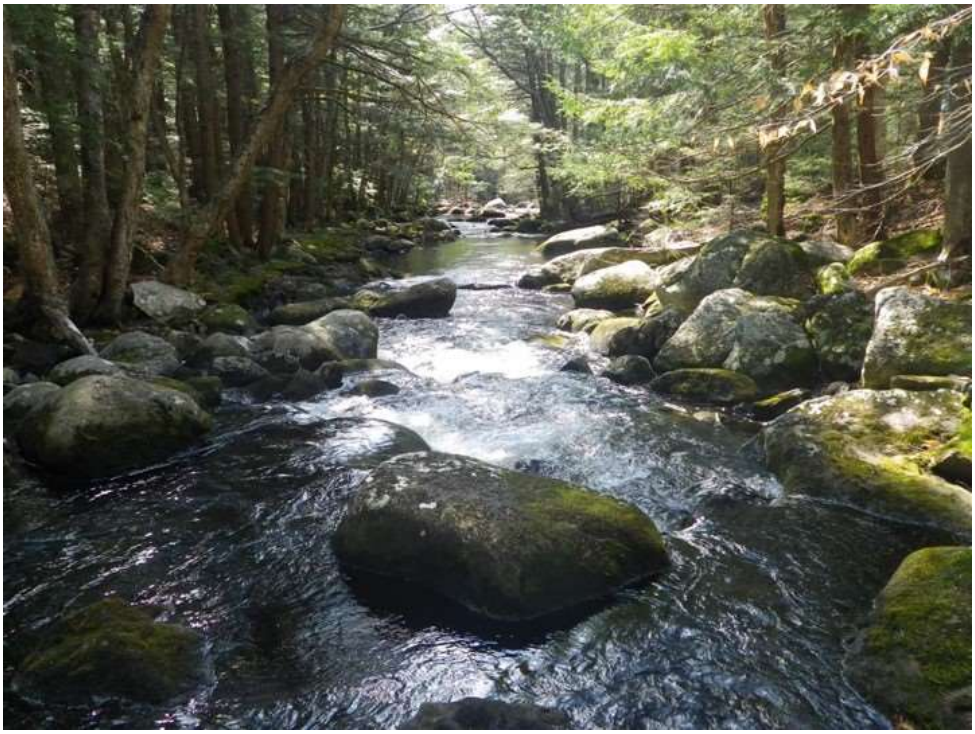
Calibration of Indices of Biotic Integrity for 300-organism macroinvertebrate riffle habitat samples in Massachusetts freshwater wadeable streams



Prepared for:
Massachusetts Department of Environmental Protection

Prepared by:
Benjamin Block
Jen Stamp
Benjamin Jessup
Tetra Tech, Inc.
73 Main Street, Room 38, Montpelier, VT 05602

December 29, 2020

Executive Summary

In an earlier phase of work, Indices of Biotic Integrity (IBIs) were calibrated for macroinvertebrate samples collected from riffle habitat in freshwater wadeable streams in all but the southeastern portion of Massachusetts (Narragansett/Bristol Lowlands, Cape Cod, and the Islands) (Jessup and Stamp 2020). The IBIs were calibrated for two regions: Western Highlands (WH) and Central Hills (CH). The IBIs were calibrated to 100-organism samples because those comprised most of the available samples at the time. However, the eventual goal was to develop 300-count versions of the IBIs.

During this exercise, we used 300-count riffle habitat samples collected by the Massachusetts Department of Environmental Protection (MassDEP) from 2012-2019 to calibrate 300-count IBIs for the WH and CH regions. We utilized the same input metrics that are in the 100-count IBIs. Steps included: 1) examining differences in distributions of metric values between 100- vs 300-count samples; 2) calculating three IBI alternatives with different metric scoring formulae and evaluating their performance; 3) evaluating differences between 100- vs. 300-count IBI scores across the three alternatives; and 4) exploring how subsample size affects numeric threshold designations and Type I and Type II error rates.

After reviewing the results for the three IBI alternatives, the MassDEP workgroup selected the option that adjusted the metric scoring formulae for richness metrics only, since these were most affected by differences in subsample size (100- vs. 300-count). For the other metrics, the workgroup decided to retain the scoring formulas that are used in the 100-count riffle habitat IBIs, since the 100-count dataset had a larger sample size and captured a wider disturbance gradient than the 300-count dataset.

The 300-count CH riffle habitat IBI effectively discriminates between reference and stressed samples (DE=100 and z-score = 2.6, compared to DE=100 and z-score = 3.0 in the 100-count CH IBI; Jessup and Stamp 2020). The 300-count WH IBI performed poorly (DE=20 and z-score = 0.34, compared to DE=88 and z-score = 1.4 in the 100-count WH IBI; Jessup and Stamp 2020). However, due to limitations with the 300-count WH stressed dataset (which only had five samples from three sites, with limited levels of disturbance), these results should be interpreted with caution. As an alternate measure of performance, we examined associations between the 300-count IBIs and disturbance variables. Compared to results from similar analyses performed on the 100-count riffle habitat IBIs (Jessup and Stamp 2020), associations tended to be slightly weaker but were mostly significant ($p<0.05$) and in keeping with the expected direction of response. The weaker associations were likely driven in part by the limitations of the 300-count dataset, which had fewer sites and a more limited disturbance gradient (in particular in the WH region, where urban land cover was poorly represented).

MassDEP will re-evaluate IBI performance after more 300-count samples are obtained and analyzed. In the meantime, MassDEP is planning to calculate both 100- and 300-count IBI scores for its samples[1] and compare results. In addition, MassDEP will continue to conduct targeted

[1]MassDEP obtains 100-count samples by randomly subsampling the 300-count samples with the "subsample" computer program.

sampling to broaden the disturbance gradients represented in each region, with a particular focus on sampling more high-stress sites in the WH region.

MassDEP will also consider potential thresholds for numeric bio-criteria as they evaluate IBI results in coming years. MassDEP does not currently have plans to pursue numeric bio-criteria in the Massachusetts surface water quality standards (SWQS) but has identified preliminary thresholds for the 100-count IBIs for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded), as described in Stamp and Jessup (2020), for use in the Consolidated Assessment and Listing Methodology (CALM) to interpret the narrative biological criteria in the SWQS. During this exercise, we took a preliminary look at whether it would be feasible to apply the same thresholds for the 300-count IBIs. Results suggest the Satisfactory/Moderately Degraded thresholds were similar (between 55-60), although reference percentiles for the 300-count IBIs tended to be slightly higher. Type I and Type II error rates were also similar between the 100-count and 300-count IBIs, with the exception of the Type II error rates in the WH dataset (which should be interpreted with caution due to limitations of the stressed samples).

Table ES-1. Metrics and scoring formulas that comprise version 1 of the Central Hills and Western Highlands 300-count riffle habitat IBIs. The metric scoring formulas highlighted in light blue differ from those used in the 100-count IBIs. These formulas were changed to account for effects of subsample size on the richness metrics.

Central Hills 300-count riffle habitat IBI		
Metric	**Response to stress**	**Scoring formula**
Total number of taxa	Decrease	100*(metric)/55.8
% EPT taxa	Decrease	100*(metric)/54.5
% Ephemeroptera individuals, excluding Caenidae and Baetidae	Decrease	100*(metric)/13.9
% Collector-filterer individuals	Increase	100*(79.9-metric)/66.9
% Predator taxa	Decrease	100*(metric)/28.5
% Intolerant taxa	Decrease	100*(metric)/39.1
Western Highlands 300-count riffle habitat IBI		
Metric	**Response to stress**	**Scoring formula**
Total number of taxa	Decrease	100*(metric)/61.8
% Plecoptera individuals	Decrease	100*(metric)/18.3
% Collector-filterer individuals	Increase	100*(50.5-metric)/40.7
% Shredder individuals	Decrease	100*(metric)/23
% Intolerant individuals	Decrease	100*(metric)/51.5
Becks Biotic Index[1]	Decrease	100*(metric)/50.6

[1]Beck's Biotic Index (Terrell and Perfetti 1996) = 2*[Class 1 Taxa]+[Class 2 Taxa] where Class 1 taxa have tolerance values of 0 or 1 and Class 2 taxa have tolerance values of 2, 3 or 4.

Acknowledgments

The index calibration process was supported by the Massachusetts Department of Environmental Protection (MassDEP) through a contract with Tetra Tech, Inc. The index development team consisted of James Meek, Allyson Yarra, Robert Nuzzo, Arthur Johnson, Robert Maietta, Joan Beskenis, and Anna Mayor, who participated in bi-weekly calls and provided feedback throughout the process. In addition, James Meek reviewed site disturbance assignments. We are very grateful for their contributions.

Table of Contents

Executive Summary i
Acknowledgments iii
1 Background 1
2 Dataset 1
3 Methods 6
3.1 Metric calculations and comparisons 6
3.2 IBI alternatives 9
3.3 Effects of subsample size on IBI scores and thresholds 10
4 Results 11
4.1 Metric comparisons 11
4.1.1 Central Hills 11
4.1.2 Western Highlands 16
4.2 IBI alternatives 20
4.3 Effects of subsample size on IBI scores and thresholds 26
5 Conclusions 32
6 Literature Cited 35

Appendices

A Disturbance Index
B Metric performance plots – 100- vs. 300-count samples
C Metric scatterplots – 100- vs. 300-count samples
D IBI scatterplots – 100- vs. 300-count samples (3 index alternatives)

Attachments

A Comparison of metric and IBI scores – Central Hills
B Comparison of metric and IBI scores – Western Highlands

List of Tables

Table 1. Number of samples and sites in each of the seven disturbance categories. 4
Table 2. Input metrics in the Central Hills and Western Highlands riffle habitat IBIs. 6
Table 3. Descriptions of the three statistics that were used to evaluate the performance of the IBI alternatives.. 7
Table 4. Distribution statistics for the Central Hills riffle habitat IBI metrics, grouped by subsample size (100- vs. 300-count). 13
Table 5. Performance statistics for Central Hills riffle habitat IBI metrics in the 300-count dataset and 100-count IBI calibration dataset (Jessup and Stamp 2020). 15
Table 6. Spearman rank correlation coefficients (r_s) for Central Hills riffle habitat IBI metrics, in 100- vs 300-count paired samples 15
Table 7. Distribution statistics for the Western Highlands riffle habitat IBI metrics, grouped by subsample size (100- vs. 300-count). 17
Table 8. Performance statistics for Western Highlands riffle habitat IBI metrics in the 300-count dataset and 100-count IBI calibration dataset (Jessup and Stamp 2020). 19
Table 9. Spearman rank correlation coefficients (r_s) for Western Highlands riffle habitat IBI metrics, in 100- vs 300-count paired samples. 19
Table 10. Descriptions of the three 300-count IBI alternatives that were considered. Table 11 contains the metric scoring formulas. 20
Table 11. Metric scoring formulas based on the 300-count dataset vs the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020). 21
Table 12. Performance statistics for the 300-count riffle habitat IBI alternatives, compared with performance statistics from the 100-count IBI calibration dataset (Jessup and Stamp 2020). 22
Table 13. Spearman rank correlation coefficients (r_s) for the 300-count IBI alternatives, compared with results from the 100-count riffle habitat IBI calibration dataset (taken from Jessup and Stamp 2020).. 25
Table 14. Precision statistics for IBI scores based on 100- vs 300-count samples, grouped by scoring scheme. 27
Table 15. Comparison of IBI scores for multiple percentiles of reference. 28
Table 16. Metrics and scoring formulas in version 1 of the Central Hills and Western Highlands 300-count riffle habitat IBIs 34

List of Figures

Figure 1. Locations of the 99 sites with 300-count riffle habitat samples. Sites are color-coded by broad disturbance category. 3
Figure 2. Range of disturbance represented in the reference, stressed, and intermediate samples in the 300-count dataset and 100-count IBI calibration dataset. 5
Figure 3. Discrimination efficiency (DE).. 8
Figure 4. Distribution of Central Hills riffle habitat IBI metric values, grouped by subsample size (100- vs. 300-count).. 14
Figure 5. Distribution of Western Highlands riffle habitat IBI metric values, grouped by subsample size (100- vs. 300-count). 18
Figure 6. Distribution of IBI scores in reference vs. stressed samples for the IBI alternatives, based on the Central Hills dataset. 23
Figure 7. Distribution of IBI scores in reference vs. stressed samples for the IBI alternatives, based on the Western Highlands dataset.. 24
Figure 8. Central Hills. Differences between IBI scores in paired 100- vs. 300-count samples for each index alternative. 26
Figure 9. Western Highlands. Differences between IBI scores in paired 100- vs. 300-count samples for each index alternative. 27
Figure 10. Central Hills. Distribution of AdjRichOnly IBI scores across the three broad disturbance categories (reference (Ref), stressed (Strs), and intermediate (Inter). 30
Figure 11. Western Highlands. Distribution of AdjRichOnly IBI scores across the three broad disturbance categories (reference (Ref), stressed (Strs), and intermediate (Inter).. 31

1 Background

The Massachusetts Department of Environmental Protection (MassDEP) began its macroinvertebrate sampling program in the 1980s. It used a target of 100-organisms for its routine sampling until 2012-2013, when it transitioned to a 300-organism target. In 2018, we started working with MassDEP on calibrating Indices of Biotic Integrity (IBIs) for macroinvertebrate samples collected from riffle habitat in freshwater wadeable streams (Jessup and Stamp 2020). At the time the riffle habitat IBIs were calibrated, most of the available samples were 100-count. Thus, the IBIs were calibrated to 100-organism samples. The goal of this exercise was to calibrate 300-count versions of the riffle habitat IBIs using the 300-count data collected from 2012-2019. The same input metrics that are in the 100-count IBIs were used in the 300-count IBIs. Steps included: 1) examining differences in distributions of metric values between 100- vs 300-count samples; 2) calculating three IBI alternatives with different metric scoring formulae and evaluating their performance; 3) evaluating differences between 100- vs. 300-count IBI scores across the three alternatives; and 4) exploring how subsample size affects numeric threshold designations and Type I and Type II error rates. Having IBIs calibrated for both subsample sizes (100- and 300-count) gives MassDEP more flexibility in ongoing assessments as they evaluate performance of the riffle habitat IBIs.

2 Dataset

The 300-count macroinvertebrate dataset consisted of riffle habitat samples collected by the MassDEP from freshwater, perennial, wadeable streams using the Rapid Bioassessment Protocol (RBP) kick net method. Samples were collected from riffle/run areas in streams with fast currents and rocky substrate. Field crews kicked or disturbed bottom sediments and caught the dislodged organisms in a net as the current carried them downstream (Barbour et al. 1999). At each site, ten kicks were taken over a 100-m reach and then composited into a single sample. Samples were collected between July 1–September 30 (± 1 week), using a kick-net with 500-µm mesh and a 46-cm wide opening. Organisms were subsampled and identified to the lowest practical level in the laboratory, by Cole Ecological, Inc.

The 100-count riffle habitat IBIs were calibrated for two regions: Western Highlands (WH) and Central Hills (CH) (Figure 1). In the 300-count riffle habitat dataset, 62 sites were in the CH and 28 were in the WH. In addition, nine sites were located in the southeastern portion of Massachusetts (Narragansett/Bristol Lowlands (NBL), Cape Cod (CC), and the Islands), but those sites were excluded from this exercise because there were insufficient samples to develop a riffle habitat IBI for this area. In both the CH and WH regions, 10 sites had multiple samples, either due to multiple years of sampling (ranging from 2 to 5 years) or random subsampling in the laboratory (referred to as lab splits). The lab splits were done in sets of three. We considered lab splits to be independent samples and included all of them in our analyses. In total, the 300-count dataset included 104 samples from the CH (including 10 sets of lab splits) and 65 samples from the WH (including 5 sets of lab splits). Each 300-count sample had an associated 100-count version that had been randomly subsampled by MassDEP with the "subsample" computer

program. Attachment A contains the list of samples that were included in the IBI calibration exercise.

During IBI calibration, it is important to capture as wide a disturbance gradient as possible. Reference sites are used to identify metric expectations with the least levels of disturbance. When a set of stressed sites are identified using criteria at the opposite end of the disturbance scale, the response of metrics along the resulting stressor gradient can be detected. The direction and strength of the response is used for selecting candidate metrics for inclusion in an IBI and properly scoring them. In this dataset, sites were assigned to disturbance categories using the process described in Appendix A. Seven disturbance variables were considered: Index of Catchment Integrity (ICI), Index of Watershed Integrity (IWI) (Thornbrugh et al. 2018, Johnson et al. 2019), percent urban land cover, density of roads, dam storage volume, percent agricultural land cover, and modeled mean rate of fertilizer application + biological nitrogen fixation + manure application. Sites were initially assigned to seven disturbance categories, ranging from Best Reference to High Stress, before being collapsed into three broader disturbance categories (reference, stressed, intermediate) for the analyses. Appendix A contains a more detailed description of the criteria and procedures that were used to assign sites to disturbance categories, and Attachments A & B include disturbance category assignments for CH and WH samples used in the analyses.

CH sites generally have higher levels of disturbance than WH sites. This difference is particularly evident when comparing percent urban land cover (Figure 2). Due to the differences in disturbance levels across the two regions and the need to obtain adequate numbers of reference and stressed sites for IBI calibration, we used slightly different thresholds to define reference and stressed in the CH and WH regions. Stressed sites in the CH were derived from the High Stress category, while in the WH, the High Stress and Stress categories were combined (Table 1). The CH reference sites were comprised of sites in the Best Reference, Reference, and Sub Reference categories; in the WH, reference sites were from the Best Reference and Reference categories (Table 1). The same disturbance category groupings were used during the calibration of the 100-count riffle habitat IBIs.

Table 1 shows the number of sites and samples in each disturbance category in the 300-count dataset, compared to the 100-count IBI calibration dataset (Jessup and Stamp 2020). As expected, the 300-count dataset was more limited, particularly in regard to the number of stressed sites. There are only three stressed sites in the 300-count dataset for the WH (vs. 70 in the 100-count IBI calibration dataset), and 12 stressed sites in the CH (vs. 63 in the 100-count IBI calibration dataset). The number of reference sites is more comparable, with 46 reference samples from 22 sites in the WH 300-count dataset (vs. 41 sites in the WH 100-count IBI calibration dataset) and 58 reference samples from 29 sites in the CH 300-count dataset (vs. 45 sites in the CH 100-count IBI calibration dataset) (Table 1).

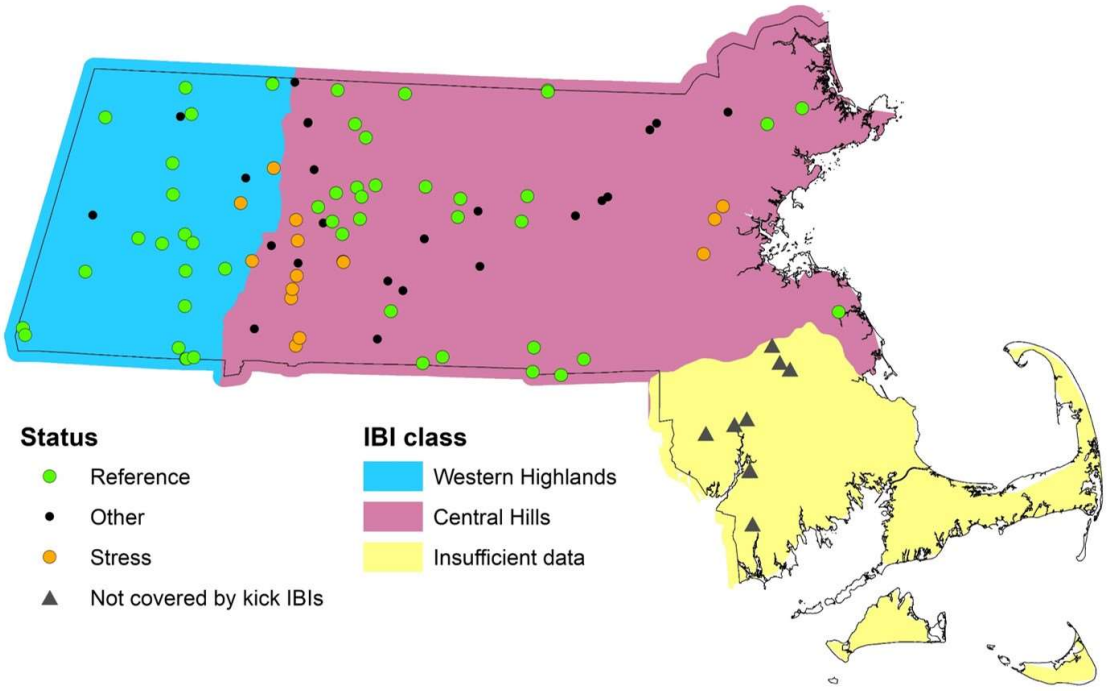


Figure 1. Locations of the 99 sites with 300-count riffle habitat samples. Sites are color-coded by broad disturbance category. There were insufficient data to develop a riffle habitat IBI for the Southeast region; therefore, sites in that region (coded as gray triangles) were excluded from the calibration exercise.

Table 1. Number of samples and sites in each of the seven disturbance categories (Best Reference to High Stress, as described Appendix A) in the 300-count dataset and 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020). Sites were later collapsed into three broader categories (reference, stressed, intermediate) for analyses. Due to the differences in disturbance levels across regions and the need to obtain adequate numbers of reference and stressed sites for IBI calibration, we used slightly different thresholds to define reference and stressed in the CH and WH regions. Stressed sites (highlighted in orange) in the CH were derived from the High Stress category, while in the WH, the High Stress and Stress categories were combined. Reference sites (highlighted in green) in the CH were comprised of sites in the Best Reference, Reference, and Sub Reference categories; in the WH, reference sites were from the Best Reference and Reference categories.

