figure 2 summarize the comparison among B-IBI and the two versions of SHIPSL with respect to sample mean and SD. Here, GS-SHIPSL is seen to behave more similarly to B-IBI than the original SHIPSL. In particular, each of its sample mean and SD has a significant negative bias. As a measure of variability, GS-SHIPSL's sample SD is half as precise as that of SHIPSL (standard error (SE) is twice as big).

Also note from the "BEFORE" column of Table 2 and the gray `o`'s of Figure 3 that negative correlation between health and site-wise bias is significant for

both B-IBI and GS-SHIPSL. As the spread of the bootstrap B-IBI or SHIPSL distribution differs from site to site, we also examine the *studentized* site-wise bias, i.e. site-wise bias divided by the SD of the bootstrap distribution. Here, the negative trend between bias and health remains. Fore *et al.* (1994) attributes this phenomenon for B-IBI to heavy compression of extreme metric values into those categories corresponding to the scores of 1 and 5. However, it does not explain the correlation for GS-SHIPSL. This is because, for either version of SHIPSL, all ten metrics are allowed to move freely in either direction of the unbounded, continuous scale for metric scores.

A closer look at the bootstrap resampling mechanism reveals that all seven taxa richness (count) metrics exhibit severe negative bias in its mean and SD (Table 3, "count" columns). In contrast, the corresponding biases are virtually non-existent for two of the three percentage metrics (one-sided p-values ≈ 0.5). To understand this phenomenon, compare the behavior of a taxa richness metric to that of a percentage metric. For example, either a single organism or 500 organisms belonging to a certain predatory taxon contributes one count to the metric *total number of taxa* (#Tx), while the corresponding contribution to the *percentage of predatory individuals* (%Pred) can drastically differ. On the other hand, adding a single organism to a zero count for the same taxon increases #Tx by one, but may have virtually no effect on %Pred.

Now, a zero count in the observed field sample for any taxa richness metric always produces zero bootstrap counts. Consequently, bootstrap values for such a metric can never exceed the observed field count, and their range is severely reduced. In the context of resampling from field samples, this limitation of the bootstrap significantly distorts the randomness that occurs across samples in practice. To better mimic the randomness in metrics resulting from sampling variability among actual field data, we correct for the bootstrap bias in mean and SD for each taxa richness metric. All three health indices are then recomputed for

the 1997 bootstrap data after this bias adjustment. Details of the bias correction procedure appear in Section 7.

The "AFTER" columns of Tables 1 and 2, together with the "after" curves of Figure 2 and the black +'s of Figure 3 show the behavior of the health indices after bias adjustment for the seven metrics. GS-SHIPSL now behaves much more closely to SHIPSL than B-IBI. Note also that although biases in B-IBI's sample mean and SD are no longer significant, the negative correlation between site-wise bias and index value remains significant. Furthermore, we see from Figure 4 that the *intrinsic variability* (c.f. Chiu and Guttorp, 2004) of B-IBI remains higher than either version of SHIPSL. It also has a larger SE, i.e. the sample SD is a less effective measure of variability for B-IBI than for GS-/SHIPSL.

Our results here indicate that, once artificial biases due to the bootstrap mechanism have been corrected for, the conclusions of Chiu and Guttorp, 2004 — with the exception of the biases in B-IBI's sample mean and SD — can be generalized to include either version of SHIPSL. Consequently, assuming that current protocols for collecting and identifying benthic organisms from the field produce metrics which effectively depict the underlying population conditions, both versions of SHIPSL have statistical properties that are (1) highly comparable, and (2) generally more desirable than those exhibited by the B-IBI. However, it is unclear how one could verify this assumption.

5.2 *Redefining SHIPSL metrics*

One way to possibly remove negative bootstrap bias (in mean and SD) for a taxa richness metric is to have it redefined as *percentage richness*. (The concept of percentage richness is similar to that of zooplankton proportions in Billheimer and Guttorp, 1995.) As no single taxa richness count can exceed the value of #Tx, all such metrics (except #Tx itself) can be converted to a percentage via division by #Tx/100. This conversion should retain most biological information contained

in the original integer-valued metric. Indeed, this is a way to standardize taxa richness metrics onto a common scale of 0 to 100. One advantage is that a percentage richness of, say, 15 may now have similar biological meanings in different geographical locations, whereas its integer-valued counterpart of, say, 3 taxa may not. In the context of the bootstrap, both numerator and denominator in the definition of percentage richness are similarly affected by negative bias; hence, the ratio between them should be relatively unaffected. Consequently, nine out of ten SHIPSL metrics should show little negative bias in its bootstrap mean and SD. In turn, the corresponding biases in SHIPSL should be reduced to a minimum.

Table 4 compares how informative are the count- and percentage-valued taxa richness metrics relative to each other. The two are highly correlated (all but one are greater than 0.8; see first column), suggesting that they contain comparable biological information about the streams. Moreover, for those cases in which this correlation is close to 0.9 or above (PleTx, LLTx, IntolTx, and ClingTx), there is little difference in correlation with urbanization between the two richness measures. For the other two cases (EphTx and TriTx), however, correlation with urbanization is greatly improved when percentage richness is used. This suggests that percentage richness may be generally more effective than count-valued taxa richness in reflecting human influence on the ecosystem being monitored. Such a characteristic is highly desired for metrics used to define health indices for biomonitoring purposes (see, for instance, Morley, 2000). Indeed, the correlation between SHIPSL and urbanization is higher when percentage richness metrics are used (-0.82 vs. -0.76, the latter of which also equals the correlation for B-IBI).

We now investigate performance issues of the two versions of SHIPSL using percentage richness metrics, based on the 10,000 bootstrap samples. Note that B-IBI is no longer considered in the comparison, as its metric scoring criteria are not available for percentage richness.