Disturbance category	Western Highlands			Central Hills		
	300-count dataset		100-count riffle IBI	300-count dataset		100-count riffle IBI
	# sites	# samples	# sites/samples*	# sites	# samples	# sites/samples*
Best Reference	7	16	7	6	9	4
Reference	15	30	34	13	28	13
Sub Reference	1	5	25	10	21	28
Intermediate	1	4	15	4	7	26
Some Stress	1	5	48	8	14	89
Stress	2	4	58	9	11	135
High Stress	1	1	12	12	14	63
Reference	22	46	41	29	58	45
Intermediate	3	14	88	21	32	250
Stress	3	5	70	12	14	63
Total	***28***	***65***	***199***	***62***	***104***	***358***

*the dataset for the 100-count riffle habitat IBI calibration was limited to one sample per site

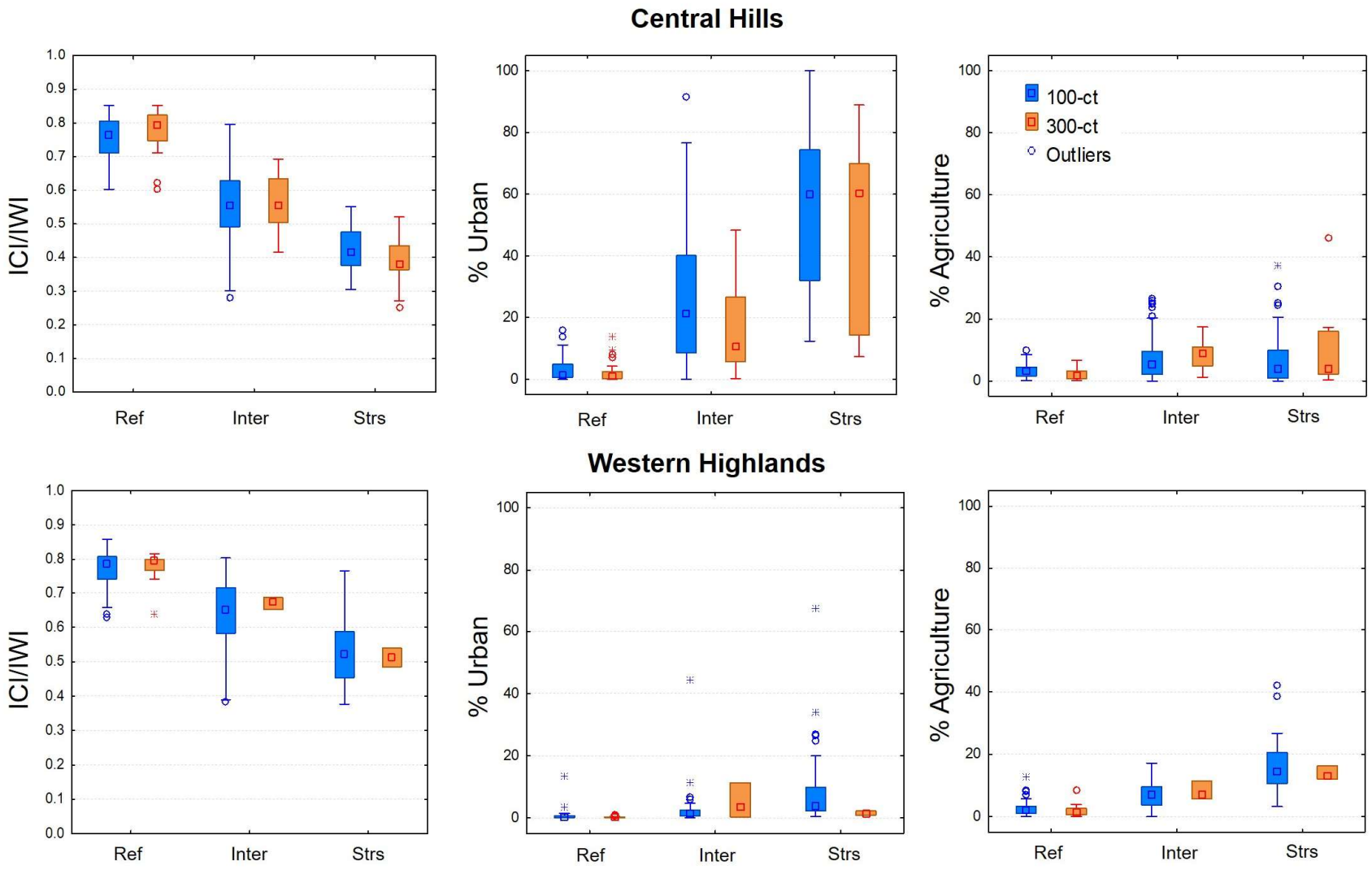


Figure 2. Range of disturbance represented in the reference, stressed, and intermediate samples in the 300-count dataset and 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020), as measured by the indices of catchment and watershed integrity (ICI and IWI, respectively) (Thornbrugh et al. 2018, Johnson et al. 2019), % urban and % agricultural land cover. For more information on the disturbance variables, see Appendix A.

3 Methods

3.1 Metric calculations and comparisons

For this exercise, we utilized the same input metrics as the 100-count riffle habitat IBIs (Table 2). When preparing the data, we followed the same procedures that were used during the calibration of the 100-count riffle habitat IBIs (described in Jessup and Stamp 2020). Metrics were calculated with the BioMonTools R package (https://github.com/leppott/BioMonTools; Leppo et al. 2020). Two of the 300-count samples had > 360 total individuals, which exceeded the ± 20% target that had been used for the 100-count IBI calibration. We used the BioMonTools R package to randomly subsample those to 300 total individuals. Metrics were calculated for the 300 and 100-count versions of the samples. We calculated the standard suite of summary statistics (minimum, mean, standard deviation, median, maximum) as well as the 5th and 95th percentiles, which were used in the IBI metric scoring formulas. We evaluated differences between 100- vs. 300-count metric values through comparisons of summary statistics, box plots, and Spearman rank correlation analyses.

We assessed metric performance by calculating three performance statistics: Discrimination Efficiency (DE) (Flotemersch et al. 2006, Maxted et al. 2000, Ofenböck et al. 2004), z-score (similar to Cohen's D; Cohen 1992), and coefficient of variation (CV). These statistics are described in Table 3. Our ability to use the DE and z-score statistics was hindered by the limited number of stressed sites in the 300-count dataset (in particular for the WH, which only had five stressed samples from three sites, and those sites did not have particularly high levels of disturbance) (Table 1).

Table 2. Input metrics in the Central Hills and Western Highlands riffle habitat IBIs.

Metric (abbrev)	Response to stress
Central Hills riffle habitat IBI	
Total number of taxa (nt_total)	Decrease
% EPT taxa (pt_EPT)	Decrease
% Ephemeroptera individuals, excluding Caenidae and Baetidae (pi_Ephem NoCaeBae)	Decrease
% Collector-filterer individuals (pi_ffg_filt)	Increase
% Predator taxa (pt_ffg_pred)	Decrease
% Intolerant taxa (pt_tv_intol)	Decrease
Western Highlands riffle habitat IBI	
Total number of taxa (nt_total)	Decrease
% Plecoptera individuals (pi_Pleco)	Decrease
% Collector-filterer individuals (pi_ffg_filt)	Increase
% Shredder individuals (pi_ffg_shred)	Decrease
% Intolerant individuals (pi_tv_intol)	Decrease
Becks Biotic Index (x_Becks)	Decrease

Table 3. Descriptions of the three statistics that were used to evaluate the performance of the IBI alternatives. The DE and z-score calculations are based on both reference and stressed samples. The CV only considers reference samples.

Statistic	Interpretation	Calculation
Discrimination Efficiency (DE)	Higher score = better. The higher the value, the better the metric or IBI is at correctly discriminating between stressed and reference samples.	Decreaser metrics: percentage of stressed values below the 25th percentile of reference site values. Increaser metrics: percentage of stressed sites that have values higher than the 75th percentile of reference values. DE can be visualized on box plots of reference and stressed values with the inter-quartile range plotted as the box (as shown in Figure 3).
z-score	Higher absolute score = better. Higher values indicate better separation between reference and stressed samples.	(Mean value of reference samples – mean value of stressed samples)/standard deviation of reference values.
Coefficient of variation (CV)	Lower score = better. Lower values mean less dispersion around the mean.	Standard deviation of reference values/mean value of reference samples.

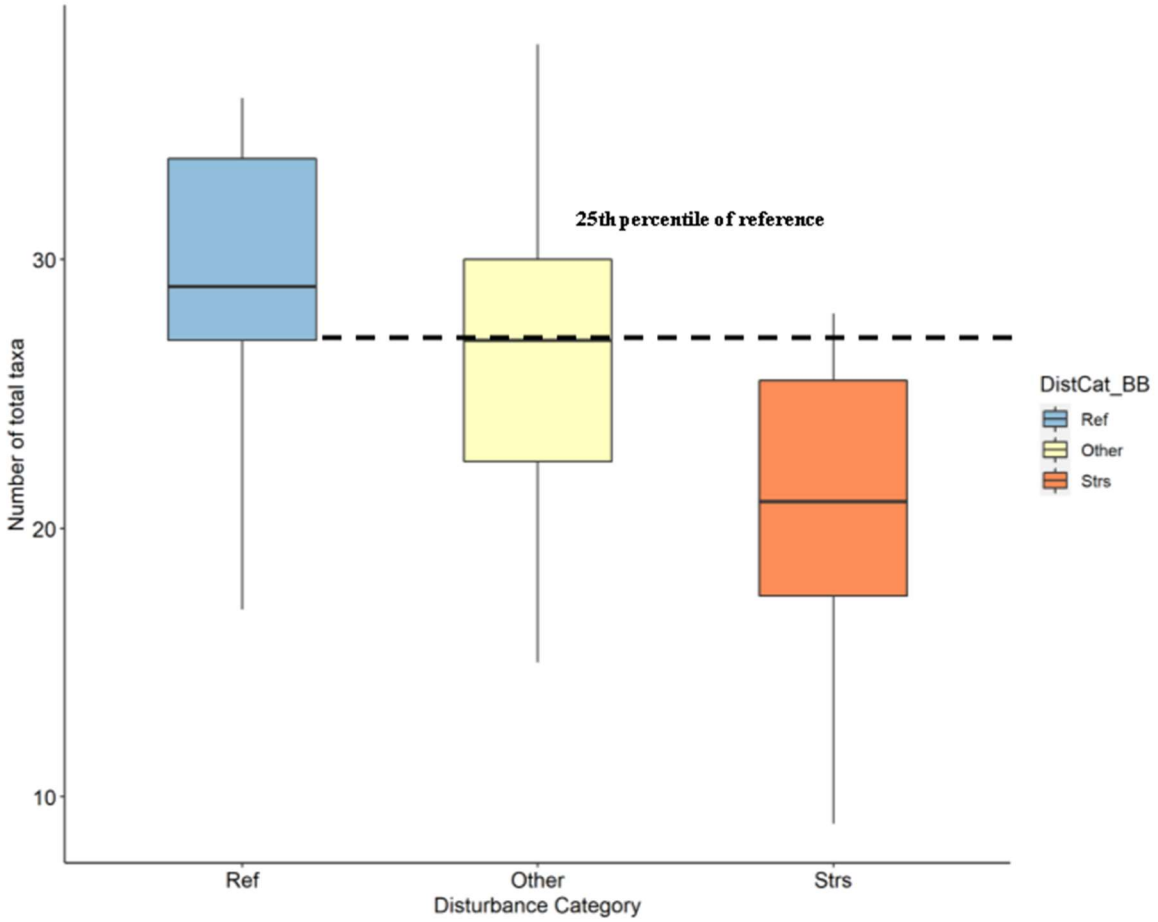


*Figure 3. Discrimination efficiency (DE). In this example, which uses the total number of taxa (a metric that decreases with stress), the 25th percentile of the reference distribution is used as the standard (and we calculate what % of stressed sites were below that threshold; for example, if 15 out of 20 stressed sites have # total taxa metric values below the threshold (in this case, 27), the DE would equal 75%; if metric values for all 20 of the stressed sites were < 27, the DE would equal 100%). If it were a metric that increased with stress, we would have used the 75th percentile of the reference distribution as the standard (and calculated what % of stressed sites were above that threshold). The formula is: DE = a/b*100, where a = number of a priori stressed sites identified as being below the degradation threshold (in this example, 25th percentile of the reference site distribution) and b = total number of stressed sites. The higher the DE, the better (the more frequent the correct association of metric values with site conditions).*

3.2 IBI alternatives

Before IBIs are calculated, metric scores are calculated because metrics are mostly on different scales and thus cannot be directly aggregated. To address this, formulas were applied to the metrics to convert them to a 0-100-point scoring scale (as in Hughes et al. 1998, and Barbour et al. 1999). The metric scoring formulas vary by the distribution of values in the IBI calibration dataset. For this exercise, the scoring scale was based on the distribution of metric values across all sites (versus reference sites only).

For metrics that decreased with increasing stress (referred to as 'decreasers'; an example is the number of intolerant taxa metric), we used the following equation, in which the 95th percentile was the upper end of the scoring scale and the minimum possible value (0) was the lower end:

$$Decreaser\ metric\ score = 100 * \frac{Metric\ value - minimum\ possible\ value}{95th\ percentile - minimum\ possible\ value}$$

Scores for metrics that increased with stress (referred to as 'increasers'; an example is the number of tolerant taxa metric) were calculated with this equation:

$$Increaser\ metric\ score = 100 * \frac{95th\ percentile - metric\ value}{95th\ percentile - 5th\ percentile}$$

Each metric was scored on a 0 to 100 scale. If metric scores came out to be < 0 or >100, they were re-set to the 0-100 scale. IBI scores were derived by averaging the metric scores.

We calculated three alternative versions of 300-count riffle habitat IBIs based on the following metric scoring schemes:

- **Adjust none** – metric scoring formulas from the 100-count riffle habitat IBI (Jessup and Stamp 2020) were applied to all metrics.
- **Adjust subset** – metric scoring formulas from the 100-count riffle habitat IBI (Jessup and Stamp 2020) were used, except for metrics that had mean values that differed by more than one standard deviation in the 100- vs. 300-count datasets.
- **Adjust all** – Metric scoring formulas based on 5th and 95th percentiles (and minimum possible scores) in the 300-count dataset were applied to all metrics.

IBI performance was evaluated with DE, z-scores, and CV (Table 3). Performance statistics were compared with those derived during the calibration of the 100-count riffle habitat IBIs (Jessup and Stamp 2020). Results for the 300-count dataset were interpreted with caution due to the limited number of sites (in particular, stressed sites). We also evaluated IBI performance by exploring the response to disturbance. We ran Spearman rank correlations to examine the relationship between the IBIs and four disturbance variables (ICI, IWI, percent urban and percent agricultural land cover) and compared results to those derived during the calibration of the 100-count riffle habitat IBIs (Jessup and Stamp 2020).

3.3 Effects of subsample size on IBI scores and thresholds

We used several approaches to evaluate the effects of subsample size on IBI scores. One involved calculating differences between IBI scores in paired 100- vs. 300-count samples for each index alternative, generating box plots and evaluating: 1) whether the 300-count samples were receiving higher IBI scores; and 2) which IBI alternative had the smallest and largest differences between 100- vs. 300-count samples. We also used scatterplots to evaluate differences in IBI scores in 100- vs. 300-count samples.

The divergence of IBI scores in 100- vs. 300-count samples was further analyzed by computing the following precision statistics:

- Mean-squared-error (MSE), which measures distances from data points to the regression line (these distances are the "errors") and squares them. MSE tells you how close a regression line is to a set of points. Higher MSE = greater divergence (less desirable).
- Root-mean-squared-error (RMSE), which is the standard deviation of the residuals. RMSE tells you how concentrated the data is around the regression line. Higher RMSE = greater divergence (less desirable).
- Coefficient of variation (CV), which is the ratio of the standard deviation to the mean. Higher CV = greater variability.
- 90% confidence interval (CI 90), which is a measure of uncertainty. Higher CI90 = greater degree of uncertainty.

We also explored differences in reference percentiles derived from the 300-count IBI alternatives vs. those calculated based on the 100-count IBI calibration dataset (Jessup and Stamp 2020). MassDEP has developed preliminary thresholds for the 100-count riffle habitat IBIs for four conditions: exceptional, satisfactory, moderately degraded, and severely degraded (described in Stamp and Jessup 2020). The threshold that distinguishes satisfactory from moderately degraded condition equals 55, which corresponds with the 10th percentile of reference in the CH 100-count dataset and the 15th percentile of reference in the WH 100-count dataset. For this exercise, we calculated the 10th, 15th, 20th, 25th, and 50th reference percentiles based on the 300-count IBI dataset and compared them to the equivalent statistics derived from the 100-count IBI dataset.

As an additional step, we assessed Type I and Type II error rates. Type I error is known as a "false positive" finding (in this case, falsely calling a site disturbed when it is not). Type II error captures "false negative" findings (or falsely calling a disturbed site undisturbed). When you decrease the probability of one error, it increases the probability of the other. A consequence of having a high Type I error rate is a higher likelihood of mistakenly subjecting undisturbed sites to potentially costly management actions, whereas having a high Type II error rate increases the likelihood of not detecting degradation. Most biomonitoring programs try to simultaneously minimize Type I and Type II errors (Breine et al. 2007), but approaches vary across entities and depend on acceptable error rates.

We assessed Type I and Type II error rates based on two thresholds:

- 55 (the 100-count riffle habitat IBI threshold distinguishing satisfactory from moderately degraded condition).

- The 10th percentile of reference in the CH 300-count dataset and the 15th percentile of reference in the WH 300-count dataset (which correspond with the reference percentiles that were used to derive the threshold of 55 for the 100-count riffle habitat IBIs)

We measured Type I error as the percentage of reference sites that fell below each threshold and Type II error as the percentage of stressed sites that had scores greater than or equal to the thresholds. We then compared the results to the error rates from the 100-count IBIs (Stamp and Jessup 2020).

4 Results

4.1 Metric comparisons

4.1.1 Central Hills

Of the six CH input metrics, the total number of taxa metric (the only richness metric in the IBI) showed the greatest difference when mean metric values from the 300-count samples were compared to the 100-count samples. On average, there were 15 more taxa in the 300-count samples (Table 4). This divergence was also evident in the box plots in Figure 4. The total taxa metric was the only metric that had mean metric values that differed by more than one standard deviation in the 100- vs. 300-count datasets. The metric with the next greatest difference was the percent predator taxa, which, on average, was ~ 3% higher in the 300-count dataset. Mean metric values for the rest of the metrics were very similar between the 100- and 300-count results (within 2%) (Table 4).

Regarding metric performance, DEs for the 300-count samples varied depending on the metric and, to a lesser degree, subsample size. DEs ranged from 57 (percent predator taxa metric) to 100 (percent intolerant taxa metric) (Table 5). When compared with DEs calculated from the 100-count versions of the samples, four of the metrics had the same DEs. Scores diverged with the percent EPT taxa metric (57 (300-count) vs. 64 (100-count)) and percent predator taxa metric (79 (300-count) vs. 50 (100-count)) (Table 5). Appendix B contains plots of metric values in stressed vs. reference samples (which makes it easier to visualize what accounts for the differences in DE scores). When DEs were compared with the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020), results were mixed. Three metrics had higher DEs in the 100-count IBI dataset and three metrics had higher DEs in the 300-count dataset, with differences ranging from 0 to 26 (Table 5).

Z-scores in the 300-count CH dataset also varied across metrics, ranging from 0.57 to 2.84 (Table 5). When compared with z-scores from the 100-count versions of the samples, the percent predator taxa metric values differed the most (values were 0.65 higher in 300-count samples), followed by the total number of taxa metric (values were 0.4 higher in 300-count samples). Higher z-scores are more desirable as they indicate a better separation of reference and stressed values. When the z-scores were compared with the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020), results varied, with four metrics having higher z-scores in the 100-count IBI dataset (Table 5). CV values in the 300-count dataset ranged from 0.15 to 1.28. When

compared with CVs from the 100-count samples, values were slightly lower in the 300-count samples. The CV of the percent predator metric improved the most, with a CV that was 0.14 lower in 300-count samples. Lower CV values are more desirable as they indicate less variation relative to their means.

We also ran a Spearman rank correlation analysis to evaluate the strength of the association between 100- vs. 300-count metric values in the CH dataset. The percent predator taxa metric had the lowest correlation coefficient (r_s=0.75). The rest had stronger associations, with r_s values ranging from 0.85 to 0.95. The percent Ephemeroptera individuals (excluding Caenidae and Baetidae) metric had the highest r_s value (Table 6). Similar patterns are evident in scatterplots of metric values in 100- vs. 300-count samples, which are provided in Appendix C.

Table 4. Distribution statistics for the Central Hills riffle habitat IBI metrics, grouped by subsample size (100- vs. 300-count). Metrics are marked in red text if the mean metric values in the 100- vs. 300-count samples differ by more than one standard deviation.

Metric	Subsample Size	Minimum	5th percentile	Mean	Standard Deviation	Median	95th percentile	Maximum
Total number of taxa (nt_total)	100ct	9.0	17.0	27.1	6.1	28.0	36.0	38.0
	300ct	15.0	26.2	42.2	9.2	44.0	55.9	62.0
% EPT taxa (pt_EPT)	100ct	11.8	15.8	35.6	12.3	34.7	57.5	68.4
	300ct	6.9	13.2	33.9	11.2	34.8	49.8	66.7
% Ephemeroptera individuals, excluding Caenidae and Baetidae (pi_Ephem NoCaeBae)	100ct	0.0	0.0	6.5	9.0	4.0	30.4	47.0
	300ct	0.0	0.0	6.7	9.3	3.7	28.6	53.3
% Collector-filterer individuals (pi_ffg_filt)	100ct	0.0	13.2	38.7	16.3	40.0	62.0	86.0
	300ct	0.6	16.3	39.0	16.0	39.1	68.7	88.2
% Predator taxa (pt_ffg_pred)	100ct	0.0	4.3	15.8	7.4	16.7	28.1	31.6
	300ct	3.3	7.2	19.0	6.8	19.0	28.5	34.8
% Intolerant taxa (pt_tv_intol)	100ct	0.0	4.2	23.3	11.3	25.0	40.5	50.0
	300ct	0.0	4.1	24.7	11.2	27.1	40.0	42.6

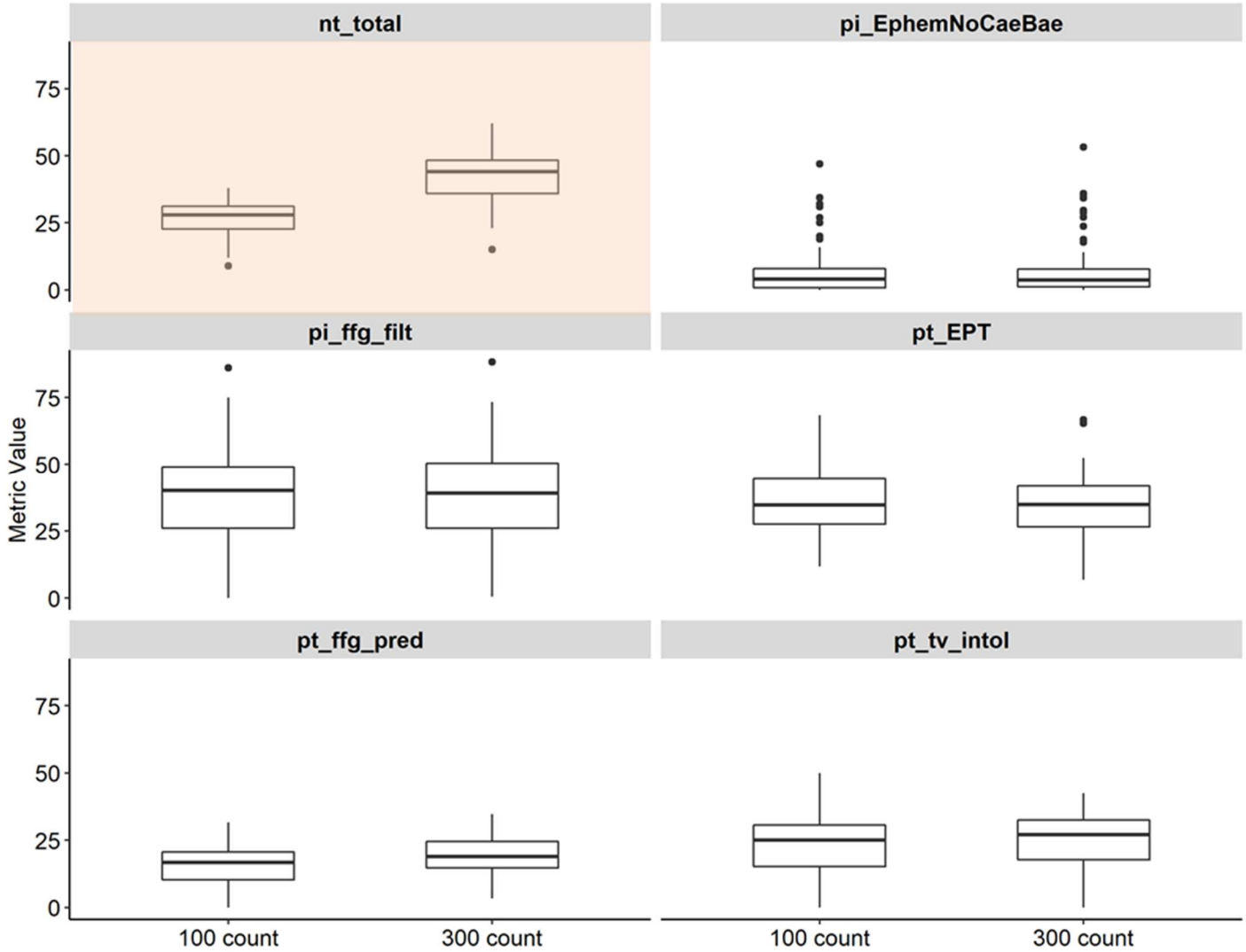


Figure 4. Distribution of Central Hills riffle habitat IBI metric values, grouped by subsample size (100- vs. 300-count). The nt_total (total number of taxa) metric is highlighted in orange because its mean metric values in 100- vs. 300-count samples differed by more than one standard deviation. Translations for the metric name abbreviations can be found in Table 4.

Table 5. Performance statistics for Central Hills riffle habitat IBI metrics in the 300-count dataset and 100-count IBI calibration dataset (Jessup and Stamp 2020). DE = discrimination efficiency and CV = coefficient of variation. The DE and z-score calculations are based on both reference and stressed samples. The CV only considers reference samples. Table 3 contains descriptions of the performance statistics.

Metric (Response to stress)	Subsample size	300-count CH dataset			*100-count IBI calibration*	
		DE	z-score	CV	*DE*	*z-score*
Total number of taxa (Dec.)	100ct	92.9	1.62	0.18	*66.7*	*1.21*
	300ct	92.9	2.03	0.15	--	--
% Ephemeroptera individuals, excluding Caenidae and Baetidae (Dec.)	100ct	71.4	0.58	1.29	*66.7*	*0.64*
	300ct	71.4	0.57	1.28	--	--
% Collector-filterer individuals (Inc.)	100ct	64.3	-0.99	0.42	*76.7*	*-1.78*
	300ct	64.3	-0.96	0.40	--	--
% EPT taxa (Dec.)	100ct	64.3	1.03	0.28	*76.7*	*1.2*
	300ct	57.1	1.15	0.24	--	--
% Predator taxa (Dec.)	100ct	50.0	0.70	0.39	*90*	*1.81*
	300ct	78.6	1.35	0.25	--	--
% Intolerant taxa (Dec.)	100ct	100.0	2.51	0.28	*100*	*2.74*
	300ct	100.0	2.84	0.24	--	--

Table 6. Spearman rank correlation coefficients (r_s) for Central Hills riffle habitat IBI metrics, in 100- vs 300-count paired samples. Correlation coefficients can range from 0 to 1, where higher values indicate stronger linear relationships.

Metric	r_s	p-value
Total number of taxa	0.85	<0.0001
% EPT taxa	0.87	<0.0001
% Ephemeroptera individuals, excluding Caenidae and Baetidae	0.95	<0.0001
% Collector-filterer individuals	0.97	<0.0001
% Predator taxa	0.75	<0.0001
% Intolerant taxa	0.89	<0.0001

4.1.2 Western Highlands

Of the six WH riffle habitat IBI input metrics, the two richness metrics – total number of taxa and Becks Biotic Index - showed the greatest difference in mean metric values in 300-count vs. 100-count samples. On average, there were 18 more taxa in the 300-count samples, and the Becks index scores were ~ 14 points higher in the 300-count dataset (Table 7). This divergence was also evident in the box plots in Figure 5. These two richness metrics were the only ones that had mean metric values that differed by more than one standard deviation in the 100- vs. 300-count datasets. Mean values of the other metrics were very similar, differing by less than <1% (Table 7).

For metric performance, the DE and z-scores were interpreted with caution in the WH dataset because they were based on so few samples (five stressed samples from three sites). DEs for the 300-count samples ranged from 0 (multiple metrics) to 80 (percent shredder individuals) (Table 8). When compared with DEs from the 100-count versions of the samples, values were the same for five of the metrics. They diverged with the number of taxa metric (20 (300-count) vs. 0 (100-count)) (Table 8). Appendix B contains plots of metric values in stressed vs. reference samples. When DEs were compared with the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020), performance in the 100-count IBI dataset was better. Five metrics had higher DE scores in the 100-count IBI (ranging from 10 to 60 points higher), while the DE for the percent shredder individuals metric was 18 points higher in the 300-count dataset (Table 8).

Absolute values of the z-scores in the 300-count WH dataset ranged from 0.06 to 0.90 (Table 8). Decreaser metrics typically receive positive z-scores (since mean metric values in reference samples are typically higher than mean metric values in stressed samples). However, in this limited dataset, three of the decreaser metrics (total number of taxa, Becks index, and percent intolerant individuals) had negative z-scores (Table 8). When compared with z-scores from the 100-count versions of the samples, four metrics had higher z-scores in the 100-count samples. When the z-scores were compared with the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020), z-scores in the 100-count IBI dataset were generally higher and occurred in the expected direction (all decreaser metrics had positive z-scores) (Table 8).

CV values in the 300-count WH dataset ranged from 0.11 to 0.7. When compared with CVs from the 100-count samples, values were slightly lower in the 300-count samples. The CV of the percent Plecoptera individuals metric improved the most, with a CV that was 0.07 lower in 300-count samples.

For the Spearman rank correlation analysis in the WH dataset, the richness metrics had the lowest correlation coefficients (Becks, $r_s = 0.76$; and total number of taxa, $r_s = 0.82$). The other metrics had r_s values ranging from 0.90 to 0.95 (Table 9). Similar patterns are evident in scatterplots of metric values in 100- vs. 300-count samples, which are provided in Appendix C.

Table 7. Distribution statistics for the Western Highlands riffle habitat IBI metrics, grouped by subsample size (100- vs. 300-count). Metrics are marked in red text if the mean metric values in the 100- vs. 300-count samples differ by more than one standard deviation.

Metric	Subsample Size	Minimum	5th percentile	Mean	Standard Deviation	Median	95th percentile	Maximum
Total number of taxa (nt_total)	100ct	19.0	23.2	32.1	4.8	33.0	38.8	41.0
	300ct	31.0	34.8	50.5	7.6	52.0	61.8	62.0
% Plecoptera individuals (pi_Pleco)	100ct	0.0	1.2	7.9	6.2	6.0	19.7	30.0
	300ct	0.0	1.0	8.2	5.9	6.6	20.6	25.3
% Collector-filterer individuals (pi_ffg_filt)	100ct	3.0	9.2	24.0	13.3	21.0	54.4	67.0
	300ct	5.6	8.1	24.3	13.1	21.9	55.0	64.1
% Shredder individuals (pi_ffg_shred)	100ct	2.0	5.0	15.4	9.4	13.0	33.6	45.0
	300ct	2.9	5.0	15.7	9.4	13.0	34.1	47.6
% Intolerant individuals (pi_tv_intol)	100ct	6.0	9.2	26.0	11.2	26.0	45.8	54.0
	300ct	6.6	9.4	26.5	10.9	26.1	45.7	54.9
Becks Biotic Index (x_Becks)	100ct	8.0	15.0	25.2	6.6	25.0	36.8	40.0
	300ct	15.0	21.8	38.8	8.2	40.0	50.6	60.0

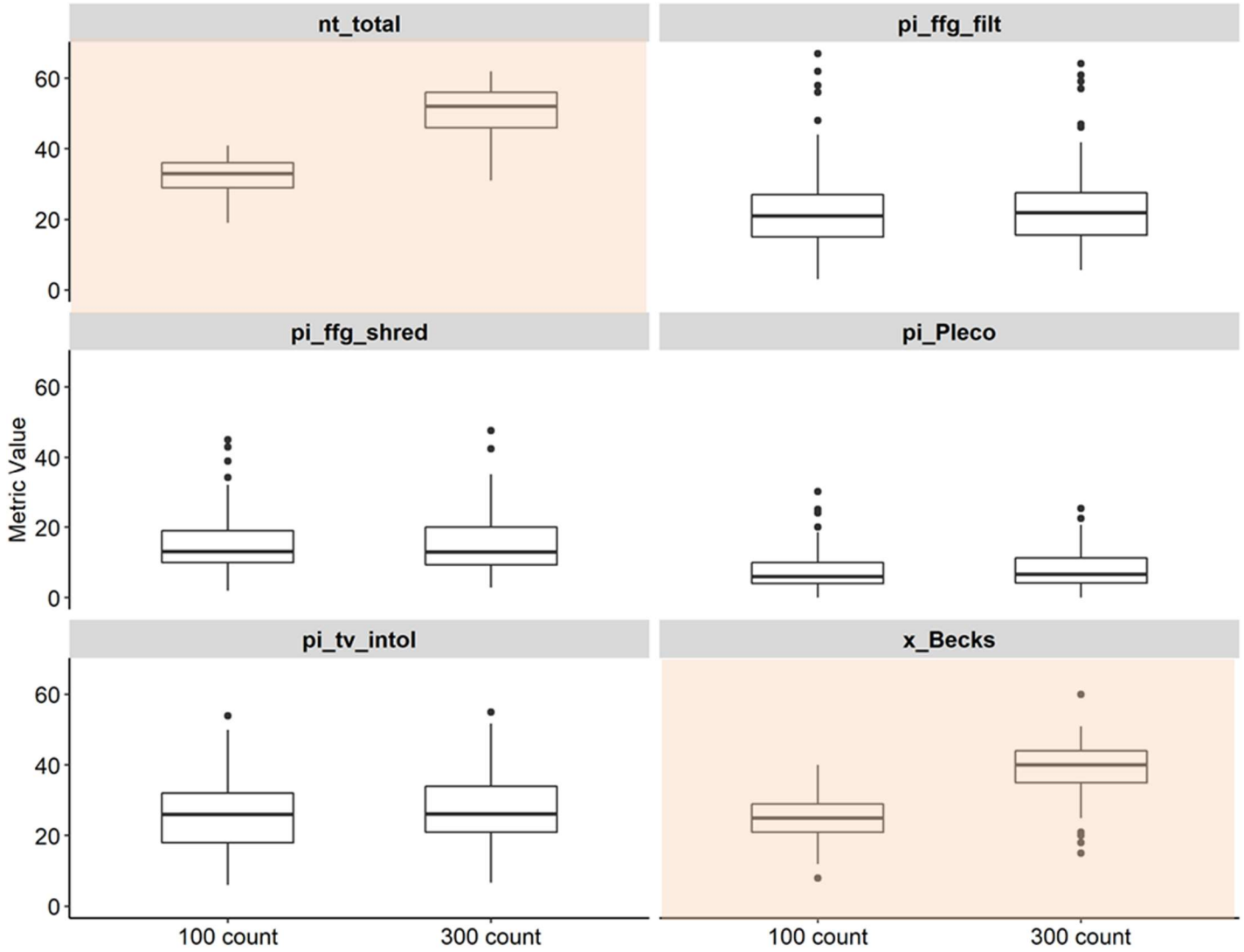


Figure 5. Distribution of Western Highlands riffle habitat IBI metric values, grouped by subsample size (100- vs. 300-count). The nt_total (total number of taxa) and x_Becks (Becks Biotic Index) metrics are highlighted in orange because their mean metric values in 100- vs. 300-count samples differed by more than one standard deviation. Translations for the metric name abbreviations can be found in Table 7.

Table 8. Performance statistics for Western Highlands riffle habitat IBI metrics in the 300-count dataset and 100-count IBI calibration dataset (Jessup and Stamp 2020). DE = discrimination efficiency and CV = coefficient of variation. The DE and z-score calculations are based on both reference and stressed samples. The CV only considers reference samples. Table 3 contains descriptions of the performance statistics.

Metric (Response to stress)	Subsample size	300-count WH dataset			*100-count IBI calibration*	
		DE	z-score	CV	*DE*	*z-score*
Total number of taxa (Dec.)	100ct	0	-1.00	0.13	*52.4*	*0.66*
	300ct	20	-0.66	0.11	--	--
% Collector-filterer individuals (Inc.)	100ct	40	-0.41	0.45	*50.0*	*-0.69*
	300ct	40	-0.38	0.45	--	--
% Shredder individuals (Dec.)	100ct	80	0.85	0.57	*61.9*	*0.84*
	300ct	80	0.89	0.55	--	--
% Plecoptera individuals (Dec.)	100ct	20	0.01	0.77	*66.7*	*1.03*
	300ct	20	0.25	0.70	--	--
% Intolerant individuals (Dec.)	100ct	0	-0.16	0.41	*59.5*	*1.02*
	300ct	0	-0.06	0.40	--	--
Becks Biotic Index (Dec.)	100ct	0	-1.44	0.21	*57.1*	*1.11*
	300ct	0	-0.90	0.15	--	--

Table 9. Spearman rank correlation coefficients (r_s) for Western Highlands riffle habitat IBI metrics, in 100- vs 300-count paired samples. Correlation coefficients can range from 0 to 1, where higher values indicate stronger linear relationships.

Metric	r_s	p-value
Total number of taxa	0.82	<0.0001
% Plecoptera individuals	0.90	<0.0001
% Collector-filterer individuals	0.92	<0.0001
% Shredder individuals	0.95	<0.0001
% Intolerant individuals	0.94	<0.0001
Becks Biotic Index	0.76	<0.0001

4.2 IBI alternatives

We explored three different index alternatives for the 300-count IBI, as described in Table 10. The IBIs use metric scoring formulae based on either the 300-count dataset or the 100-count riffle habitat IBIs (Jessup and Stamp 2020). Both sets of scoring formulae are shown in Table 11.

Table 10. Descriptions of the three 300-count IBI alternatives that were considered. Table 11 contains the metric scoring formulas.

Metric scoring scheme	**Description**
Adjust none (AdjNone)	Metric scoring formulas from the 100-count riffle habitat IBI (Jessup and Stamp 2020) were applied to all metrics.
Adjust Richness Metrics Only (AdjRichOnly)	Metric scoring formulas based on 5th and 95th percentiles in the 300-count dataset were applied to richness metrics*. Metric scoring formulas from the 100-count riffle habitat IBI (Jessup and Stamp 2020) were applied to the other metrics. Richness metrics: • Central Hills - total number of taxa • Western Highlands – total number of taxa, Becks Biotic Index
Adjust all (AdjAll)	Metric scoring formulas based on 5th and 95th percentiles in the 300-count dataset were applied to all metrics.

*Richness metrics were selected because they differed the most between 100- vs 300-count samples (their mean metric values differed by more than 1 standard deviation).

When evaluating IBI performance, the DE and z-scores were interpreted with caution due to the limited size of the 300-count dataset. In the CH dataset, all three IBI alternatives were effective at discriminating between reference and stressed samples. DEs were all 100 and z-scores ranged from 2.61 (AdjNone) to 2.65 (AdjRichOnly). CVs were low for all alternatives (ranging from 0.14 to 0.15) (Table 12). When index discrimination was compared with the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020), the DEs were the same (100) and the 100-count IBI had a higher z-score (3 vs. 2.6) (Table 12). Figure 6 shows the distribution of IBI scores in reference vs. stressed samples for each of the three IBI alternatives in the CH dataset.

It was difficult to evaluate the performance of the WH IBI alternatives based on DE and z-scores due to the very low number of stressed sites. DE scores were low (ranging from 20 to 60), as were z-scores (ranging from 0.28 to 0.56). The 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020) had a DE of 88 and a z-score of 1.4 (Table 12). The relatively high IBI values in the five stressed samples were responsible for the low DE scores in the 300-count IBI alternatives, as shown in Figure 7. CVs were similar across IBI alternatives, with the AdjNone IBI having the lowest value (0.16 compared to 0.18) (Table 12).

Table 11. Metric scoring formulas based on the 300-count dataset vs the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020). The scoring formula for 'decreaser' metrics = 100(Metric value – minimum possible value)/(95th percentile-minimum) and the formula for 'increaser' metrics = 100*(95th percentile-metric value)/(95th percentile-5th percentile). The minimum possible value for these metrics is 0. To simplify the formulas, the 0's in the 'decreaser' formulas are not shown.*

Region	**Metric (Response to stress)**	**Percentiles and scoring formulae based on the 300-count riffle habitat IBI dataset**			***Percentiles and scoring formulae based on the original 100-count IBI calibration dataset (Jessup and Stamp 2020)***		
		5th	**95th**	**Scoring formulae**	***5th***	***95th***	***Scoring formulae***
Central Hills	Total number of taxa (Dec.)	26.2	55.9	100*(metric)/ 55.8	*11.0*	*34.9*	*100*(metric)/ 34.9*
	% EPT taxa (Dec.)	13.2	49.8	100*(metric)/ 49.8	*10.6*	*54.5*	*100*(metric)/54.5*
	% Ephemeroptera individuals, excluding Caenidae and Baetidae (Dec.)	0.0	28.6	100*(metric)/ 28.5	*0.0*	*13.9*	*100*(metric)/13.9*
	% Collector-filterer individuals (Inc.)	16.3	68.7	100*(68.7-metric)/ 52.4	*13.0*	*79.9*	*100*(79.9-metric)/66.9*
	% Predator taxa (Dec.)	7.2	28.5	100*(metric)/ 28.5	*0.0*	*28.5*	*100*(metric)/28.5*
	% Intolerant taxa (Dec.)	4.1	40.0	100*(metric)/ 40	*0.0*	*39.1*	*100*(metric)/39.1*
Western Highlands	Total number of taxa (Dec.)	34.8	61.8	100*(metric)/ 61.8	*21.0*	*38.8*	*100*(metric)/38.8*
	% Plecoptera individuals (Dec.)	0.95	20.6	100*(metric)/ 20.6	*0.0*	*18.3*	*100*(metric)/18.3*
	% Collector-filterer individuals (Inc.)	8.11	55.03	100*(55.03-metric)/ 46.9	*9.8*	*50.5*	*100*(50.5-metric)/40.7*
	% Shredder individuals (Dec.)	5.03	34.1	100*(metric)/ 34.1	*1.2*	*23.0*	*100*(metric)/23*
	% Intolerant individuals (Dec.)	9.35	45.65	100*(metric)/ 45.6	*6.1*	*51.5*	*100*(metric)/51.5*
	Becks Biotic Index (Dec.)	21.8	50.6	100*(metric)/ 50.6	*12.0*	*36.8*	*100*(metric)/36.8*

Table 12. Performance statistics for the 300-count riffle habitat IBI alternatives, compared with performance statistics from the 100-count IBI calibration dataset (Jessup and Stamp 2020). DE = discrimination efficiency and CV = coefficient of variation. The DE and z-score calculations are based on both reference and stressed samples. The CV only considers reference samples. Table 3 contains descriptions of the performance statistics.