Recall the speculation that the new definition of richness should reduce nega-

tive bias among all taxa richness metrics except #Tx. The "percentage" columns of Table 3 show that this is generally true, except for %EphTx, %LLTx, and %IntolTx. The latter two have negative biases that are borderline significant at a 5% level. Interestingly, the former is the only percentage richness metric showing a positive bias, and it is highly significant. This phenomenon is explained in Section 5.3 below.

Does improved behavior of taxa richness metrics give rise to a more effective SHIPSL? The "o.v." columns of Table 1 show that variability in SHIPSL is reduced by using percentage richness. As in the case of count-valued taxa richness (before bias adjustment), GS-SHIPSL has bootstrap sample SD's that are, on average, comparable to those for ordinary SHIPSL (see "mean" columns). Judging by the left-most and right-most columns of p-values, bias in sample SD is virtually removed for GS-SHIPSL, although bias in sample mean remains significant. The bias in SD for ordinary SHIPSL is slightly worsened but remains insignificant. However, Table 2 ("percentage" columns) shows that site-wise bias is now significantly correlated with stream health (5% level) for either version of SHIPSL. We investigate this undesirable property in the next section.

5.3 *Taxon abundance and bootstrap bias*

While SHIPSL based on percentage richness is demonstrated to be more accurate and precise than that based on count-valued richness, a few concerns remain: (i) The severe negative bias in #Tx unquestionably contributes to the negative bias of SHIPSL even when all other taxa richness metrics have been converted to percentages; (ii) instead of removing bias for %EphTx, the conversion now adds positive bias to this metric; and (iii) the significant correlation between site-wise bias and stream health even when percentage richness is used remains unexplained.

Scatterplots of metric bias vs. observed SHIPSL using percentage richness appear in Figure 5. They can be regarded as a plot of site-wise bias vs. SHIPSL,

broken down by metric. The points labeled "a" and "b" correspond respectively to the sites MI1 (Miller Creek) and TH1 (Thornton Creek), and are influential values across all metrics. Upon their removal, we immediately see that (iii) above is a result of the high negative correlation between metric and SHIPSL for the metrics #Tx, %ClingTx, %LLTx, and %IntolTx.

Of the ten SHIPSL metrics, ClingTx, IntolTx, and LLTx have the lowest abundance. That is, each taxon classified as clinger, intolerant, or long-lived yields zero to very few individuals in an observed field sample regardless of the site's overall health conditions. Therefore, many such taxa are often entirely missing in a bootstrap sample, potentially causing a large negative bias in the count-valued richness for the bootstrap sample. In turn, #Tx (which comprises all these taxa) is also subject to a large negative bias. More specifically, let N be, say, #LLTx and D be #Tx, respectively, of the observed sample. Let $\varepsilon_n, \varepsilon_d \leq 0$ be bootstrap biases in N and D, respectively. Then, the bootstrap bias in %LLTx is

$$B = \frac{N - |\varepsilon_n|}{D - |\varepsilon_d|} - \frac{N}{D}\,.$$

Now, for a healthy site, N is not small, although it's contribution is from one or two individuals observed to belong to each long-lived taxon. The bootstrap then easily misses these taxa, so that $|\varepsilon_n| \approx N$. Note that D is large for a healthy site. Hence,

$$B \approx \frac{0}{D - |\varepsilon_d|} - \frac{N}{D} < 0\,. \tag{4}$$

However, low abundance can cause a positive bias also. For a degraded site, N is very small, and zero counts for many long-lived taxa imply $\varepsilon_n \approx 0$. Also, D is small, and $|\varepsilon_d| \geq |\varepsilon_n|$. Thus,

$$B \approx \frac{N}{D - |\varepsilon_d|} - \frac{N}{D} \geq 0\,. \tag{5}$$

Of course, (4) and (5) hold for any low abundance metric, and the value of B generally varies from positive to negative as the health conditions vary from bad

to good. This explains the severe negative correlation between metric bias and health for %ClingTx, %IntolTx, and %LLTx, and in turn, for #Tx.

A similar argument explains the positive bias in %EphTx. Of the three groups of flies, *Ephemeroptera* is up to twice as abundant as *Plecoptera* and *Trichoptera*. Thus, using the same notation as above, $\varepsilon_n \approx 0$ while $\varepsilon_d < 0$. Hence, (5) also applies to %EphTx. However, note that had #Tx not been negatively biased, %EphTx would not have been positively biased. In practice, high abundance of *Ephemeroptera* should not be a disadvantage.

For studying statistical properties of SHIPSL properly, however, we need to adjust for the distortion of #Tx, %EphTx, %ClingTx, %IntolTx, and %LLTx. Indeed, upon their removal in the computation of SHIPSL, the bias in sample mean and SD and the negative correlation between studentized site-wise bias and health virtually disappear (Table 5). For comparison, B-IBI scores are recomputed with the (count-valued versions of the) same five metrics removed. Table 5 shows that although biases in B-IBI mean and SD are no longer significant, negative correlation between bias and health remains. While removing these metrics for SHIPSL or B-IBI may be unwise in practice, this exercise demonstrates that:

(a) the negative biases in B-IBI sample mean and SD become insignificant after adjusting for metric biases (Section 5.1), or after removing very low / high abundance metrics;

(b) the negative correlation for B-IBI between site-wise bias and health remains despite bias adjustments for or removal of metrics;

(c) (a) and (b) suggest that bias-related issues for the B-IBI are largely due to the compression of tail values in its metric scoring scheme, particularly for high abundance metrics (#Tx and #EphTx);

(d) such issues for either version of SHIPSL are mainly due to the bootstrap

in the presence of very low abundance metrics, and not to any inherent limitations of the scoring mechanism;

(e) SHIPSL based on percentage richness have properties that are more desirable than when count-valued taxa richness is used.