Region	Scoring scheme	DE	Z- score	CV
Central Hills	AdjNone	100	2.61	0.14
	AdjRichOnly	100	2.65	0.15
	AdjAll	100	2.64	0.15
	100-count IBI calibration dataset	***100***	***3.02***	***--***
Western Highlands	AdjNone	60	0.56	0.16
	AdjRichOnly	20	0.34	0.18
	AdjAll	20	0.28	0.18
	100-count IBI calibration dataset	***88.1***	***1.40***	***--***

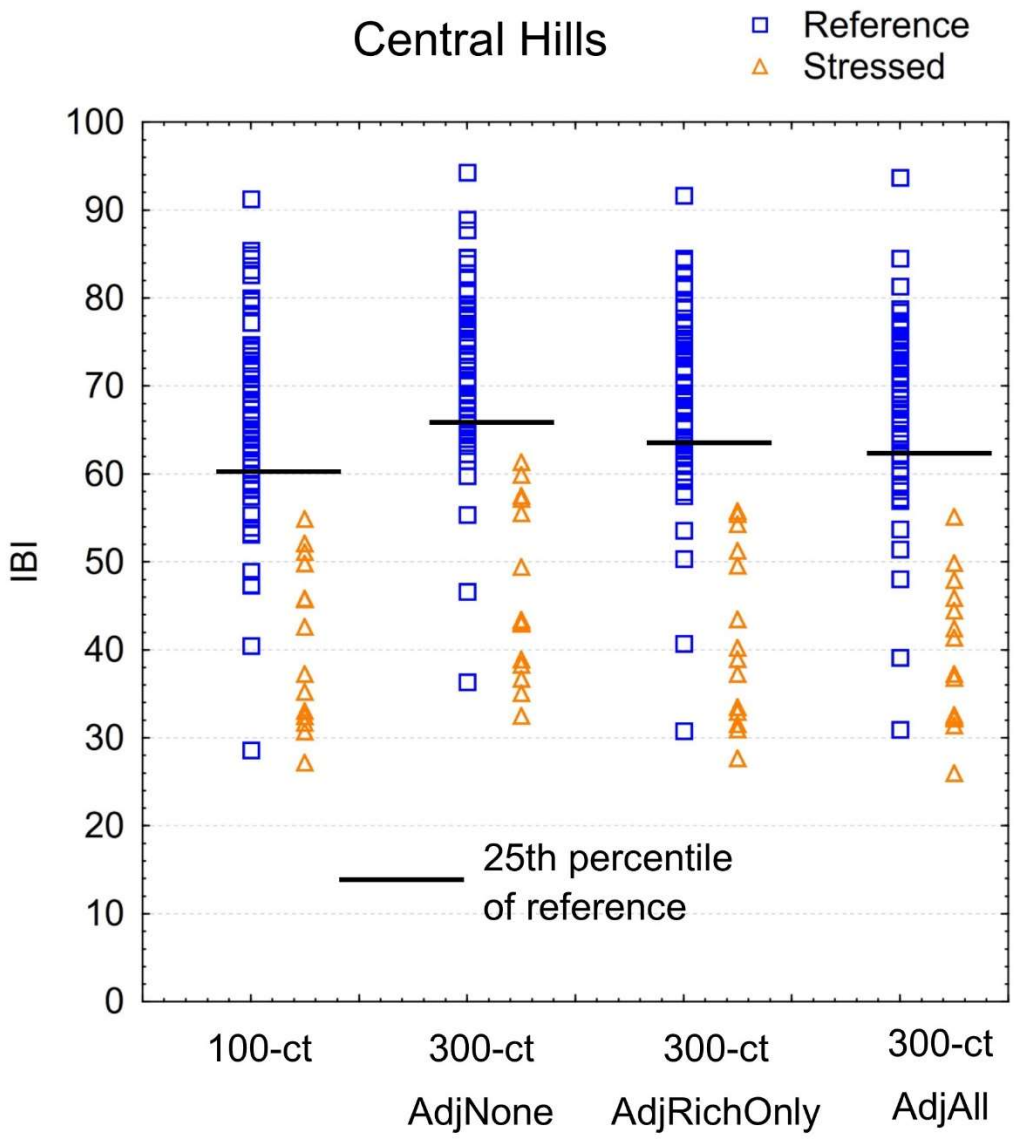


Figure 6. Distribution of IBI scores in reference vs. stressed samples for the IBI alternatives, based on the Central Hills dataset. IBI scores for the 100-count version of the samples (calculated using the AdjNone scoring scheme) were also included. Because IBI scores for all the stressed sites were less than the 25th percentile of reference, DEs were 100.

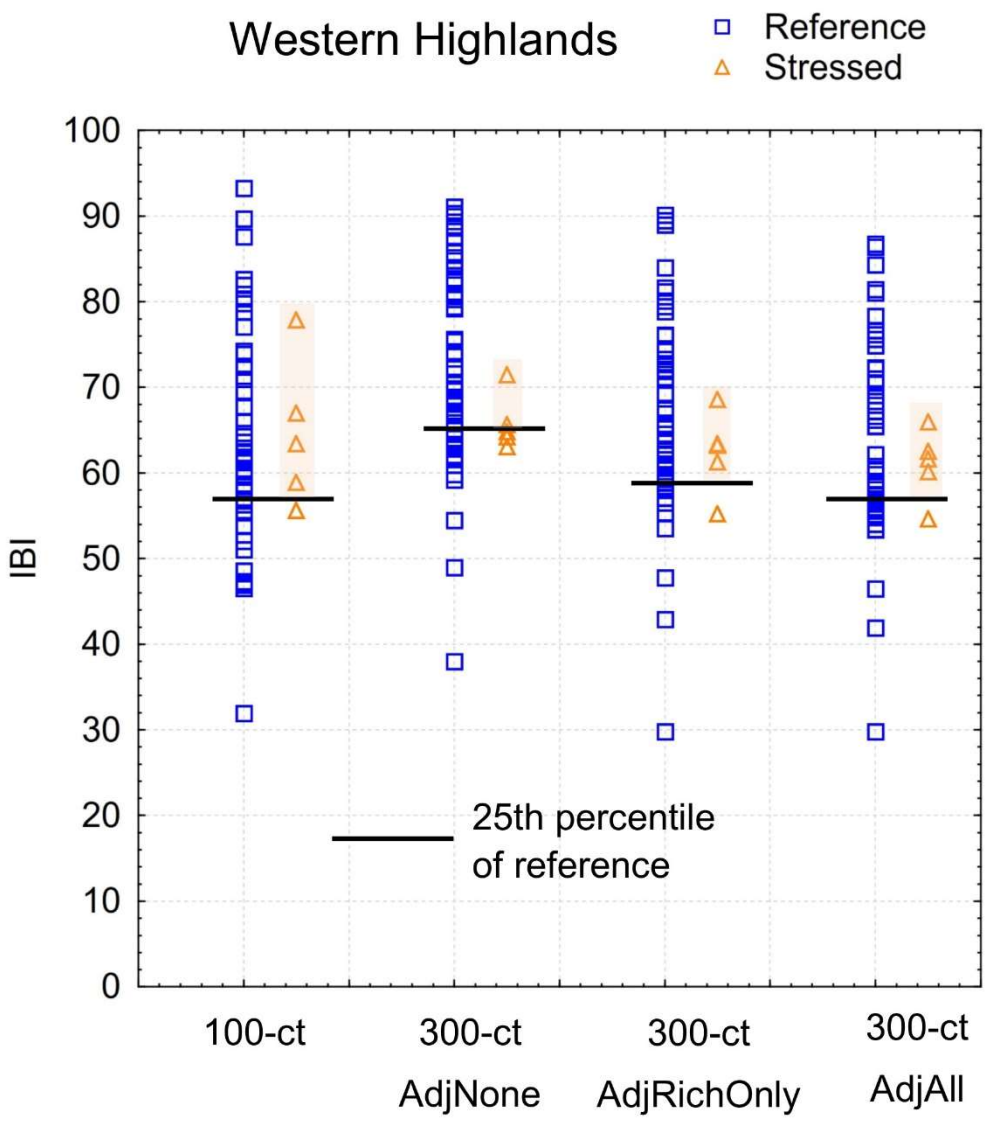


Figure 7. Distribution of IBI scores in reference vs. stressed samples for the IBI alternatives, based on the Western Highlands dataset. IBI scores for the 100-count version of the samples (calculated using the AdjNone scoring scheme) were also included. Many of the IBI scores for the stressed sites were greater than the 25th percentile of reference, so DEs were low (≤60).

As an alternate way to assess IBI performance, we ran a Spearman rank correlation analysis to assess the response of the three IBI alternatives to four disturbance variables (ICI, IWI, percent urban, and percent agriculture) and compared the correlation coefficients to a comparable analysis run on the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020). In the CH dataset, aside from the percent agriculture metric, r_s values were lower in the 300-count IBIs. However, the relationships were fairly strong ($|\geq 0.46|$) and were in the expected direction. With the agricultural metric, the 300-count IBI alternatives performed better (both in magnitude as well as the direction of response; the 100-count riffle habitat IBI went against expectations, as it had a positive instead of negative relationship with percent agriculture).

In the WH dataset, results were generally similar. The r_s values were slightly lower for the 300-count IBIs vs. the 100-count IBI (Jessup and Stamp 2020) (Table 12). Unlike the CH results, two IBIs (AdjRichOnly and AdjAll) were not significantly correlated with the IWI ($p>0.05$). The AdjNone IBI had the highest r_s values for each variable in the WH dataset.

Table 13. Spearman rank correlation coefficients (r_s) for the 300-count IBI alternatives, compared with results from the 100-count riffle habitat IBI calibration dataset (taken from Jessup and Stamp 2020). Correlations in red text are significant ($p<0.05$). Descriptions of the disturbance variables can be found in Appendix A.

Region	Scoring scheme	Spearman rank correlations (r_s)			
		ICI v. 2.1	**IWI v. 2.1**	**% Urban (Ws*)**	**% Agricultural (Cat*)**
Central Hills	AdjNone	0.47	0.56	-0.59	-0.21
	AdjRichOnly	0.46	0.54	-0.58	-0.20
	AdjAll	0.48	0.56	-0.61	-0.23
	100-count IBI calibration dataset	***0.54***	***0.70***	***-0.72***	***0.20***
Western Highlands	AdjNone	0.38	0.27	-0.48	-0.39
	AdjRichOnly	0.28	0.21	-0.42	-0.31
	AdjAll	0.27	0.20	-0.42	-0.31
	100-count IBI calibration dataset	***0.50***	***0.46***	***-0.50***	***-0.34***

*Ws = total watershed scale; Cat = local catchment scale

4.3 Effects of subsample size on IBI scores and thresholds

When we evaluated differences between IBI scores in 300- vs. 100-count samples on a sample-by-sample basis, the AdjNone IBIs showed the greatest difference. AdjNone median IBI scores were ~5 points higher in the 300- vs. 100-count samples in the CH dataset and ~8 points higher in the WH dataset (Figures 8 and 9, respectively). Scores were most similar for the AdjAll IBI, with median scores ~1 point lower than their 100-count counterparts in both regions. The AdjRichOnly IBI scores fell in-between, with median scores for the 300-count IBIs ~1 point higher in the CH dataset and ~2 points higher in the WH dataset. These patterns are also evident in the 100- vs. 300-count IBI scatterplots in Appendix D. Attachments A & B contain IBI alternative scores for the 300- vs. 100-count versions of each sample in the CH and WH datasets, respectively.

Differences in IBI alternatives were also evident in the precision statistics. The AdjNone IBI had the highest values, which indicate a greater divergence between IBI scores in 100- vs. 300-count samples. Values for the AdjRichOnly and AdjNone IBIs were very similar, and were lower than the AdjNone IBI results, indicating that they are less affected by subsample size (Table 14).

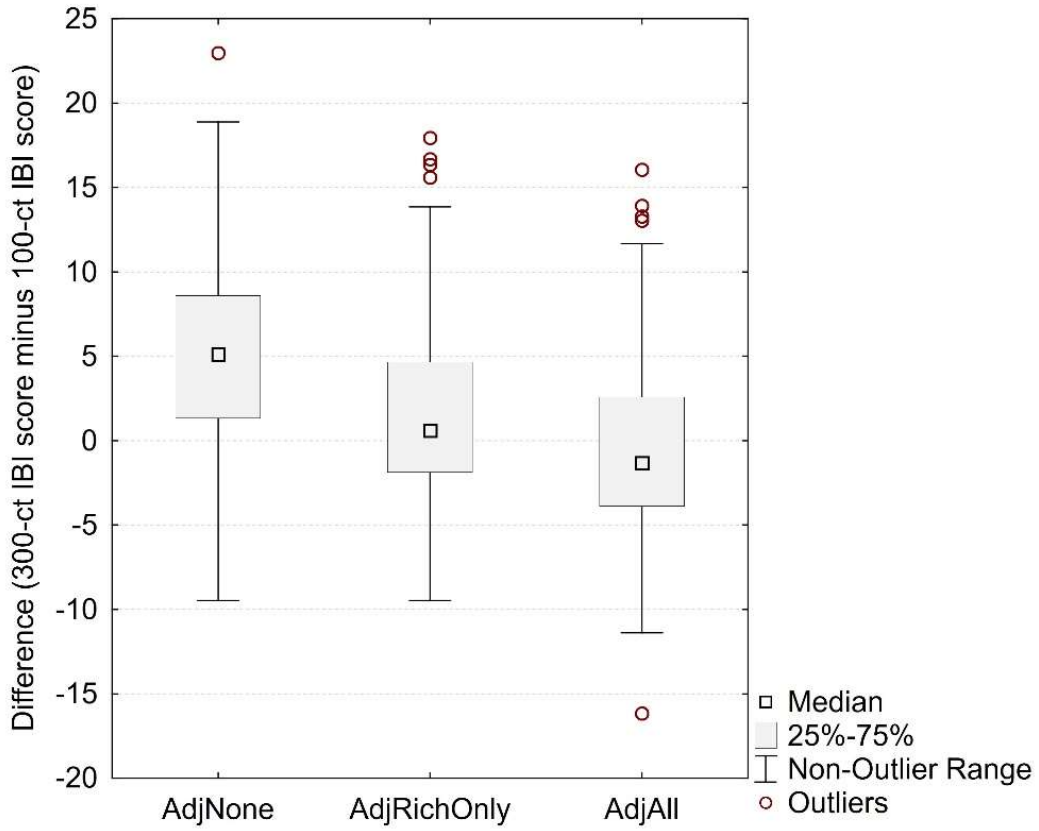


Figure 8. Central Hills. Differences between IBI scores in paired 100- vs. 300-count samples for each index alternative.

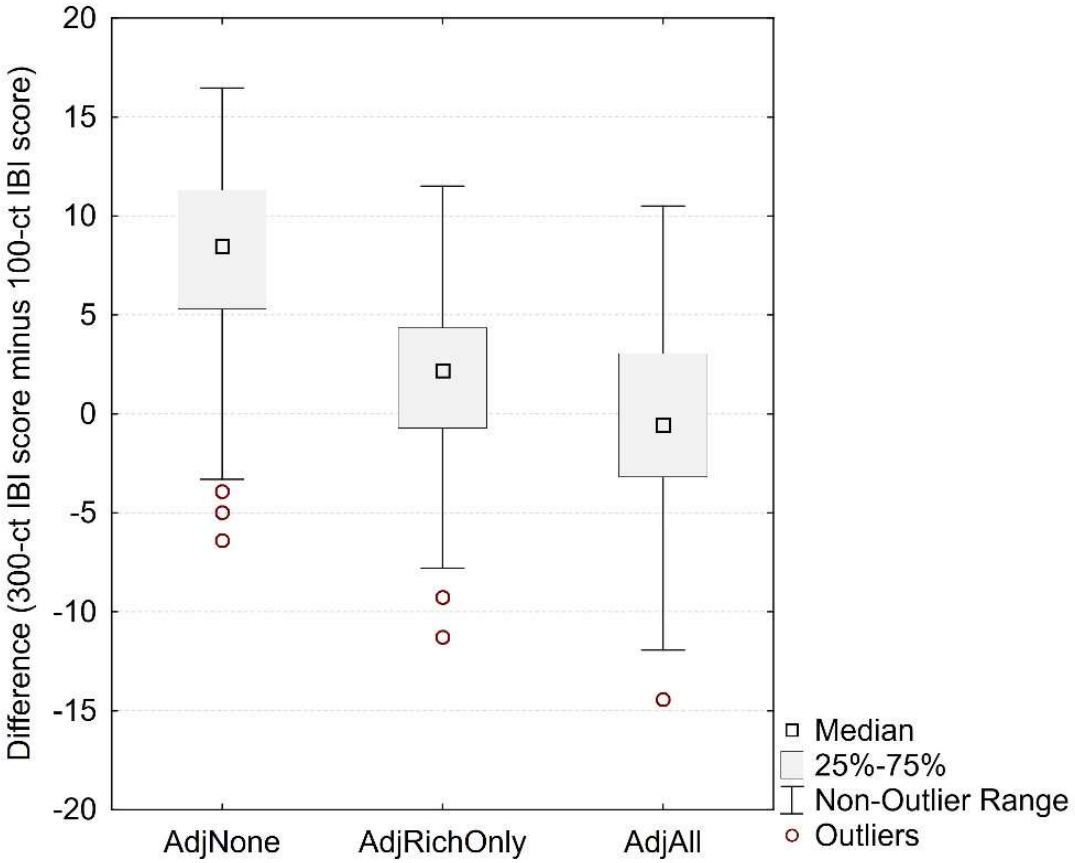


Figure 9. Western Highlands. Differences between IBI scores in paired 100- vs. 300-count samples for each index alternative.

Table 14. Precision statistics for IBI scores based on 100- vs 300-count samples, grouped by scoring scheme. MSE = mean squared error – from one-way ANOVA; Overall Mean = mean index score for the combined 100ct and 300ct samples; RMSE = root mean squared error – from one-way ANOVA; CV = coefficient of variation (RMSE/ Total Mean); CI 90 = 90% confidence interval.

Region	IBI Scores Comparison	MSE	Overall Mean	RMSE	CV	CI 90
Central Hills	100ct vs AdjNone	30.5	61.3	5.5	9.0	9.1
	100ct vs AdjRichOnly	17.4	59.6	4.2	7.0	6.9
	100ct vs AdjAll	17.1	58.4	4.1	7.1	6.8
Western Highlands	100ct vs AdjNone	41.7	65.2	6.5	9.9	10.6
	100ct vs AdjRichOnly	12.0	61.2	3.5	5.7	5.7
	100ct vs AdjAll	13.5	62.3	3.7	5.9	6.0

When we evaluated reference statistics calculated based on the 300-count IBI alternatives vs. those calculated from the 100-count riffle habitat IBI calibration dataset (Jessup and Stamp 2020), differences were greatest at the lowest percentiles (in this case, the 10th percentile), with 300-count IBIs having higher values than those derived from the 100-count IBI calibration dataset (Table 15). The reference percentiles for the AdjNone IBI scores differed the most, with the 10th percentile being ~8 points higher than the 100-count equivalent in the CH dataset and ~14 points higher in the WH dataset. The AdjAll IBI reference percentiles were most similar to the 100-count IBI values, while AdjRichOnly IBI percentiles fell in-between.

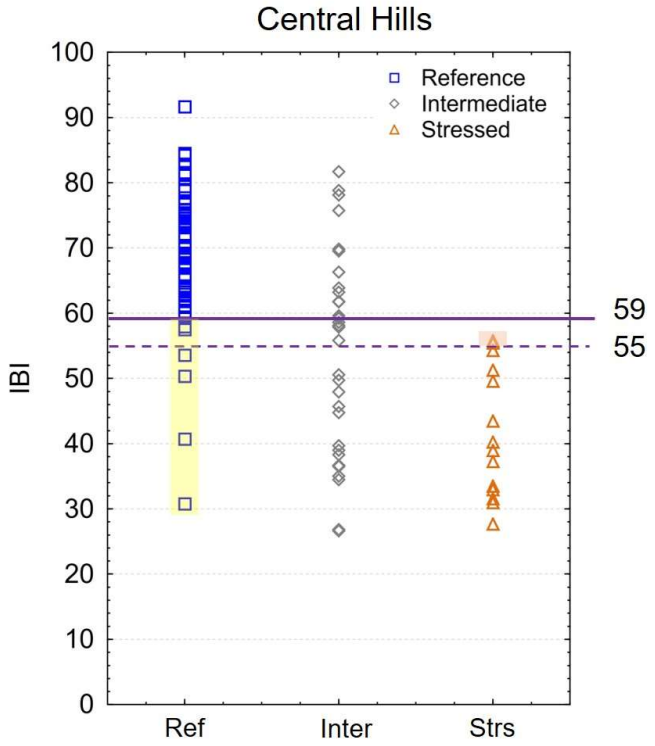
Of particular interest were differences between the 10th percentile of reference in the CH dataset and the 15th percentile of reference in the WH dataset, as these were used to derive the preliminary 100-count riffle habitat IBI threshold of 55, which distinguishes between satisfactory vs. moderately degraded condition (for more information, see the 100-count IBI Thresholds document; Stamp and Jessup 2020). In the CH dataset, the 10th percentile of reference for the AdjRichOnly IBI was ~4 points higher (59 (300-count) vs. 55 (100-count)) and ~2 points higher for the AdjAll IBI (Table 15). In the WH dataset, the 15th percentile of reference for the AdjRichOnly IBI was ~3 points higher (58 (300-count) vs. 55 (100-count)), and ~0.5 points higher for the AdjAll IBI (Table 15).

Table 15. Comparison of IBI scores for multiple percentiles of reference. The percentiles highlighted in green (10th in Central Hills and 15th in Western Highlands) correspond to the percentiles currently being used for the 100-count IBI satisfactory/moderately degraded condition threshold in each region. For more information, see the Thresholds report (Stamp and Jessup 2020).

Percentile of Reference	**AdjNone (300ct)**	**AdjRichOnly (300ct)**	**AdjAll (300ct)**	**AdjNone (100ct)**	***100-count IBI calibration dataset (Jessup and Stamp 2020)***
Central Hills IBI scores					
10th	62.9	59	57.2	53.3	*55.2*
15th	64.6	61.4	58.6	56.5	*58.5*
20th	65.7	62.5	59.9	58.5	*61*
25th	66.1	63.8	61.8	60	*64.5*
50th	72.6	69	67.4	67.4	*69.4*
Western Highland IBI scores					
10th	60.8	55.3	53.7	47.9	*47*
15th	62.7	57.7	55.4	51.8	*55*
20th	63.3	58.5	56	55.4	*57.8*
25th	65.3	59	56.8	56.8	*59.6*
50th	70.5	64.9	61.5	61.7	*62.7*

When we calculated Type I and II error rates for the AdjRichOnly IBI based on the two sets of thresholds (55 vs. the 10th percentile of reference in the CH and 15th percentile of reference in the WH), error rates in the CH dataset were equal to or less than the rates reported in the 100-count riffle habitat IBI threshold document (Stamp and Jessup 2020). When 55 was used as the threshold, four of the 58 reference samples scored below the threshold (equaling a Type I error rate of 6.9%, which is less than the 100-count rate of 11.1%) and two of the 14 stressed samples scored above the threshold (Type II error = 14.3%, which is the same as the 100-count rate) (Figure 10). When using 59 as the threshold (which equals the 10th percentile of reference based on 300-count CH dataset), the Type I error rate increased to 10% and the Type II error rate dropped to 0 (meaning IBI scores for all stressed samples were less than 59).

For the WH dataset, when 55 was used as the threshold, four of the 46 reference samples scored below the threshold (equaling a Type I error rate of 8.7%, which is less than the 100-count rate of 14.6%) and all five of the stressed samples scored above the threshold (Type II error = 100%, which is much higher than the 100-count rate of 17.1%). When we switched to a threshold of 58 (which equals the 15th percentile of reference based on 300-count WH dataset), the Type I error rate increased to 15% (approximately the same as the 100-count rate) and the Type II error rate decreased to 80% (still much higher than the 100-count rate).

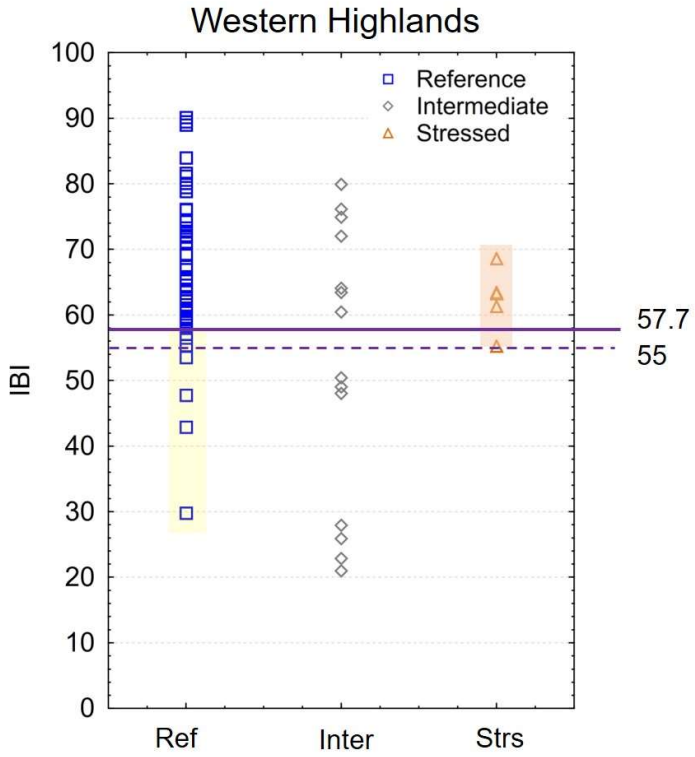


Type I error
- Erroneously calling a site disturbed
- Calculated as the % of Ref samples with IBI scores < the threshold. For example, if 59 is the threshold, 6 out of 58 ref samples (or 10.3%) have IBI scores < 59

Type II error
- Erroneously calling a disturbed site undisturbed
- Calculated as the % of stressed samples with IBI scores ≥ the threshold. For example, if 59 is the threshold, 0 out of 14 strs samples (or 0%) have IBI scores ≥ 59

IBI threshold	Rationale	Number of samples in each disturbance group						Type I error	Type II error	Difce
		≥ Threshold			< Threshold					
		Ref	Inter	Strs	Ref	Inter	Strs			
59	10th percentile of reference - AdjRichOnly IBI, 300-ct dataset	52	13	0	6	19	14	10.3%	0%	10.3%
55	10th percentile of reference - original 100-ct kick IBI calibration dataset	54	17	2	4	15	12	6.9%	14.3%	7.4%
		Error rates, 100-ct IBI, using threshold of 55						***11.1%***	***14.3%***	***3.2%***

Figure 10. Central Hills. Distribution of AdjRichOnly IBI scores across the three broad disturbance categories (reference (Ref), stressed (Strs), and intermediate (Inter). The table summarizes Type I and II error rates and the number and percentages of samples in each disturbance category that fell above or below two thresholds (59, which is based on the 10th percentile of reference vs 55, which is the satisfactory/moderately degraded condition threshold in the 100-count IBI. Cells highlighted in yellow show Type I error rates; light orange cells show Type II error rates.



Type I error
- Erroneously calling a site disturbed
- Calculated as the % of Ref samples with IBI scores < the threshold. For example, if 57.7 is the threshold, 7 out of 46 ref samples (or 15.2%) have IBI scores < 57.7

Type II error
- Erroneously calling a disturbed site undisturbed
- Calculated as the % of stressed samples with IBI scores ≥ the threshold. For example, if 57.7 is the threshold, 4 out of 5 strs samples (or 80%) have IBI scores ≥ 57.7

IBI threshold	Rationale	Number of samples in each disturbance group						Type I error	Type II error	Difce
		≥ Threshold			< Threshold					
		Ref	Inter	Strs	Ref	Inter	Strs			
57.7	15th percentile of reference - AdjRichOnly IBI, 300-ct dataset	39	7	4	7	7	1	15.2%	80%	64.8%
55	15th percentile of reference - original 100-ct kick IBI calibration dataset	42	7	5	4	7	0	8.7%	100%	91.3%
	Error rates, 100-ct IBI, using threshold of 55							***14.6%***	***17.1%***	***2.5%***

Figure 11. Western Highlands. Distribution of AdjRichOnly IBI scores across the three broad disturbance categories (reference (Ref), stressed (Strs), and intermediate (Inter). The table summarizes Type I and II error rates and the number and percentages of samples in each disturbance category that fell above or below two thresholds (59, which is based on the 10th percentile of reference vs 55, which is the satisfactory/moderately degraded condition threshold in the 100-count IBI. Cells highlighted in yellow show Type I error rates; light orange cells show Type II error rates.

5 Conclusions

The purpose of this exercise was to develop 300-count versions of the 100-count riffle habitat IBIs that were recently developed for freshwater perennial wadeable streams in the WH and CH regions (Jessup and Stamp 2020). We evaluated three 300-count IBI alternatives, each of which had the same input metrics as the 100-count riffle habitat IBIs but differed in that they used different combinations of metric scoring formulas (ranging from no adjustments to adjusting all scoring formulas to reflect the distribution of values in the 300-count dataset). Compared to the 100-count IBI calibration dataset, the 300-count dataset was limited in that it had fewer samples and did not capture as wide a disturbance gradient. This affected our confidence in two of the performance statistics (DE and z-score), which are based on stressed samples. The lack of stressed sites was a bigger issue in the WH vs. CH dataset, as the WH stressed dataset only had five samples from three sites, and the level of disturbance at those sites was not particularly high. Thus, performance statistics for the 300-count samples were interpreted with caution, particularly in the WH dataset.

When we evaluated subsample size effects on individual metrics, we found that differences were most evident in the richness metrics, which was expected given how higher numbers of taxa generally occur in samples in which more individuals are counted (Gotelli and Graves 1996). There is one richness metric in the CH IBI (total number of taxa) and two richness metrics in the WH IBI (total number of taxa and Becks Biotic Index). Both had mean metric values that differed by more than one standard deviation in 100- vs. 300-count samples. Subsample size effects were much smaller in the percent taxa and percent individuals metrics. Metric performance (as measured by DE, z-score, and CV) varied depending on the metric but, overall, was fairly similar between 100- vs. 300-count samples. The most consistent pattern was that CVs in the 300-count samples were typically lower than the 100-count samples (lower CV values are more desirable as they indicate less variation relative to their means).

Regarding scoring differences across the three IBI alternatives, when 100-count IBI metric scoring formulas were applied to the 300-count samples without making any adjustments (AdjNone), median IBI scores were ~5 points higher in the 300 vs. 100-count samples in the CH dataset and ~8 points higher in the WH dataset. When adjustments were made to the richness metrics to account for subsample size effects (AdjRichOnly), it reduced these differences. Differences were further reduced when all six metrics were adjusted (AdjAll) based on distribution statistics (5^{th}/95^{th} percentiles) in the 300-count dataset.

When we evaluated the effectiveness of the three IBI alternatives at discriminating between reference and stressed samples, their performance was fairly similar. In the CH dataset, all three IBIs performed well, with DEs of 100 (which was also the DE for the 100-count CH IBI; Jessup and Stamp 2020). Z-scores were slightly higher for the 100-count CH IBI (3.0 vs. 2.6). For the WH dataset, results for the three IBI alternatives were similar but were much poorer than the CH IBIs. The AdjNone IBI performed the best, with a DE of 60 compared to DEs of 20 for the other two IBIs (vs. a DE of 88 for the 100-count IBI; Jessup and Stamp 2020). Z-scores were also low ($\leq$ 0.6, compared to a z-score of 1.4 for the 100-count WH IBI). However, due to the known limitations with the WH stressed dataset, these results were interpreted with caution.

When we examined associations between the three IBI alternatives and disturbance variables, the three IBIs performed similarly in the strength and direction of the relationships with the ICI, IWI, percent urban, and percent agricultural land cover. Compared to results from similar analyses performed on the 100-count riffle habitat IBIs (Jessup and Stamp 2020), associations tended to be slightly weaker but were mostly significant ($p<0.05$) and in keeping with the expected direction of response. The weaker associations were likely driven in part by the limitations of the 300-count dataset, which had fewer sites and a more limited disturbance gradient, particularly in regard to urban land cover in the WH region.

After reviewing the results, the MassDEP workgroup selected the AdjRichOnly IBI (Table 16). They felt there was a clear need to adjust metric scoring formulas for the richness metrics to account for subsample size effects. Due to limitations with the 300-count dataset, they wanted to minimize the number of adjustments that were being made to scoring formulas for the other metrics (which are less affected by subsample size). Instead, they wanted scoring formulas for those metrics to be based on the larger, more robust 100-count dataset that was used to calibrate the 100-count IBIs.

MassDEP will interpret results from the 300-count CH and WH IBIs with caution, recognizing that their performance will need to be reevaluated in coming years after they obtain more 300-count samples. In the meantime, MassDEP is planning to calculate IBI scores for both 100- and 300-count versions of the CH and WH samples, which can be done using the MassIBI Tools Shiny app calculator (https://tetratech-wtr-wne.shinyapps.io/MassIBItools/). MassDEP will continue to evaluate both the 100- and 300-count riffle habitat IBIs as new data are collected, asking questions such as: are results in keeping with expectations? At what types of sites are the IBIs performing well? Where are they performing poorly and why? In addition, MassDEP is planning to continue conducting targeted sampling to broaden the disturbance gradients represented in each region, with a particular focus on sampling more high-stress sites in the WH region.

MassDEP will also consider potential thresholds for numeric bio-criteria as they evaluate IBI results in coming years. MassDEP does not currently have plans to pursue numeric bio-criteria in the SWQS but has identified preliminary thresholds for the 100-count riffle habitat IBIs for four biological condition categories (exceptional condition, satisfactory condition, moderately degraded, and severely degraded), as described in Stamp and Jessup (2020), for use in the Consolidated Assessment and Listing Methodology (CALM) to interpret the narrative biological criteria in the SWQS. During this exercise, we took a preliminary look at whether it would be feasible to use the same thresholds for the 300-count IBIs. Results suggest the satisfactory/moderately degraded thresholds were similar (between 55-60), although reference percentiles for the 300-count IBIs tended to be slightly higher. Type I and Type II error rates were also similar between the 100-count and 300-count IBIs, with the exception of the Type II error rates in the WH dataset (which should be interpreted with caution due to limitations of the stressed samples).

Table 16. Metrics and scoring formulas in version 1 of the Central Hills and Western Highlands 300-count riffle habitat IBIs. These are based on the AdjRichOnly scoring scheme (described in Table 10). The metric scoring formulas highlighted in light blue differ from those used in the 100-count IBIs. These formulas were changed to account for effects of subsample size on the richness metrics.

Central Hills 300-count riffle habitat IBI		
Metric	**Response to stress**	**Scoring formula**
Total number of taxa *	Decrease	100*(metric)/55.8
% EPT taxa	Decrease	100*(metric)/54.5
% Ephemeroptera individuals, excluding Caenidae and Baetidae	Decrease	100*(metric)/13.9
% Collector-filterer individuals	Increase	100*(79.9-metric)/66.9
% Predator taxa	Decrease	100*(metric)/28.5
% Intolerant taxa	Decrease	100*(metric)/39.1
Western Highlands 300-count riffle habitat IBI		
Metric	**Response to stress**	**Scoring formula**
Total number of taxa *	Decrease	100*(metric)/61.8
% Plecoptera individuals	Decrease	100*(metric)/18.3
% Collector-filterer individuals	Increase	100*(50.5-metric)/40.7
% Shredder individuals	Decrease	100*(metric)/23
% Intolerant individuals	Decrease	100*(metric)/51.5
Becks Biotic Index*	Decrease	100*(metric)/50.6

6 Literature Cited

Barbour, M.T., J. Gerritsen, B.D. Snyder, and J.B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C.

Breine, J., Maes, J., Quataert, P., Van den Bergh, E., Simoens, I., Thuyne, G., and C. Belpaire. 2007. A fish-based assessment tool for the ecological quality of the brackish Schelde estuary in Flanders (Belgium). Hydrobiologia 575: 141-159.

Cohen, J. 1992. A power primer. Psychological Bulletin, 112(1):155.
Flotemersch, J.E., Leibowitz, S.G., Hill, R.A., Stoddard, J.L., Thomas, M.C., & Tharme, R.E. 2016. A watershed Integrity Definition and Assessment approach to Support Strategic Management of Watersheds. River Research and Applications, 32, 1654–1671.

Gotelli, N. J. and G. R., Graves. 1996. Null models in ecology. Washington, DC: Smithsonian Institution Press.

Hughes, R. M., P. R. Kaufmann, A. T. Herlihy, T. M. Kincaid, L. Reynolds, and D. P. Larsen. 1998. A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences. 55(7):1618-1631.

Jessup, B., and J. Stamp. 2020. Development of Indices of Biotic Integrity for Assessing Macroinvertebrate Assemblages in Massachusetts Freshwater Wadeable Streams. Prepared for the Massachusetts Department of Environmental Protection.

Johnson, Z.G., Leibowitz, S. and R. Hill. 2019. Revising the index of watershed integrity national maps. Science of The Total Environment. 10.1016/j.scitotenv.2018.10.112.

Leppo, E. 2020. BioMonTools: Tools for biomonitoring and bioassessment; metric calculation for benthic macroinvertebrates, fish, and periphyton. https://github.com/leppott/BioMonTools.

Maxted, J.R., M.T. Barbour, J. Gerritsen, V. Poretti, N. Primrose, A. Silvia, D. Penrose, and R. Renfrow. 2000. Assessment framework for mid-Atlantic coastal plain streams using benthic macroinvertebrates. Journal of the North American Benthological Society 19(1):128–144.

Ofenböck, T., O. Moog, J. Gerritsen, and M. Barbour. 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. In Integrated Assessment of Running Waters in Europe (pp. 251-268). Springer Netherlands.

Stamp, J., and B., Jessup. 2020. Establishing numeric biological condition thresholds. Prepared for the Massachusetts Department of Environmental Protection.

Thornbrugh, D. J., Leibowitz, S.G., Hill, R. A., Weber, M. H., Johnson, Z.C. Olsen, A. R., Flotemersch, J. E., Stoddard, J. L., & Peck, D. V. 2018. Mapping watershed integrity for the conterminous United States. Ecological Indicators, 85, 1133-1148.

Appendix A

Derivation of site disturbance category designations

Disturbance category designations for most sites had already been made during the calibration of the 100-count riffle habitat IBIs, using the procedures described in Section 3 of the 100-count riffle habitat IBI report (Jessup and Stamp 2020). For this exercise, there were 31 new sites sampled in 2019 that needed disturbance category assignments. During development of the MassDEP low gradient IBI (which took place after the calibration of the 100-count riffle habitat IBI), the process (outlined in Figure A1) stayed the same but the following changes were made:

- We switched to version 2.1 of the ICI and IWI (in place of version 1) and adjusted the ICI and IWI metric thresholds to account for this change
- We switched to the National Land Cover Database (NLCD) 2016 land cover metrics (in place of NLCD 2011)
- We used two spatial scales (local catchment (Cat) and watershed (Ws)[1]) instead of one

When assigning the 2019 sites to disturbance categories, we first spatially associated the biological sampling sites with the National Hydrography Dataset (NHD) Plus Version 2 (NHDPlusV2) geospatial layer (McKay et al. 2012)[2] by performing an intersect procedure with Geographic Information System software (ArcGIS 10.7.1). This created an attribute table that included the list of biological sampling stations and unique identifiers for the NHDPlusV2 catchments (COMID/FEATUREID). The COMID was then used to link the biological sampling sites with USEPA's Stream-Catchment (StreamCat) Dataset[3] (Hill et al. 2016), which is the source of the disturbance variables (Table A1). Table A2 shows the thresholds that were used when assigning metric scores to each site. The metric scores were then considered in combination, using the 'combination rules' described below, and sites were assigned to one of seven preliminary disturbance categories, ranging from Best Reference to Highly Stressed. The preliminary designations were then reviewed by James Meek from MassDEP, who either confirmed or changed the designations. In addition, a new column was added to the table with designations collapsed into three broader disturbance categories (reference, stressed, intermediate), as folows:

- Reference
    - Western Highlands (WH) - Best Reference + Reference
    - Central Hills (CH) - Best Reference + Reference + Sub Reference
- Stressed
    - WH - High Stress + Stress
    - CH - High Stress

[1]Upstream watershed scale (Ws) includes the local catchment plus the accumulated area of all upstream catchments; local catchment scale (Cat) is defined as the landscape area draining to a single stream segment, excluding upstream contributions.

[2] https://www.epa.gov/waterdata/get-nhdplus-national-hydrography-dataset-plus-data#Download

[3] https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset-0

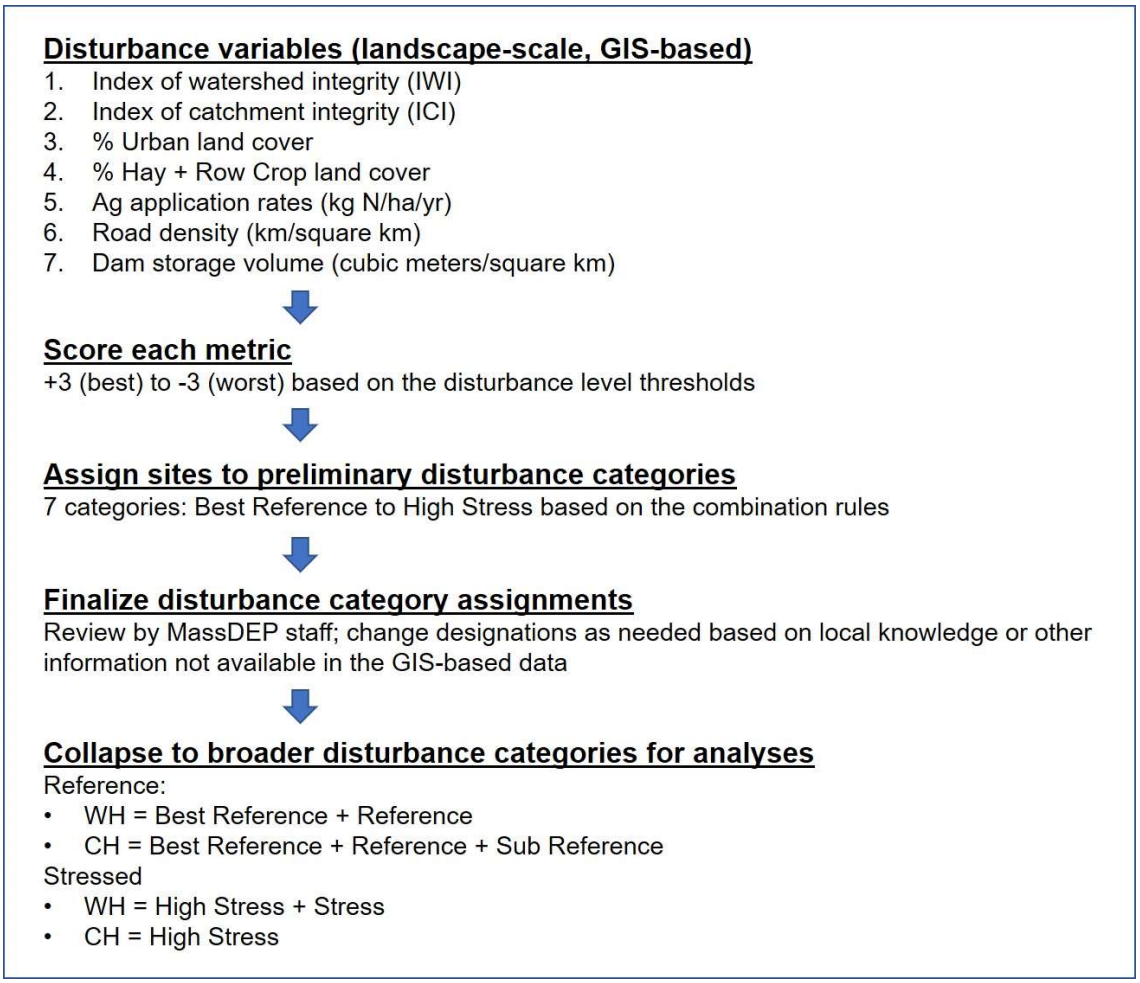


Figure A1. Process for assigning sites to disturbance categories. Information on variable selection and development of the disturbance gradient can be found in Jessup and Stamp (2020).

Table A1. Seven disturbance variables from the USEPA StreamCat dataset (Hill et al. 2016) were used to assign sites to preliminary disturbance categories. Entries were marked as '2019' vs. 'previous' to show where changes were made when the new sites (sampled in 2019) were assessed vs. sites that had been previously assessed during the calibration of the 100-count riffle habitat IBI (Jessup and Stamp 2020).

Disturbance variable	Abbrev	Spatial scale	Source	Units	Description
Index of catchment integrity	ICI	Local catchment (Cat)	2019: version 2.1[a] Previous: version 1	0 (worst) -1 (best)	Measure of overall watershed condition, based on six components: hydrologic regulation, regulation of water chemistry, sediment regulation, hydrologic connectivity, temperature regulation, and habitat provision
Index of watershed integrity	IWI	Upstream watershed (Ws)			
% Urban land cover	PctUrbLMH	2019: maximum value across both scales (Cat, Ws) Previous: Ws only	2019: NLCD 2016[b]; Previous: NLCD 2011	percent (0-100)	Percent of area classified as developed, high + medium + low-intensity land use (NLCD classes 24+23+22)
Road density	RdDens	2019: maximum value across both scales (Cat, Ws) Previous: Cat only	Road layer = 2010 Census Tiger Lines	km/km^2	Density of roads within area
% Agricultural land cover	PctHayCrop	2019: maximum value across both scales (Cat, Ws) Previous: Cat only	2019: NLCD 2016[b]; Previous: NLCD 2011	percent (0-100)	Percent of area classified as hay and crop land use (NLCD classes 82+81)
Mean rate of fertilizer application + biological nitrogen fixation + manure application	[CBNF]+[Fert] +[Manure]	2019: maximum value across both scales (Cat, Ws) Previous: Ws only	EnviroAtlas	mean rate kg N/ ha/yr	[Mean rate of biological nitrogen fixation from the cultivation of crops (CBNF)] + [Mean rate of synthetic nitrogen fertilizer application to agricultural land within area (Fert)] + [Mean rate of manure application to agricultural land from confined animal feeding operations within area (Manure)]
Dam storage volume	DamNrmStor	2019: maximum value across both scales (Cat, Ws) Previous: Ws only	Army Corps of Engineers (ACOE)	m^3/km^2	Volume all reservoirs per unit area. Based on typical volumes stored within reservoirs (NORM_STORA in NID)

[a]version 1 - Thornbrugh et al. 2018; version 2.1 - Johnson et al. 2019
[b]NLCD data download - https://www.mrlc.gov/data

Table A2. Metric scoring thresholds for the 2019 sites. These are the same thresholds that were used during calibration of the 100-count riffle habitat IBI (Jessup and Stamp 2020) except for the IWI and ICI, which were adjusted to account for changes in the distribution of scores in version 2.1 compared to version 1.