6 DISCUSSION

Prompted by reactions to the introduction of SHIPSL by Chiu and Guttorp (2004), this article has tried to address recently arisen issues which concern the use of SHIPSL and other health indices for bioassessment of fresh water systems. We have pointed out the need for a set of universal protocols under which the process of quantifying health conditions — from the notion of *pristine* and *degraded* waters, to the choice of index metrics and their scoring, to the collection of data, to the interpretation of index values — should ideally be agreeable upon regardless of local geography and policies. To this end, GS-SHIPSL and the notion of percentage richness have been introduced in this article. Of course, the choice of a health index is merely one element of an intricate biomonitoring scheme. However, the comparison among SHIPSL, GS-SHIPSL, and B-IBI brings to light how this choice may help to achieve part of the grand goal of developing a sound system for ecological assessment.

In particular, both versions of SHIPSL involve unbounded metric scoring scales that prevent negative correlation between bias in index and stream health. In practice, GS-SHIPSL is more tractable over time / space, and likely more biologically meaningful than ordinary SHIPSL. This is because GS-SHIPSL utilizes an invariable "gold standard" that can be made universal for gauging metrics. Provided that field sampling protocols yield samples that effectively reflect underlying biological conditions, GS-SHIPSL is statistically comparable to SHIPSL. Furthermore, the use of *percentage* instead of *count-valued* taxa richness metrics is

shown to increase the precision of both versions of SHIPSL, and to better reflect human impact. Thus, percentage richness is highly recommended for B-IBI and similar health indices which comprise many taxa richness metrics.

However, two issues remain unresolved. Firstly, we have observed how bootstrap results can be distorted by the presence of very low abundance metrics. In practice, repeated sampling of benthic organisms may be similarly affected. For instance, consider a taxon unobserved in a field sample replicate. Although future replicates are not restricted to yield zero counts as would bootstrap resamples, how may one determine whether organisms of this taxon actually exist at this sampled site? If they do not, there is no variability or bias whatsoever for the taxon frequency. However, if they do but are low in abundance, it may take many replicates before a non-zero frequency is observed, and bias among a small to moderate number of replicates is almost undoubtedly negative. Therefore, for some sites, conditions gauged by such metrics are possibly irreproducible in a handful of field samples. It may appear that the practicality of metrics involving low abundance taxa as indicators of biological conditions is questionable.

Secondly, current biomonitoring practices remain highly geographically dependent. Any effort for enhancing the universality of a health index would be made in vain unless a common "language" for describing and quantifying health is available to different geographical regions.

7 TECHNICAL DETAILS

First, some notation is needed for discussing the bias correction procedure of Section 5.1. For metric j that measures taxa richness, let

$$\begin{aligned} y_{ijk} &= \text{metric value observed for } i\text{-th site's } k\text{-th field replicate} \\ \overline{y}_{jk} &= \left\{ \begin{array}{c} \text{metric's observed sample mean} \\ \text{over sites for replicate } k \end{array} \right\} = \frac{1}{n} \sum_{i=1}^{n} y_{ijk} \end{aligned}$$

$$
\begin{aligned}
s_{jk} &= \left\{ \begin{array}{c} \text{metric's observed sample SD} \\ \text{over sites for replicate } k \end{array} \right\} = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ijk} - \overline{y}_{jk})^2 \\
y^*_{ijk,r} &= r\text{-th boostrap metric value for } i\text{-th site's replicate } k \\
\overline{y}^*_{jk,r} &= \left\{ \begin{array}{c} \text{metric's sample mean over sites} \\ \text{for } r\text{-th bootstrap replicate } k \end{array} \right\} = \frac{1}{n} \sum_{i=1}^{n} y^*_{ijk,r} \\
s^*_{jk,r} &= \left\{ \begin{array}{c} \text{metric's sample SD over sites} \\ \text{for } r\text{-th bootstrap replicate } k \end{array} \right\} = \frac{1}{n-1} \sum_{i=1}^{n} (y^*_{ijk,r} - \overline{y}^*_{jk,r})^2
\end{aligned}
$$

Now, bias for this metric (j) is corrected for each replicate, as follows. (See **Remarks I** below for notes on bias correction when metric values are averaged over replicates.) Let

$$
\begin{aligned}
\overline{\overline{y}}^*_{jk} &= \left\{ \begin{array}{c} \text{bootstrap estimate of metric mean} \\ \text{for } k\text{-th replicate} \end{array} \right\} = \frac{1}{10\,000} \sum_{r=1}^{10\,000} \overline{y}^*_{jk,r} \\
u^*_{jk} &= \left\{ \begin{array}{c} \text{estimate of bias in metric mean} \\ \text{for } k\text{-th replicate} \end{array} \right\} = \overline{\overline{y}}^*_{jk} - \overline{y}_{jk} \\
\overline{s}^*_{jk} &= \left\{ \begin{array}{c} \text{estimate of expected value of metric SD} \\ \text{for } k\text{-th replicate} \end{array} \right\} = \frac{1}{10\,000} \sum_{r=1}^{10\,000} s^*_{jk,r} \\
v^*_{jk} &= \left\{ \begin{array}{c} \text{correction factor for bias in metric SD} \\ \text{for } k\text{-th replicate} \end{array} \right\} = \frac{s_{jk}}{\overline{s}^*_{jk}} \\
t^*_{ijk,r} &= \left\{ \begin{array}{c} \text{bias-corrected metric value} \\ \text{for } i\text{-th site's replicate } k \\ \text{in } r\text{-th bootstrap sample} \end{array} \right\} = \begin{array}{c} v^*_{jk}(y^*_{ijk,r} - \overline{y}^*_{jk,r}) \\ + \\ (\overline{y}^*_{jk,r} - u_{jk}) \end{array}
\end{aligned} \tag{6}
$$

Subsequently, bootstrap values for all three stream health indices are recomputed using $t^*_{ijk,r}$'s. In particular, suppress the subscript r and see that $\overline{t}^*_{ij}$ replaces $\overline{y}^*_{ij}$ in (3) for each bootstrap replicate.

The non-parametric bootstrap with bias adjustment as described above becomes semi-parametric in that the first two moments of the bootstrap distribution are coerced to coincide with those of the observed replicate.

Remarks I

One may wish to reduce computation by using a u_j^* and a v_j^* which are not replicate-specific. Possible definitions would involve estimating bias in the mean and SD (over sites) of the metric values already averaged over replicates. Unless with care, however, such correction factors could lead to circular definitions of $z_{ij,r}^*$ and $z_{ij,r}^{(g)*}$, thereby defeating the purpose of separately defining a gold-standard SHIPSL. On the other hand, circularity is avoided by using (6), as averaging the term $v_{jk}^*(y_{ijk,r}^* - \overline{y}_{jk,r}^*)$ over k in the definition of $z_{ij,r}^{(g)*}$ unlikely leads to canceling of many terms.

Remarks II

Here, calculations of metric values, bias estimation, and bias adjustment altogether become very computationally intensive for the non-parametrically bootstrapped field samples. Indeed, it is possible to employ a different type of bootstrap than that used by Fore *et al.* (1994), Blocksum (2003), and Chiu and Guttorp (2004). For example, values of metric j can be semi-parametrically bootstrapped from, say, a Poisson($\widehat{\lambda}_j$) distribution (for taxa richness; see Billheimer and Guttorp, 1995), or a log-normal($\widehat{\mu}_j, \widehat{\sigma}_j^2$) distribution (for a percentage metric), where $\widehat{\lambda}_j$ and $\{\widehat{\mu}_j, \widehat{\sigma}_j^2\}$ are estimates based on the original field samples. Note that metrics are correlated. Hence, resampling metric values must involve a model for the dependence structure among metrics. It is, however, beyond the scope of this article to develop a suitable dependence model for such a purpose.

ACKNOWLEDGMENTS

We thank the staff at the U.S. Environmental Protection Agency (EPA) Western Ecology Division (WED) for their valuable suggestions.

REFERENCES

Auerbach AJ. 1982. The index of leading indicators: "measurement without theory," thirty-five years later. *The Review of Economics and Statistics* **64**: 589–595.

Billheimer D, Guttorp G. 1995. Zooplankton proportion estimates from non-unifrom sample volumes. *Envrionmental and Ecological Statistics* **2**: 117–124.

Blocksum KA. 2003. A performance comparison of metric scoring methods for a multimetric index for mid-Atlantic highlands streams. *Environmental Management* **31**: 670–682.

Chiu G, Guttorp P. 2004. *Stream Health Index for the Puget Sound Lowland.* NRCSE Technical Report Series No. 078.

Fore LS, Karr JR, Conquest LL. 1994. Statistical properties of an index of biological integrity used to evaluate water resources. *Canadian Journal of Fisheries and Aquatic Sciences* **51**: 1077–1087.

Karr JR. 1998. Rivers as sentinels: Using the biology of rivers to guide landscape management. In *River Ecology and Management: Lessons from the Pacific Coastal Ecosystems,* Naiman RJ, Bilby RE (eds). Springer, New York; 502–528.

Lackey, RT. 2003. Appropriate use of ecosystem health and normative science in ecological policy. In *Managing for Healthy Ecosystems,* Rapport DJ, Lasley WL, Rolston DE, Nielsen NO, Qualset CO, Damania AB (eds). Lewis, Boca Raton; 175–186.

Morley, SA. 2000. *Effects of Urbanization on the Biological Integrity of Puget Sound Lowland Streams: Restoration with a Biological Focus.* M.S. thesis: University of Washington.

Zarnowitz V. 1992. *Business Cycles: Theory, History, Indicators, and Forecasting.* The University of Chicago Press: Chicago.

Zarnowitz V, Boschan C. 1975. Cyclical indicators: an evaluation and new leading indexes. In *Business Conditions Digest* (May 1975). U.S. Department of Commerce.

	count richness									**percentage richness**				
	o.v.	*BEFORE bias adjustment*				*AFTER bias adjustment*				o.v.				
		bootstrap			p	bootstrap			p		bootstrap			p
		mean	bias	SE		mean	bias	SE			mean	bias	SE	
SHIPSL														
sample mean	0	0	0	0	NA	0	0	0	NA	0	0	0	0	NA
sample SD	8.17	8.12	−0.05	0.12	0.33	identical to 3 decimal places			0.33	7.16	6.94	−0.22	0.22	0.16
GS-SHIPSL														
sample mean	0	−1.68	−1.68	0.18	0.00	−0.05	−0.05	0.18	0.40	0	−0.39	−0.39	0.22	0.04
sample SD	8.17	7.69	−0.48	0.24	0.02	8.13	−0.04	0.26	0.43	7.16	7.06	−0.10	0.44	0.41
B-IBI														
sample mean	27.67	26.95	−0.72	0.26	0.00	27.60	−0.07	0.31	0.41	NA				
sample SD	9.06	8.46	−0.60	0.35	0.04	8.65	−0.41	0.35	0.12					

Table 1: **Bias in sample mean and sample SD.** "o.v." denotes observed value from field samples. p-values are one-sided, computed based on normal approximations to the bootstrap distributions.