Metric Scores	IWI (2.1)	ICI (2.1)	% Urban	% Hay/Crop	Fertilizer application	Road density	Dam storage volume
+3 (least disturbed)	≥ 0.85	≥ 0.85	≤ 1	≤ 1	≤ 0.5	≤ 1.5	≤ 0.1
+2	< 0.85 and ≥ 0.80	< 0.85 and ≥ 0.80	> 1 and ≤ 2	> 1 and ≤ 2	> 0.5 and ≤ 1	> 1.5 and ≤ 2	> 0.1 and ≤1,000
1	< 0.80 and ≥ 0.70	< 0.80 and ≥ 0.70	> 2 and ≤ 5	> 2 and ≤ 5	> 1 and ≤ 2.5	> 2 and ≤ 3	> 1000 and ≤ 10,000
0	< 0.70 and > 0.60	< 0.70 and > 0.60	> 5 and < 10	> 5 and < 10	> 2.5 and < 5	> 3 and < 5	> 10,000 and < 50,000
-1	≤ 0.60 and > 0.50	≤ 0.60 and > 0.50	≥ 10 and < 40	≥ 10 and < 15	≥ 5 and < 7.5	≥ 5 and < 7.5	≥ 50,000 and < 100,000
-2	≤ 0.50 and > 0.40	≤ 0.50 and > 0.40	≥ 40 and < 60	≥ 15 and < 20	≥ 7.5 and < 10	≥ 7.5 and < 10	≥ 100,000 and < 200,000
-3 (most disturbed)	≤ 0.40	≤ 0.40	≥ 60	≥ 20	≥ 10	≥ 10	≥ 200,000

The following metric 'combination rules' were used to assign sites to one of seven preliminary disturbance categories, ranging from Best Reference to Highly Stressed:

- **Best Reference** –
  - All metrics meet the +2 scoring thresholds

- **Reference** –
  - All metrics meet the +1 scoring thresholds

- **Sub Reference –**
  - All metrics meet the 0 scoring thresholds and at least five metrics receive positive scores (> 0)

- **Intermediate –**
  - All metrics meet the 0 scoring thresholds and no more than four metrics receive positive scores

- **Some Stress** –
  - One or two metrics receive a score of -1 and the rest (at least five) receive positive scores or scores of 0; ***OR***
  - One metric receives a score of -2, another receives a score of -1, and the rest receive scores of 0 or higher

- **Stressed** –
  - Three or more metrics receive scores of -1 or -2; ***OR***
  - At least one metric receives a score of -3, and no more than three metrics receive negative scores

- **High Stress** -
  - At least one metric receives a score of -3, and at least four other metrics receive negative scores

Literature cited

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., 2016. The Stream-Catchment (StreamCat) dataset: a database of watershed metrics for the Conterminous United States. J. Am. Water Res. Assoc. 52 (1), 120–128

Jessup, B., and J. Stamp. 2020. Development of Indices of Biotic Integrity for Assessing Macroinvertebrate Assemblages in Massachusetts Freshwater Wadeable Streams. Prepared for the Massachusetts Department of Environmental Protection.

Johnson, Zachary & G. Leibowitz, Scott & Hill, Ryan. (2018). Revising the index of watershed integrity national maps. Science of The Total Environment. 10.1016/j.scitotenv.2018.10.112.

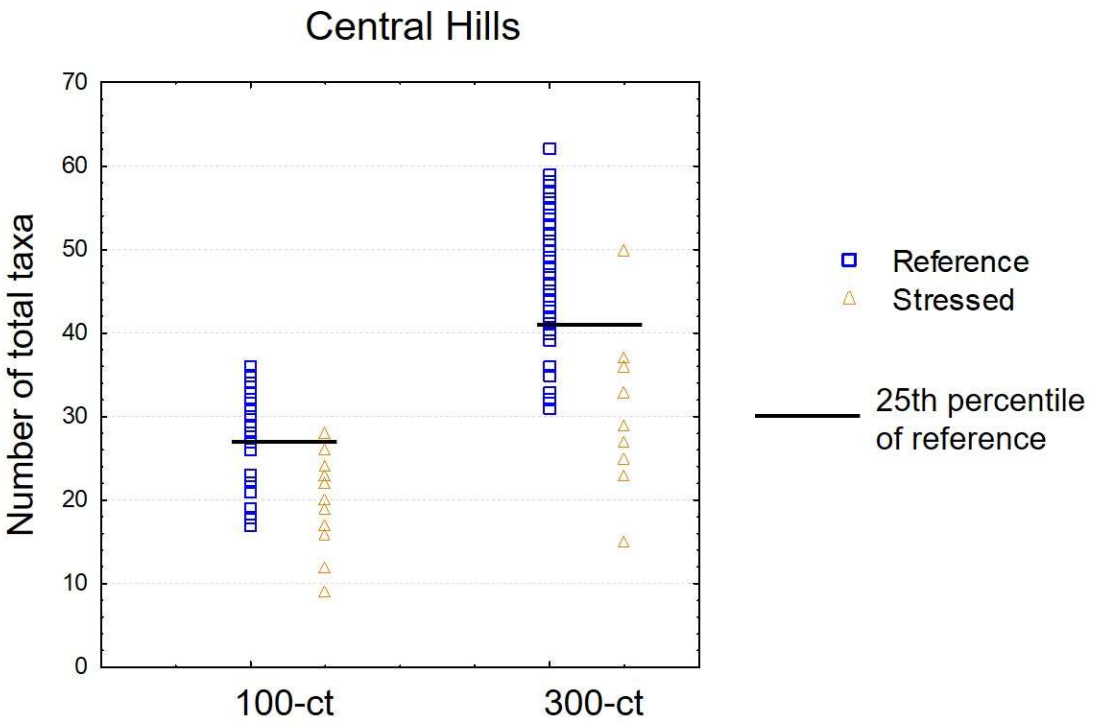
McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., Reah, A., 2012. NHDPlus Version 2: User Guide. U.S. Environmental Protection Agency

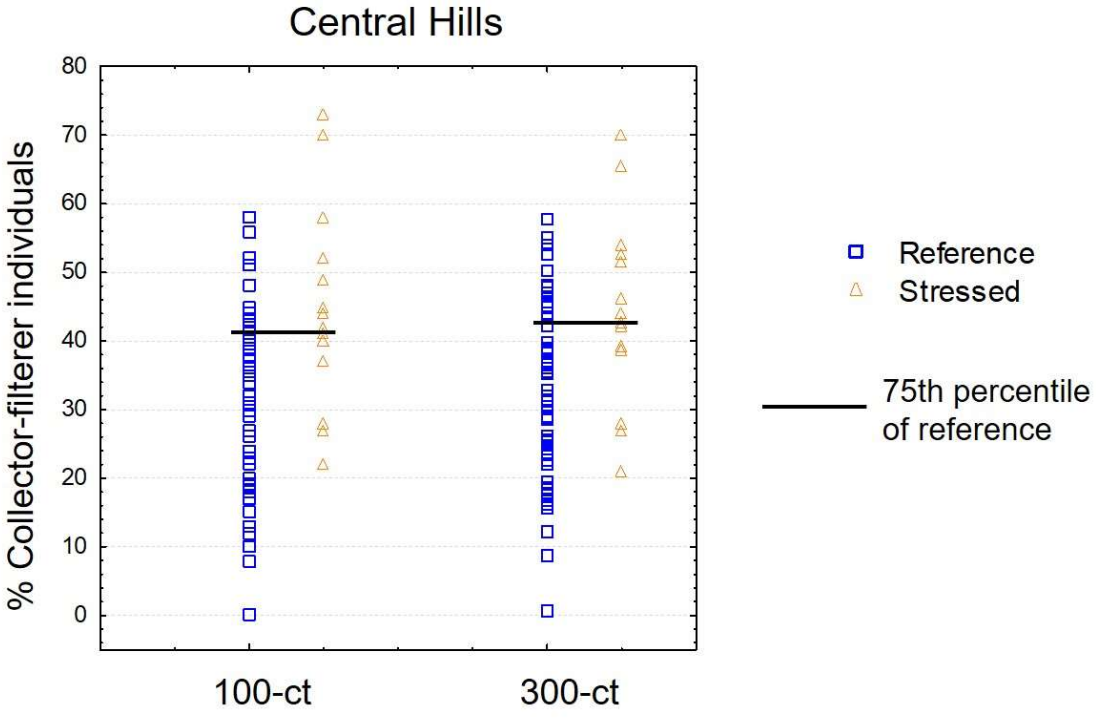
Thornbrugh, D. J., Leibowitz, S.G., Hill, R. A., Weber, M. H., Johnson, Z.C. Olsen, A. R., Flotemersch, J. E., Stoddard, J. L., & Peck, D. V. 2018. Mapping watershed integrity for the conterminous United States. Ecological Indicators, 85, 1133-1148

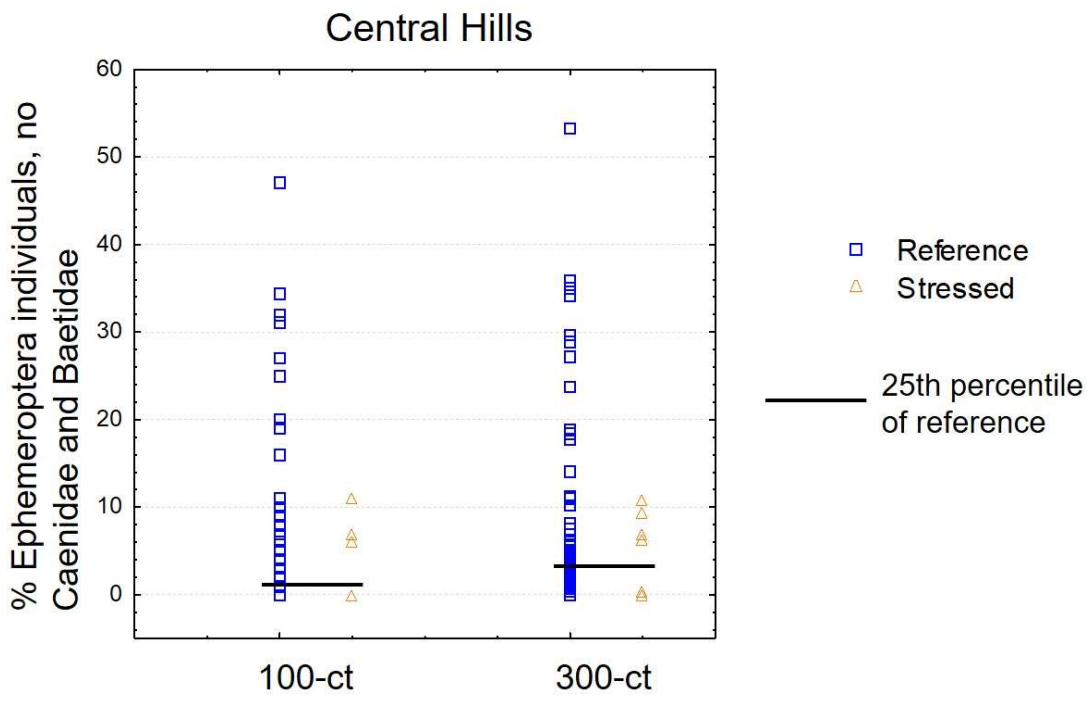
Appendix B

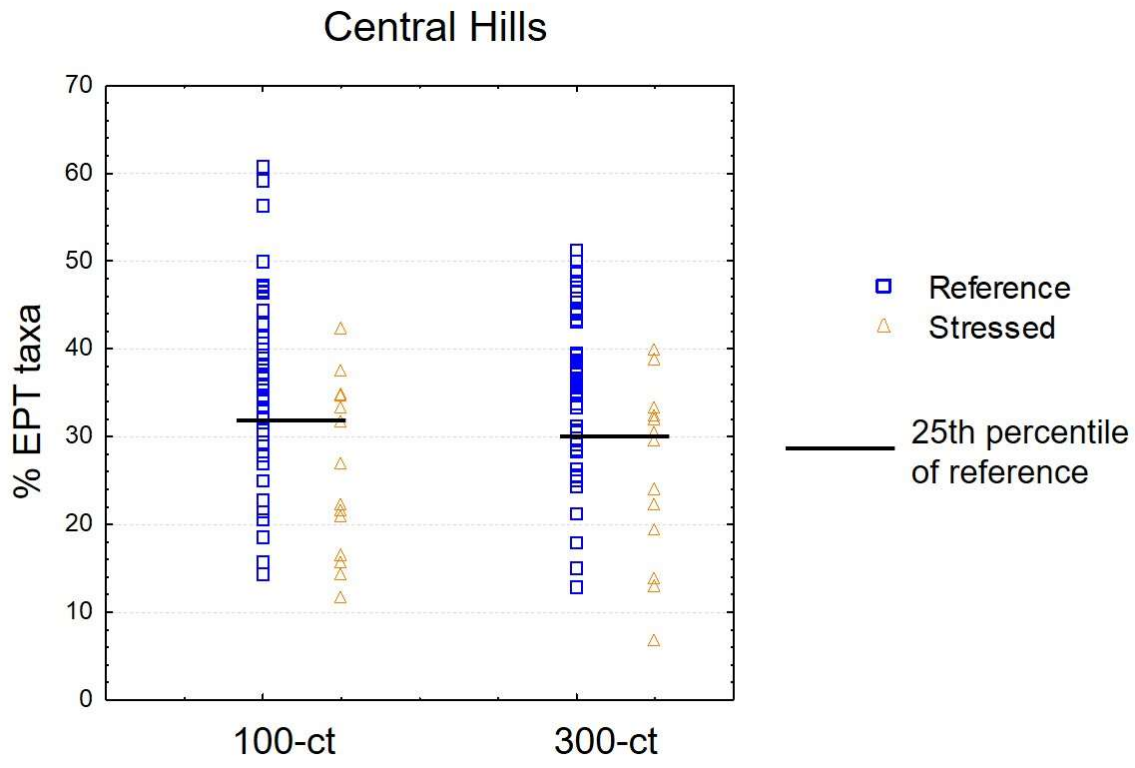
Metric performance plots - 100 vs. 300-count samples

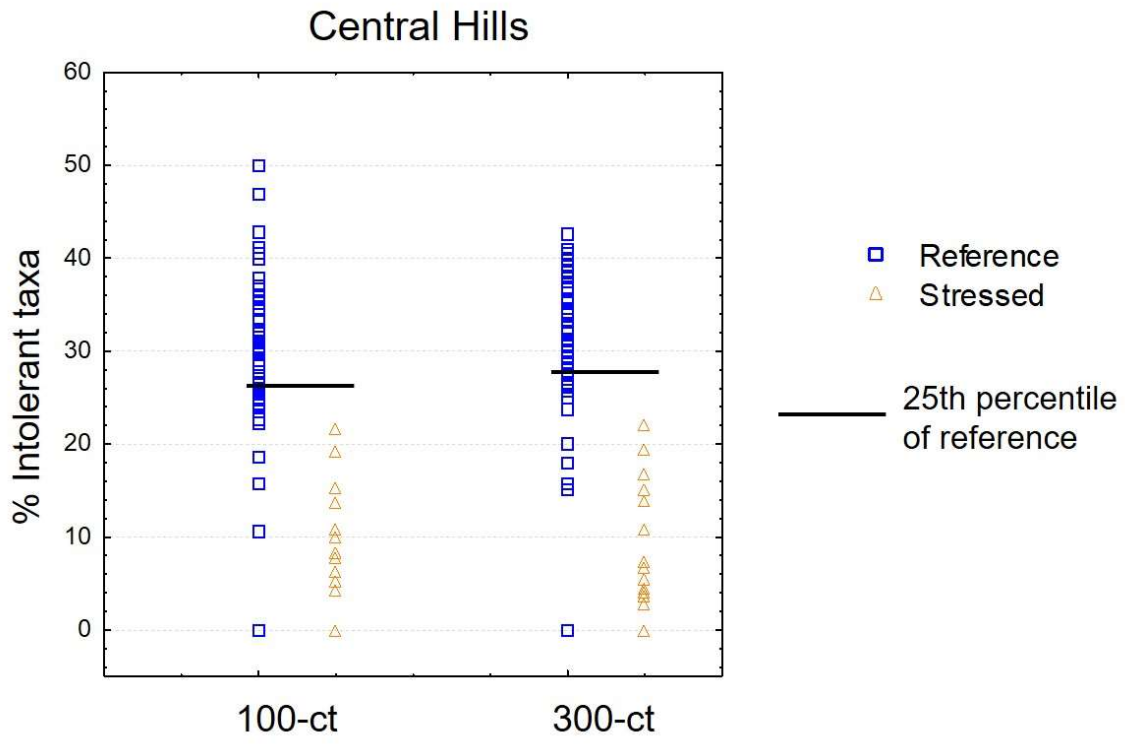
These plots show distributions of metric values in reference vs. stressed samples in the Central Hills and Western Highlands. Discrimination efficiency (DE) scores are derived based on the number of stressed samples that have values $\geq 25^{th}$ percentile of reference for 'decreaser' metrics (which decrease as stress increases) and the number of stressed samples that have values $\leq 75^{th}$ percentile of reference for the 'increaser' metrics. There is only one 'increaser' metric in the IBIs (percent collector-filterer individuals).

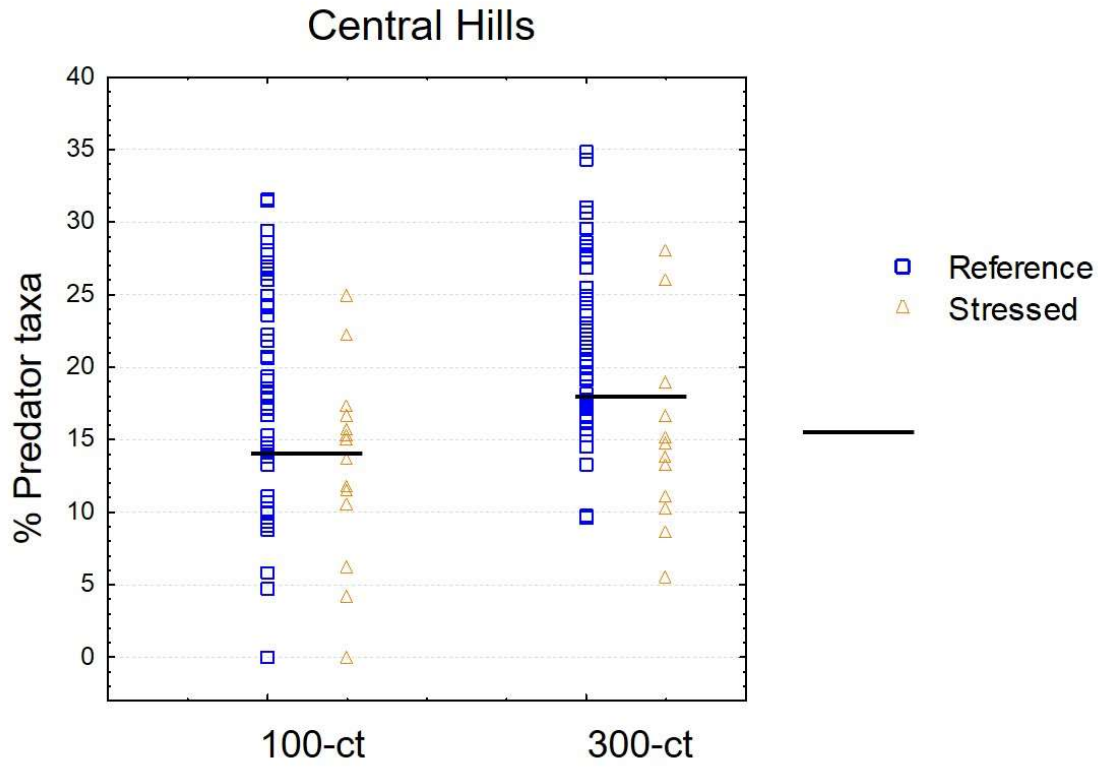

Central Hills
Number of total taxa
70
60
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
25th percentile
of reference

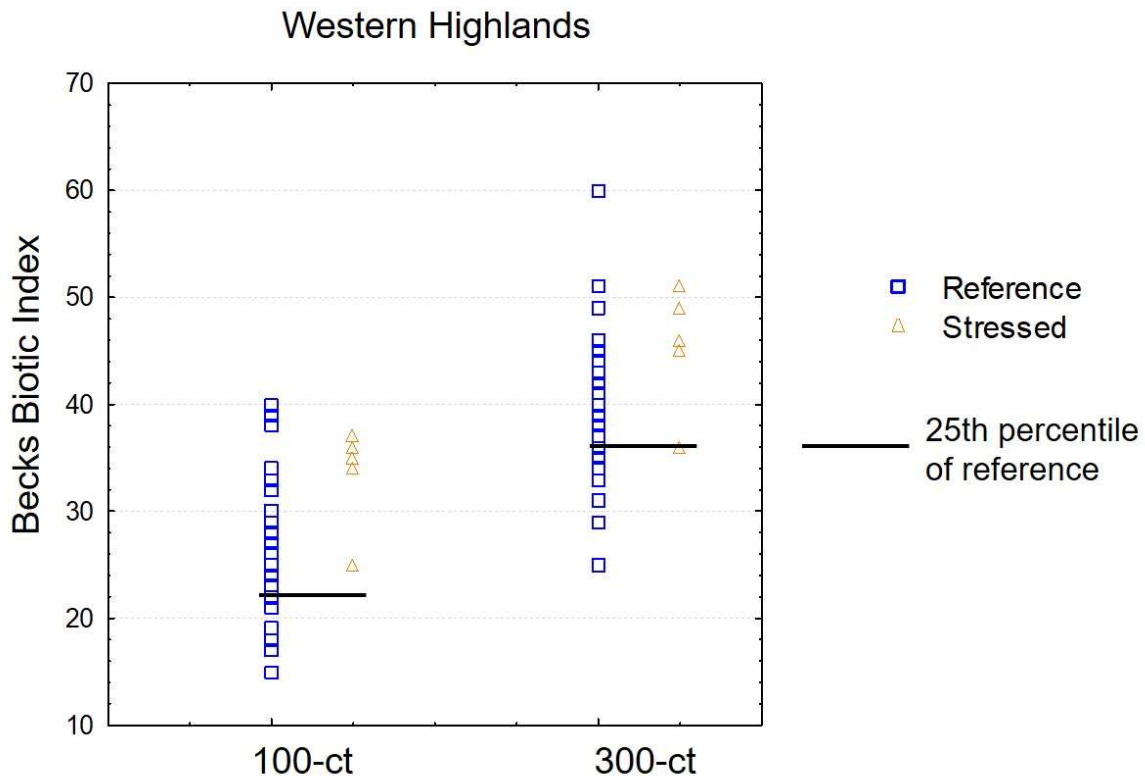

Central Hills
% Collector-filterer individuals
80
70
60
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
75th percentile
of reference