index	count richness				percentage richnesss	
	BEFORE bias adjustment		AFTER bias adjustment			
	corr(bias, health)	1-sided *p*-value	*corr(bias, health)*	1-sided *p*-value	*corr(bias, health)*	1-sided *p*-value
SHIPSL	−0.227	0.18	−0.228	0.18	−0.639	0.00
	[−0.214]	[0.20]	[−0.214]	[0.20]	[−0.546]	[0.01]
GS-SHIPSL	−0.763	0.00	−0.209	0.20	−0.412	0.04
	[−0.805]	[0.00]	[−0.188]	[0.23]	[−0.617]	[0.00]
B-IBI	−0.632	0.00	−0.493	0.02	——	——
	[−0.603]	[0.00]	[−0.493]	[0.02]	——	——

Table 2: **Correlation between bias and health.** Entries under "*corr(bias, health)*" are correlation coefficients between site-wise bootstrap bias and observed index value. (See Figure 3.) Entries in square brackets ([ ]) are computed for studentized site-wise biases.

taxa richness metric		count richness			percentage richness		
		observed	*bootstrap bias*	*p*	*observed*	*bootstrap bias*	*p*
total # taxa (#Tx)	mean	23.26	**−2.23**	0.00	----------	----------	----------
	SD	6.65	**−0.32**	0.01			
Ephemeroptera taxa (EphTx)	mean	4.74	**−0.28**	0.00	19.93	**1.09**	0.00
	SD	1.76	**−0.12**	0.00	4.31	−0.06	0.42
Plecoptera taxa (PleTx)	mean	3.89	**−0.37**	0.00	15.58	−0.06	0.42
	SD	1.66	**−0.12**	0.01	5.78	0.20	0.27
Trichoptera taxa (TriTx)	mean	4.00	**−0.39**	0.00	16.88	−0.18	0.34
	SD	1.37	−0.06	0.19	2.83	0.51	0.19
long-lived taxa (LLTx)	mean	3.02	**−0.37**	0.00	11.94	−0.41	0.08
	SD	1.47	**−0.15**	0.00	5.38	−0.14	0.26
intolerant taxa (IntolTx)	mean	0.20	**−0.05**	0.02	0.71	−0.14	0.07
	SD	0.49	**−0.10**	0.04	1.61	−0.22	0.14
clinger taxa (ClingTx)	mean	12.26	**−1.19**	0.00	50.35	−0.31	0.27
	SD	4.46	**−0.32**	0.00	11.55	0.49	0.32

non-taxa richness metric		*observed*	*bootstrap bias*	*p*
% predatory individuals (%Pred)	mean	5.34	0.00	0.50
	SD	3.76	0.00	0.33
% tolerant individuals (%Tol)	mean	74.49	0.00	0.50
	SD	14.24	0.00	0.46
% individuals in 3 most dominant taxa (%Dom3)	mean	38.21	−0.01	0.06
	SD	11.36	0.00	0.20

Table 3: **Bias in mean and SD of SHIPSL metrics.** One-sided p-values are based on (1) normal approximations to the bootstrap distributions of metric mean and SD, and (2) unrounded observed and bootstrap values. Numbers in bold are significant biases at a 5% level.

taxa richness	*correlation between*		
metric	# and %	# and urbanization	% and urbanization
EphTx	0.83	−0.21	−0.47
PleTx	0.93	−0.75	−0.73
TriTx	0.73	−0.58	−0.74
LLTx	0.93	−0.86	−0.83
IntolTx	0.99	−0.40	−0.41
ClingTx	0.88	−0.76	−0.73

Table 4: **Correlations involving percentage richness.** First "correlation" column is that between the count- (#) and percentage-values (%) of the taxa richness metric indicated under "richness." Second and third columns are correlation between taxa richness (# or %, as indicated by the column heading) and the *percentage of urbanized area* from Morley, 2000.

	bootstrap		1-sided	*corr(bias, health)*	1-sided
	bias	SE	*p*-value		*p*-value
SHIPSL (% richness)				−0.38	0.06
sample mean	0	0	NA		
sample SD	0.02	0.19	0.46		
GS-SHIPSL (% richness)				0.11	0.33
sample mean	−0.12	0.19	0.26		
sample SD	0.23	0.35	0.25		
B-IBI (# richness)				−0.62	0.00
sample mean	−0.18	0.19	0.18		
sample SD	−0.30	0.24	0.10		

Table 5: **Biases after *Tx, ClingTx, IntolTx* and *LLTx* are removed.** Entries under "*corr(bias, health)*" are based on studentized site-wise biases.

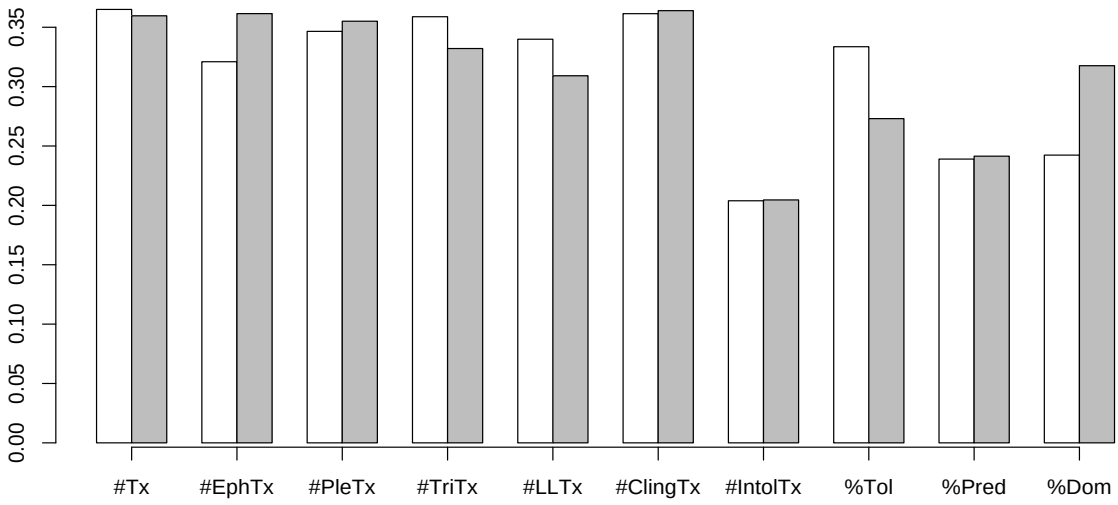

Figure 1: **First PCA loadings of standardized metrics.** 1997 loadings are in white, and 1998 loadings are in gray. See Table 3 for full metric names.

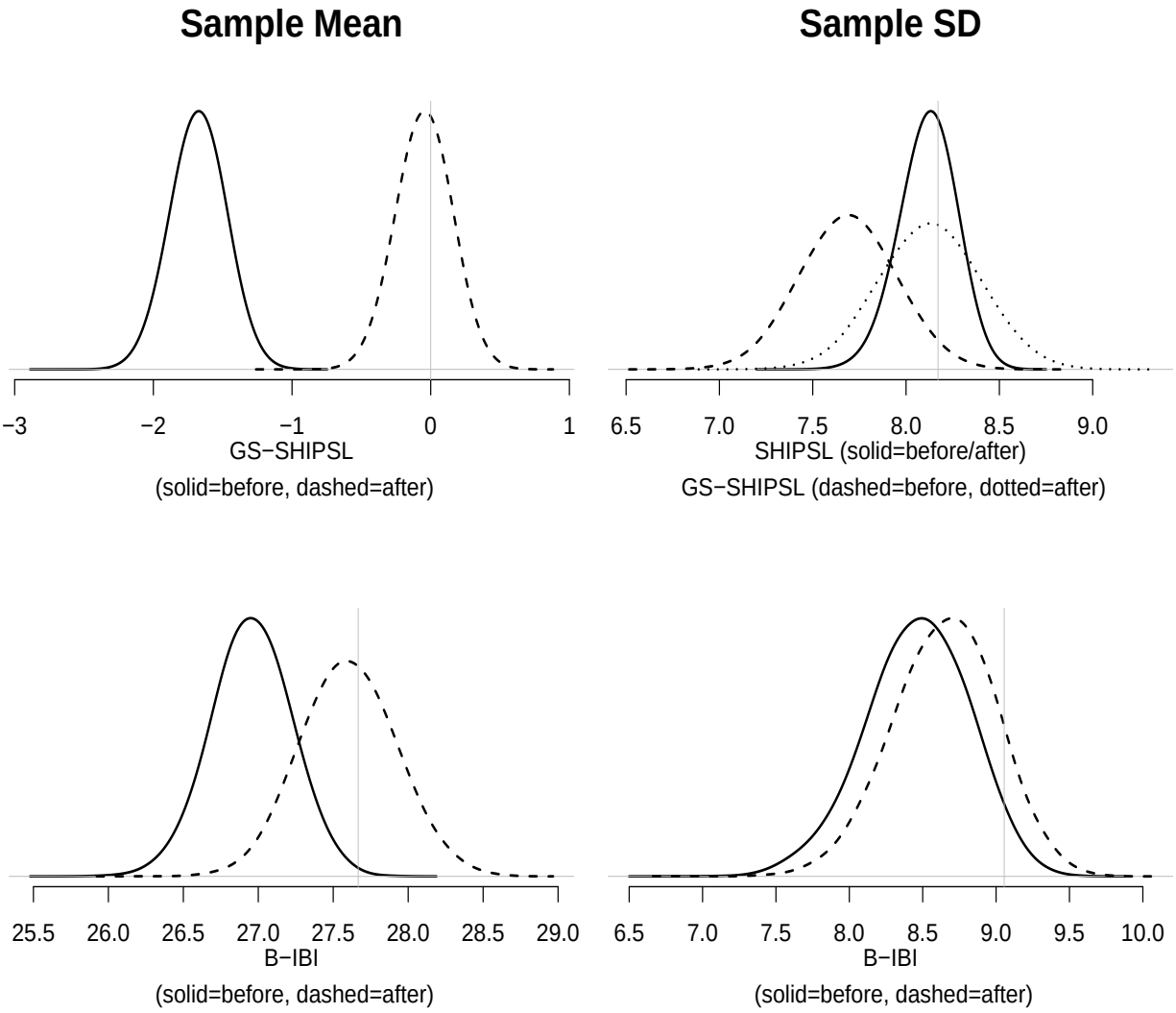

Figure 2: **Effect of bias correction for count-valued taxa richness metrics on bootstrap sample mean and SD.** (Also see Table 1.) Displayed are bootstrap distributions of sampled mean (left) and sampled SD (right) before and after bias correction. Vertical lines in gray denote values observed from 1997 field samples. Note that bias correction has virtually no effect on the sample SD for ordinary SHIPSL.