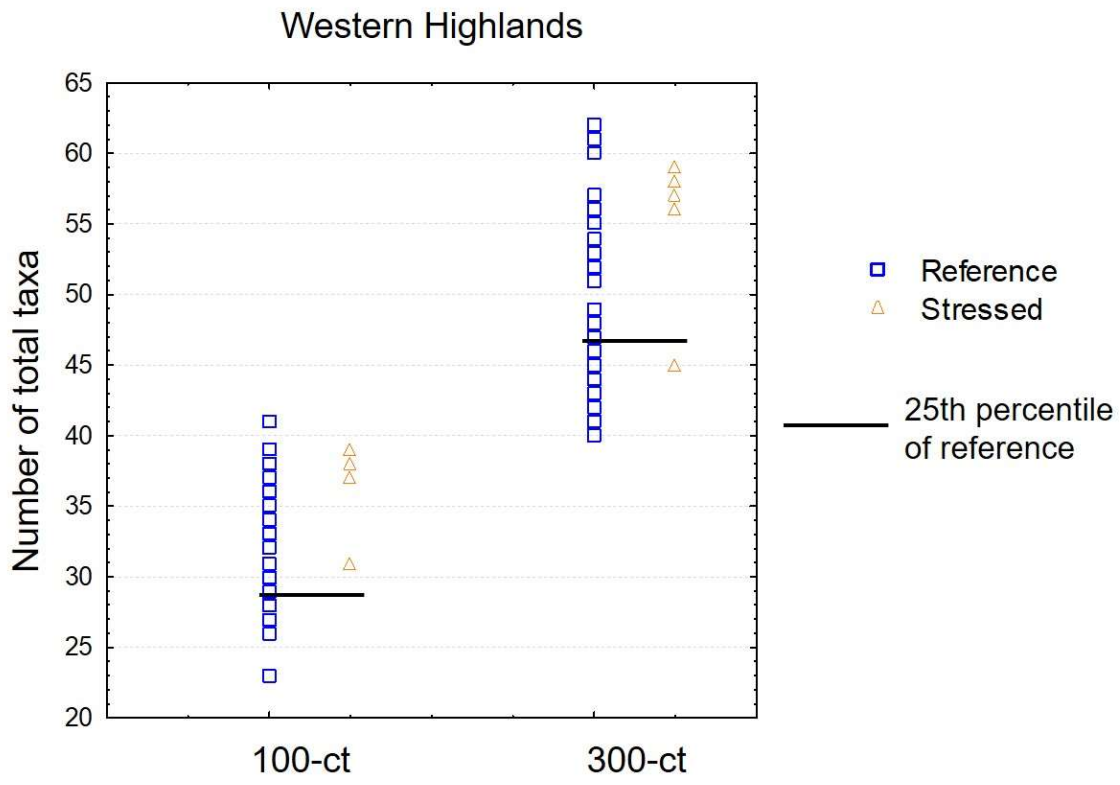

Central Hills
% Ephemeroptera individuals, no Caenidae and Baetidae
60
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
25th percentile of reference

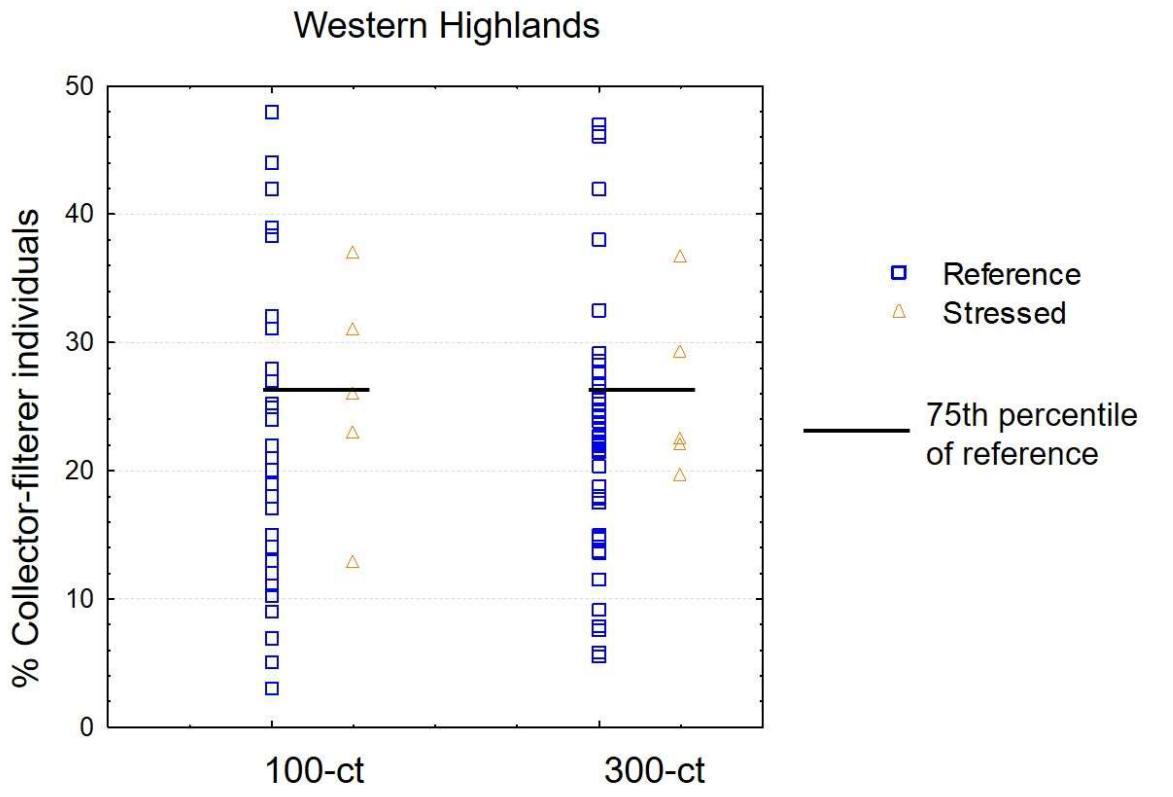

Central Hills
% EPT taxa
70
60
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
25th percentile of reference

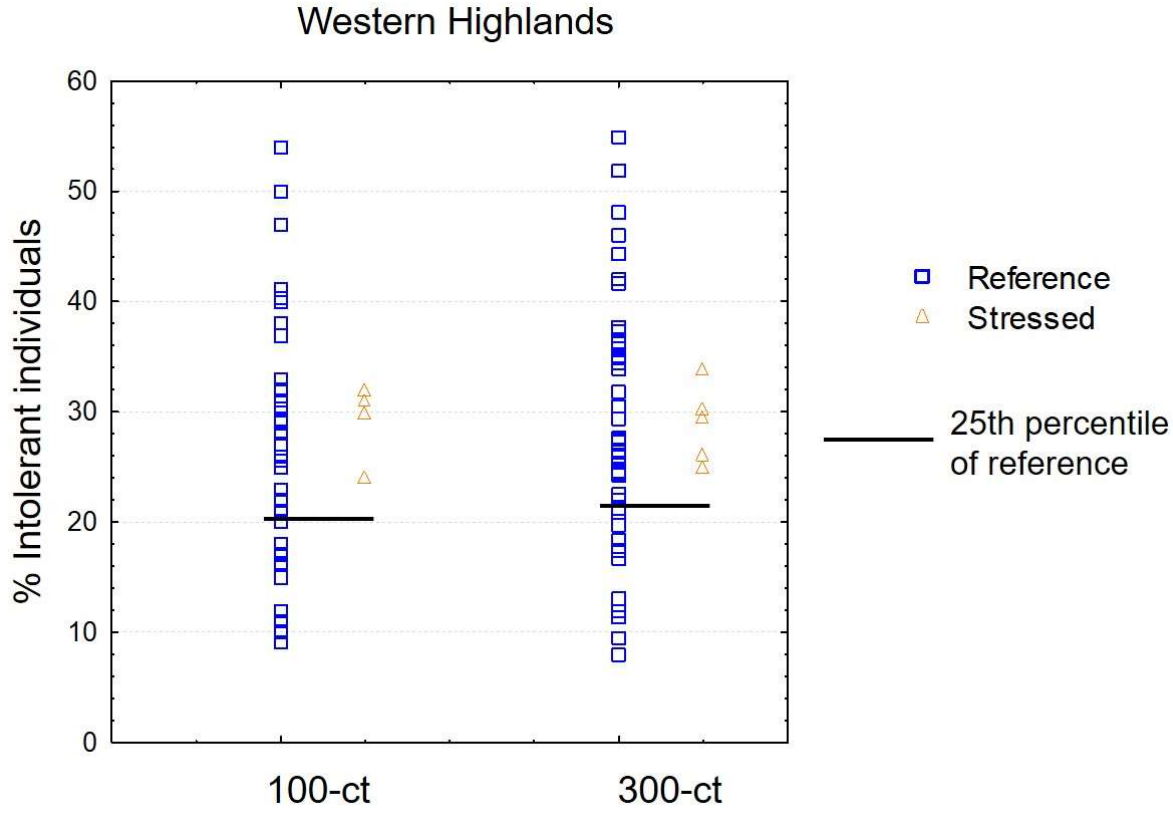

Central Hills
% Intolerant taxa
60
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
25th percentile
of reference

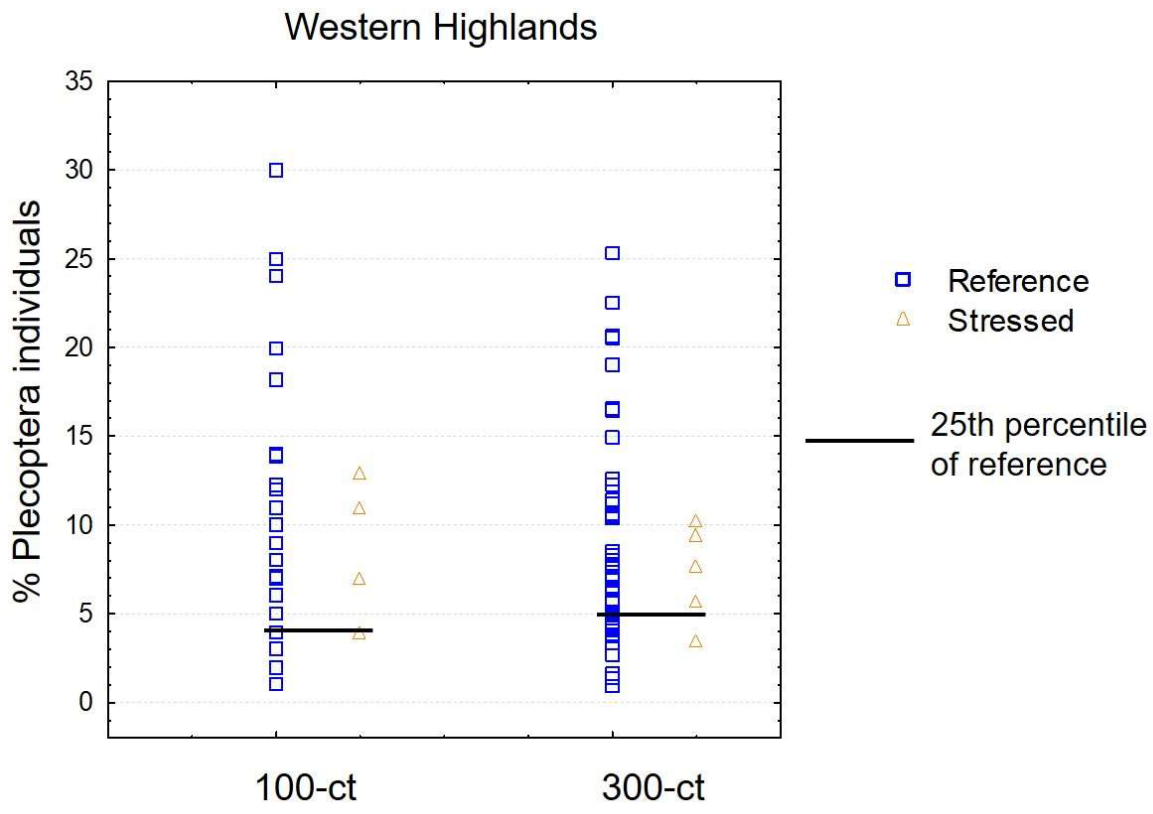

Central Hills
% Predator taxa
40
35
30
25
20
15
10
5
0
100-ct
300-ct
Reference
Stressed

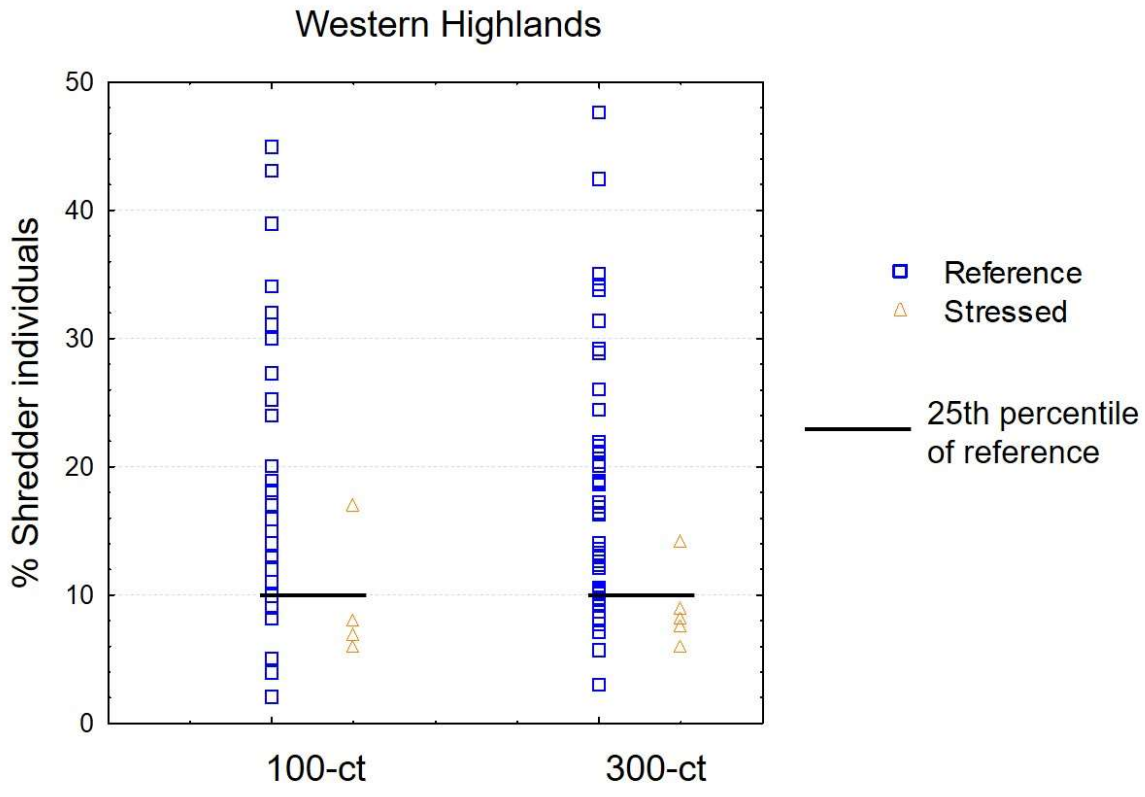

Western Highlands
Becks Biotic Index
70
60
50
40
30
20
10
100-ct
300-ct
Reference
Stressed
25th percentile
of reference


Western Highlands
Number of total taxa
65
60
55
50
45
40
35
30
25
20
100-ct
300-ct
Reference
Stressed
25th percentile
of reference


Western Highlands
% Collector-filterer individuals
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
75th percentile
of reference


Western Highlands
% Intolerant individuals
60
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
25th percentile
of reference


Western Highlands
% Plecoptera individuals
35
30
25
20
15
10
5
0
100-ct
300-ct
Reference
Stressed
25th percentile
of reference

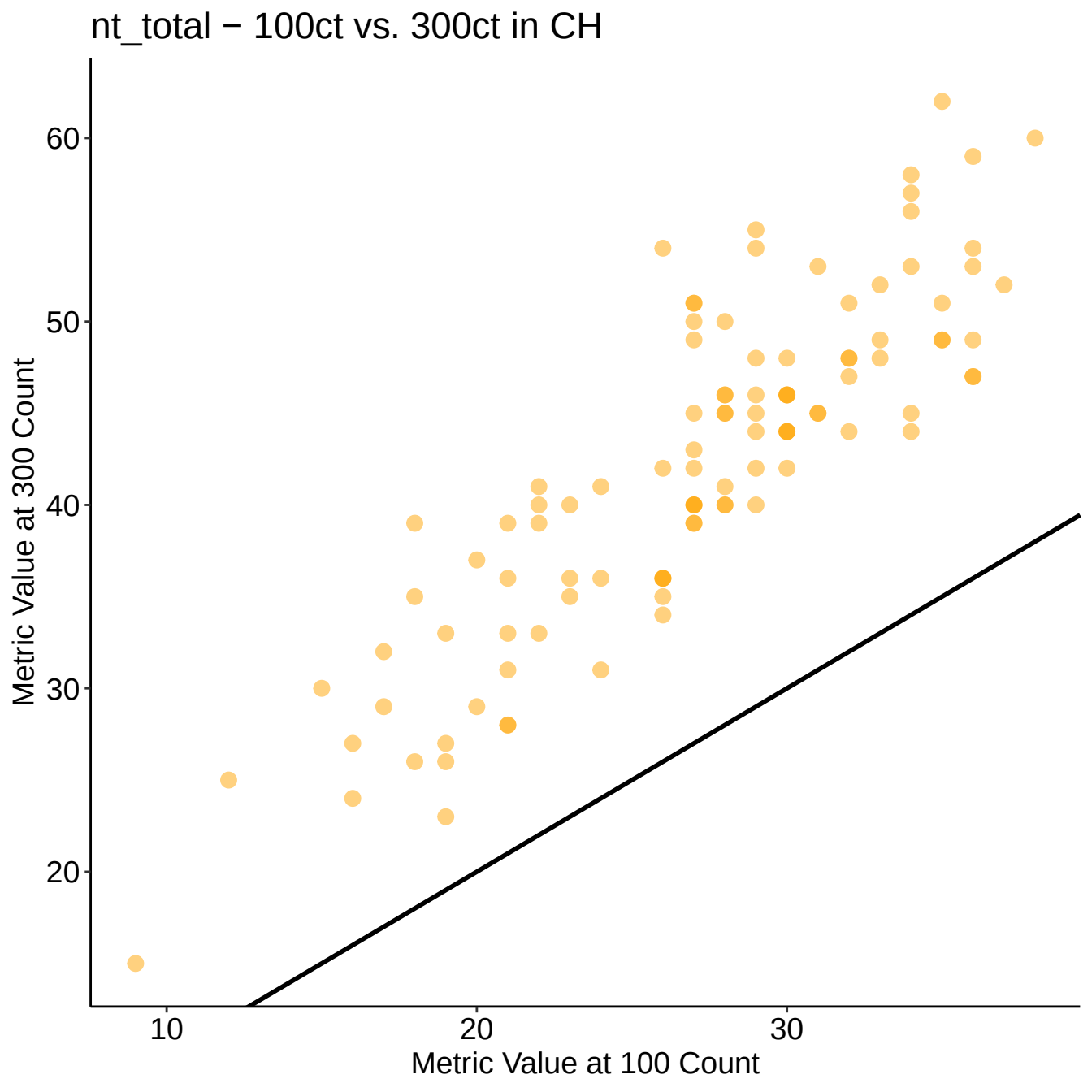

Western Highlands
% Shredder individuals
50
40
30
20
10
0
100-ct
300-ct
Reference
Stressed
25th percentile
of reference