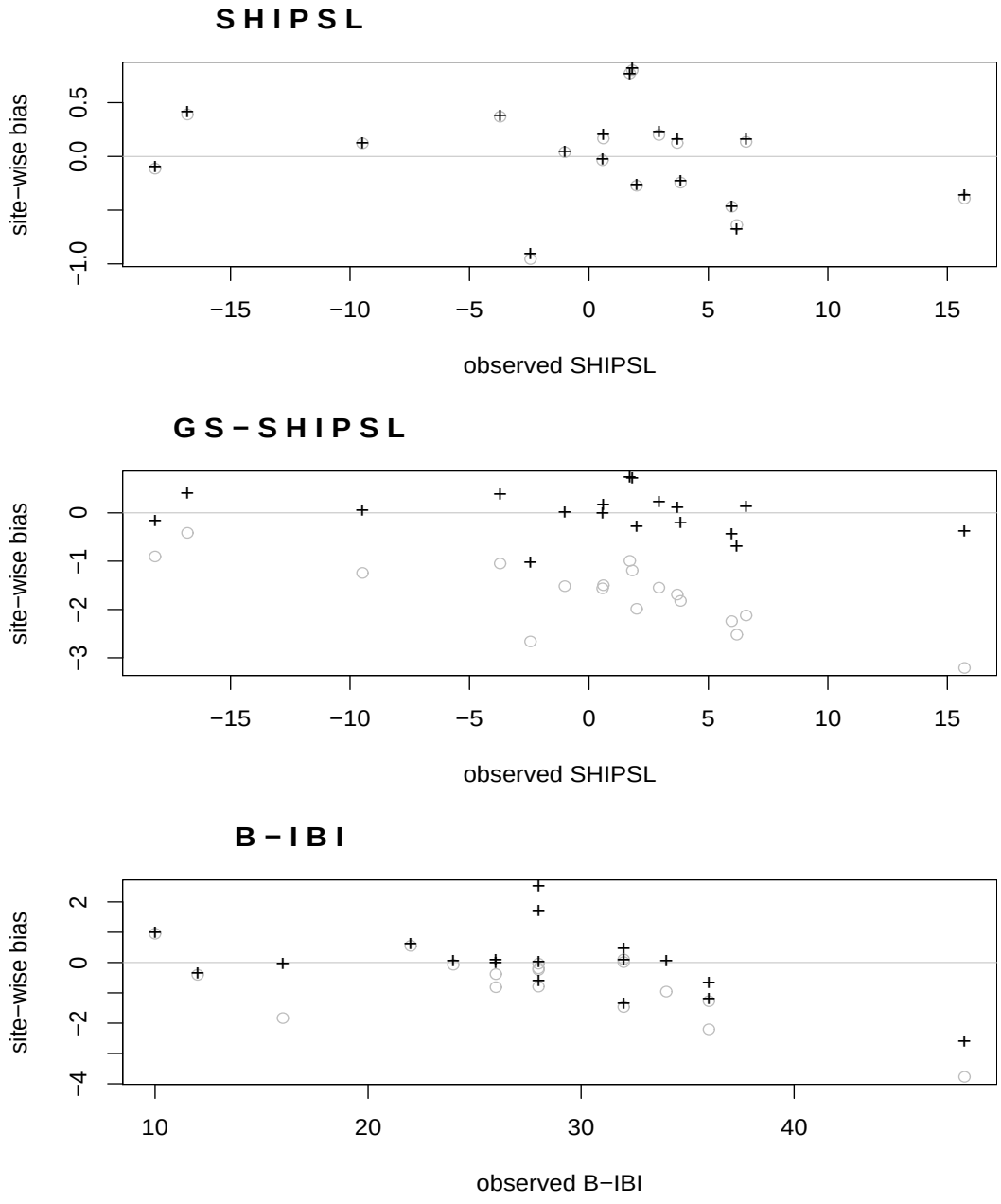

Figure 3: **Effect of bias correction for count-valued taxa richness metrics on site-wise bias.** (Also see Table 2.) Displayed are scatterplots of site-wise bias vs. observed index value for the 1997 data, before ("o") and after ("+") bias correction.

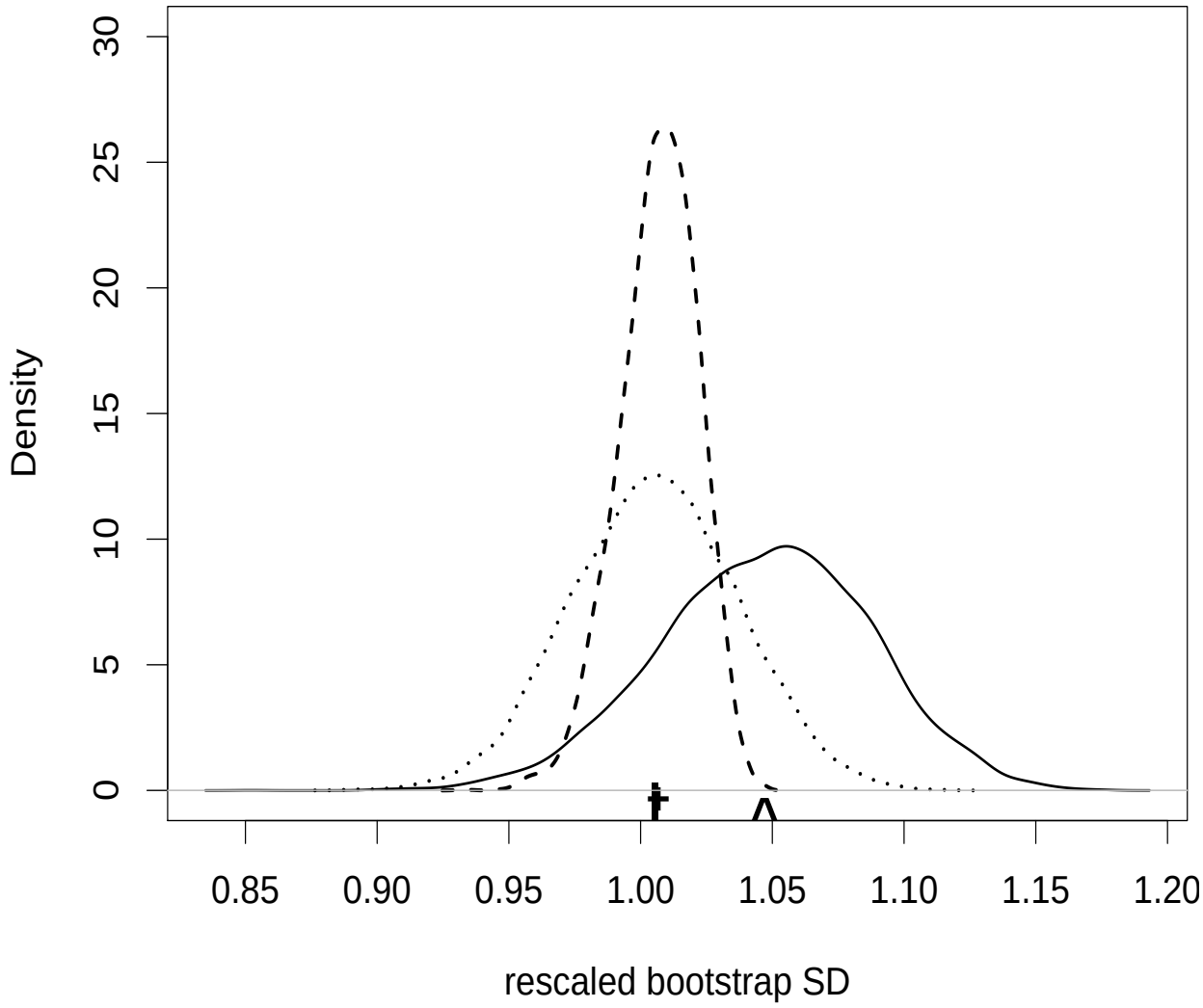

Figure 4: **Intrinsic variability.** Displayed are bootstrap distributions for sample SD of B-IBI's (—), SHIPSL's (- - -), and GS-SHIPSL's (. . .), after (1) biases in count-valued taxa richness metrics have been corrected for, and (2) resulting index values have been rescaled to allow sensible comparison between the indices' precision. (See Chiu and Guttorp, 2004 for (2).) The means of the respective distributions are marked by " ^ ", "+", and "|".

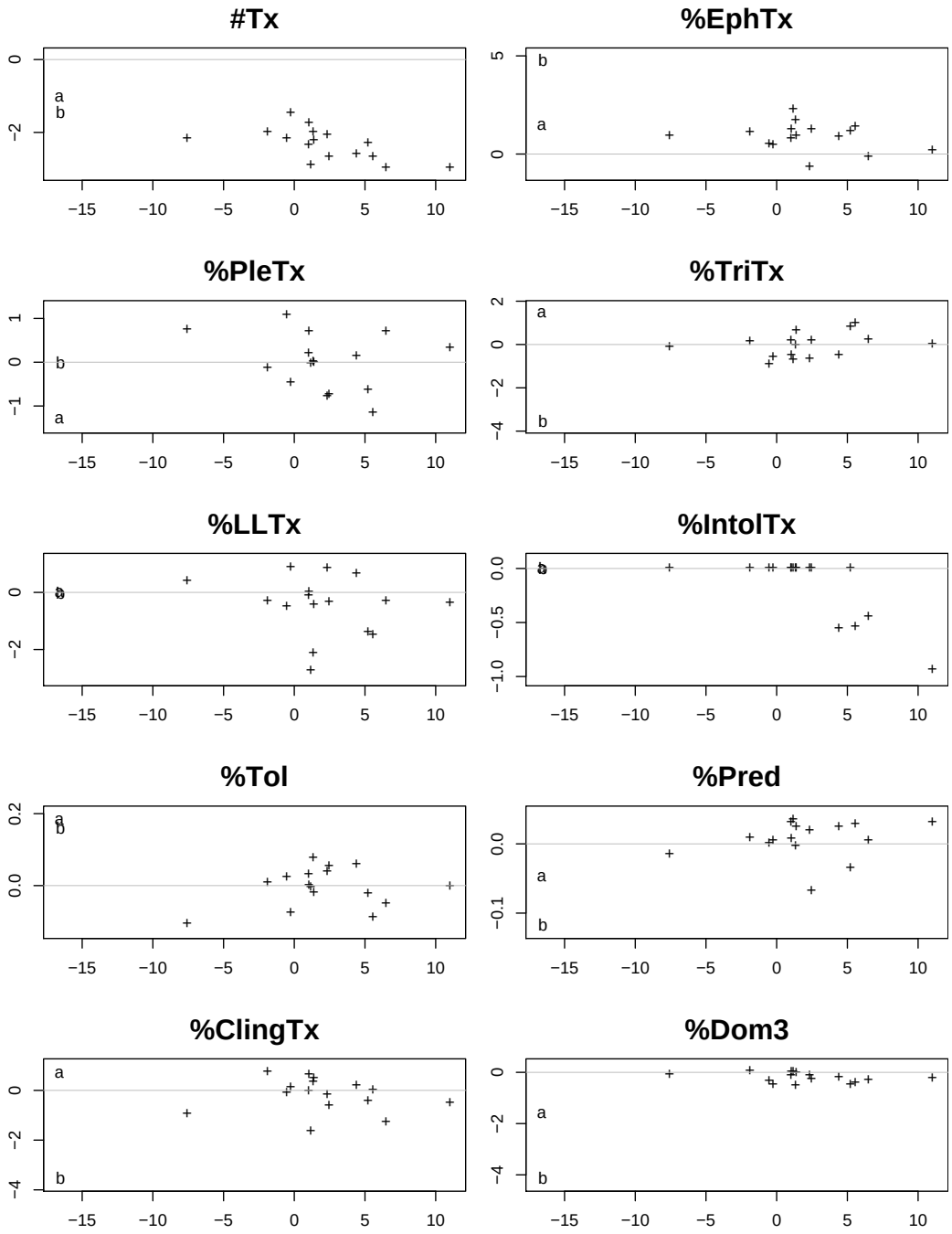

Figure 5: **Metric bias and stream health.** Displayed are scatterplots of (site-wise) metric bias vs. observed SHIPSL based on percentage richness. (See Table 3 for full metric names.) Points labeled "a" and "b" correspond respectively to Miller Creek and Thornton Creek.