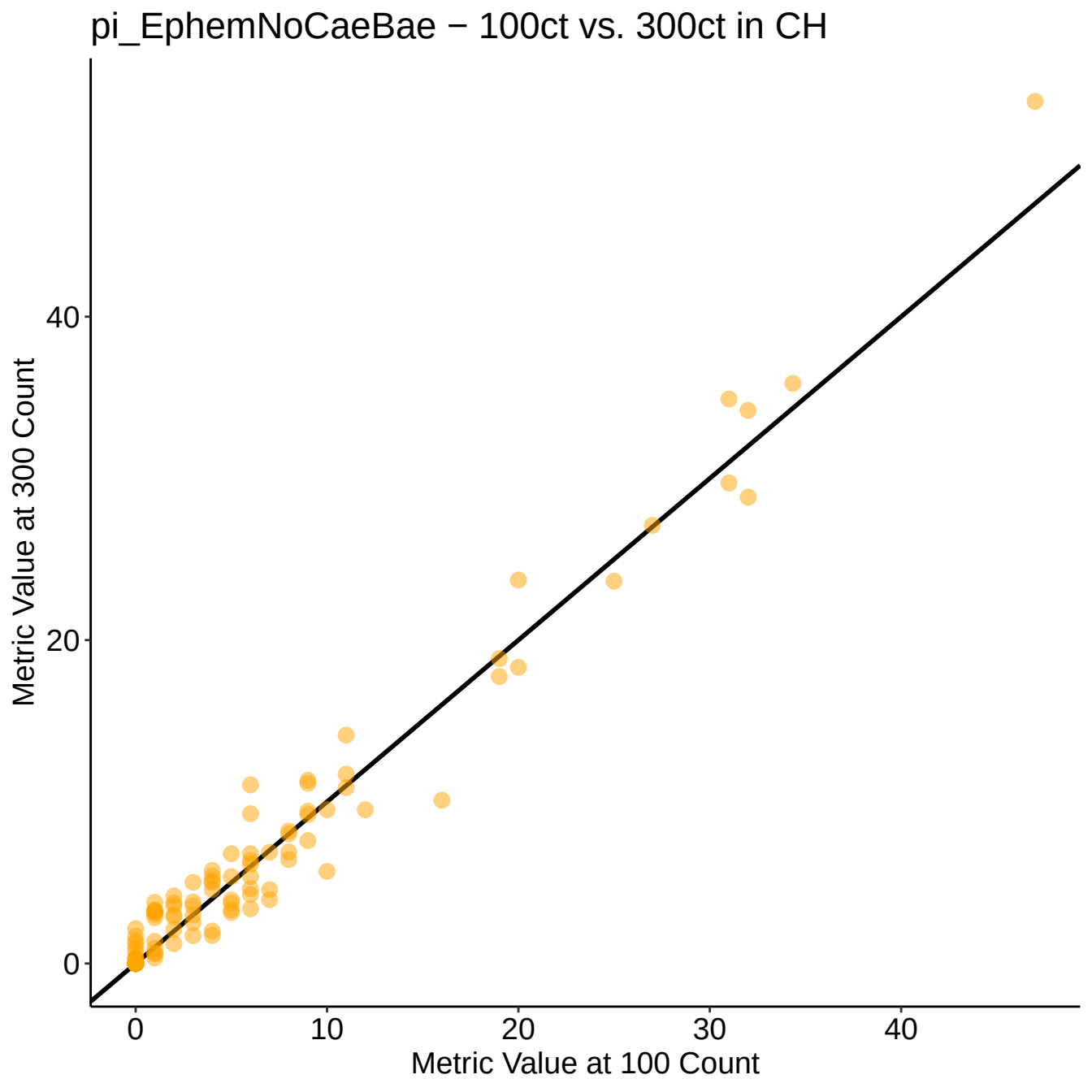
Appendix C

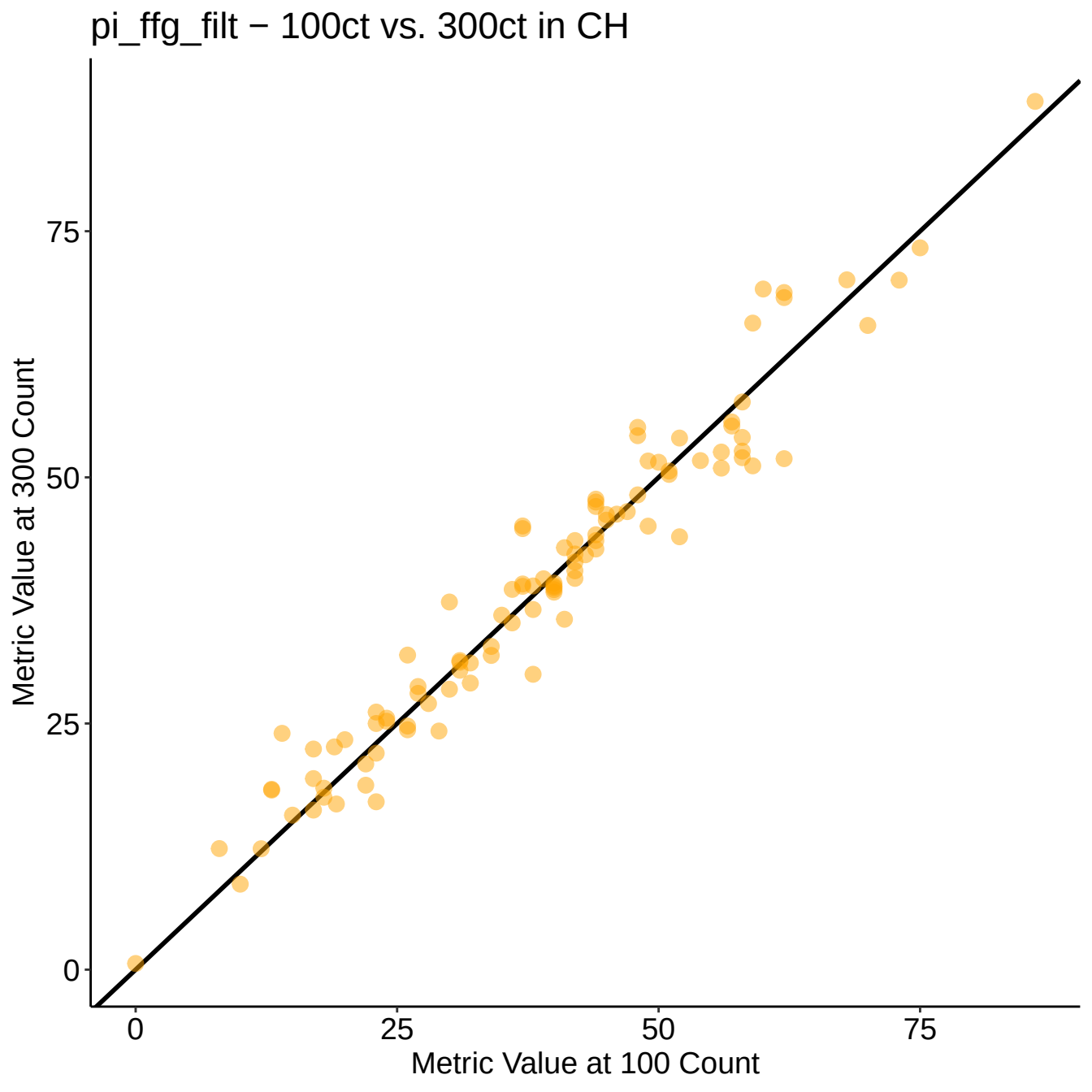
Comparison scatterplots of metric values in the Central Hills and Western Highlands. Metric values were calculated for 100- and 300-count paired samples. The black line indicates a slope of 1 which would mean that metric values would be equal between subsample sizes on a sample-by-sample basis.

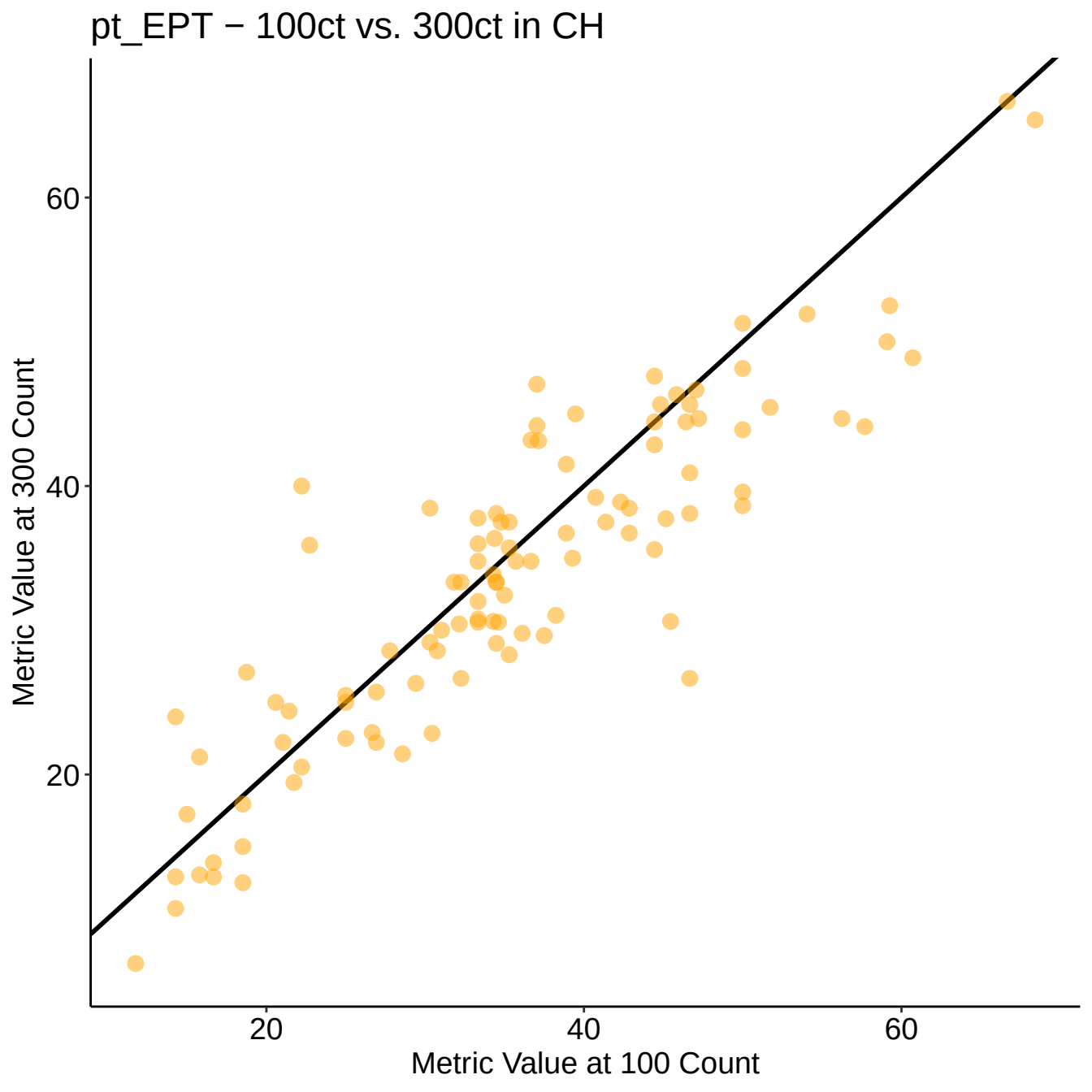
Table 1. Input metrics in the Central Hills and Western Highlands IBIs. DE = discrimination efficiency. Trend is the direction of metric response with increasing stress.

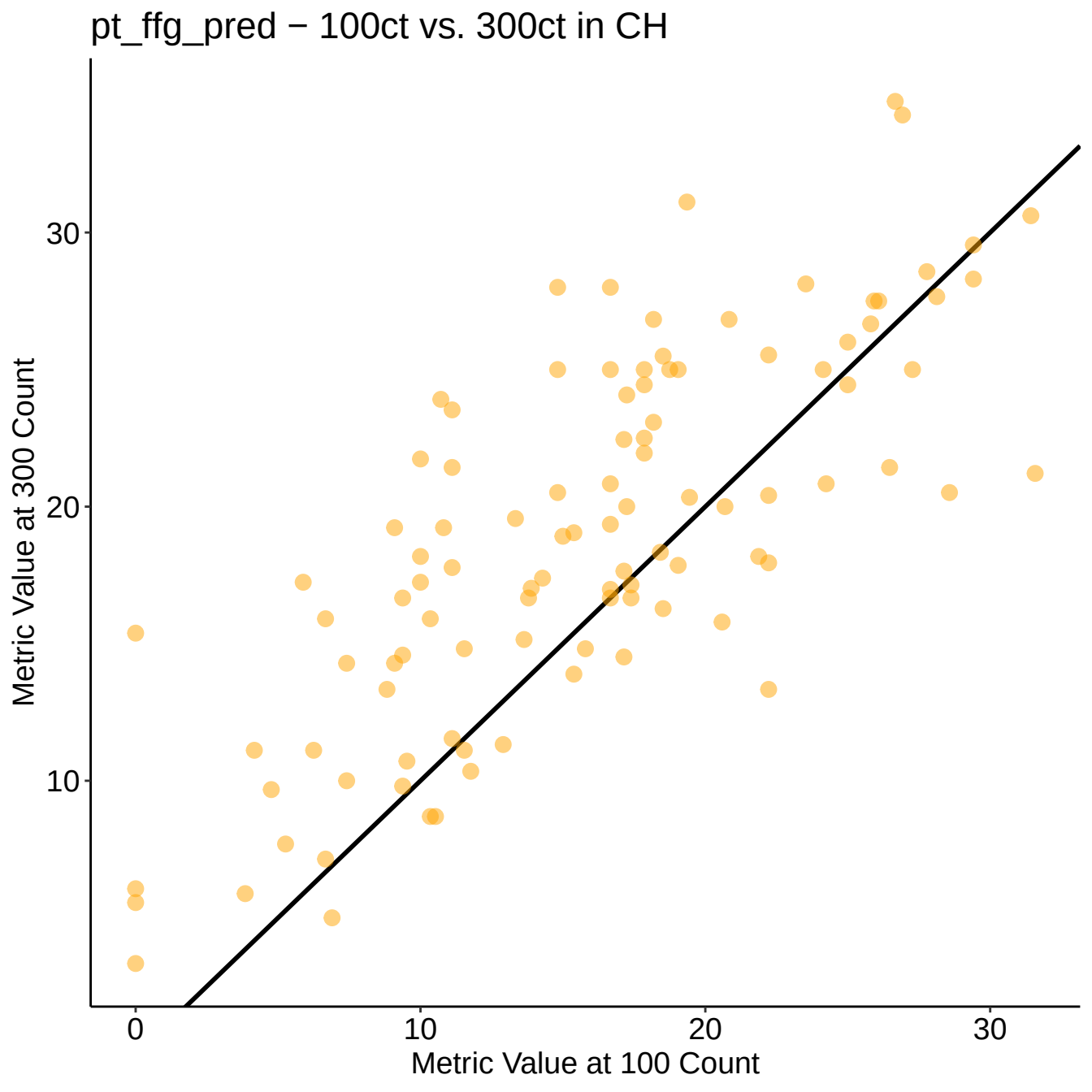
Western Highlands kick net IBI	
Metric (abbrev)	**Response to stress**
Number of total taxa (nt_total)	Decrease
% Plecoptera individuals (pi_Pleco)	Decrease
% Collector-filterer individuals (pi_ffg_filt)	Increase
% Shredder individuals (pi_ffg_shred)	Decrease
% Intolerant individuals (pi_tv_intol)	Decrease
Becks Biotic Index (x_Becks)	Decrease
Central Hills kick net IBI	
Metric (abbrev)	**Response to stress**
Number of total taxa (nt_total)	Decrease
% EPT taxa (pt_EPT)	Decrease
% Ephemeroptera individuals, excluding Caenidae and Baetidae (pi_Ephem NoCaeBae)	Decrease
% Collector-filterer individuals (pi_ffg_filt)	Increase
% Predator taxa (pt_ffg_pred)	Decrease
% Intolerant taxa (pt_tv_intol)	Decrease

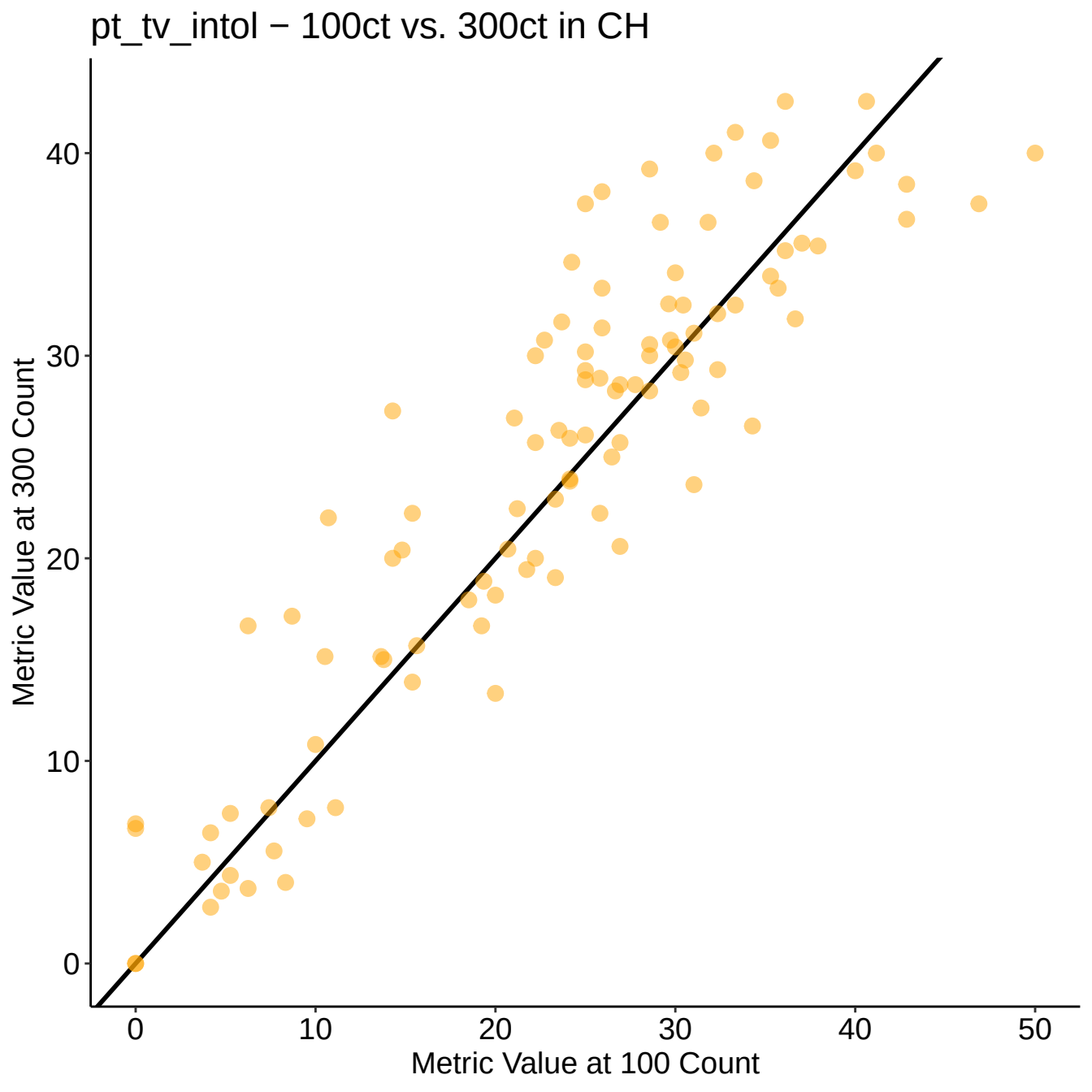

nt_total − 100ct vs. 300ct in CH
Metric Value at 300 Count
60
50
40
30
20
10
20
30
Metric Value at 100 Count

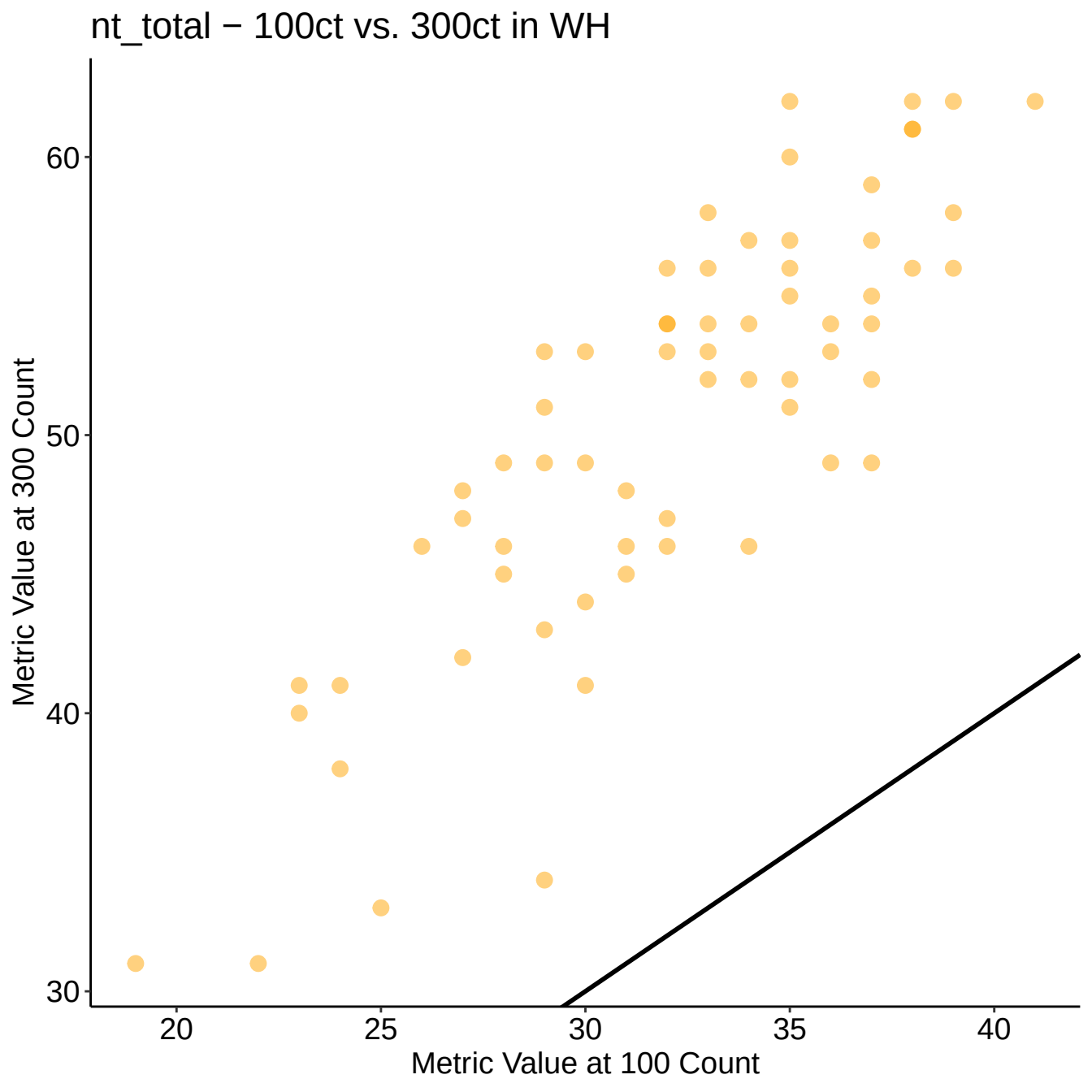

pi_EphemNoCaeBae − 100ct vs. 300ct in CH
Metric Value at 300 Count
0
20
40
0
10
20
30
40
Metric Value at 100 Count

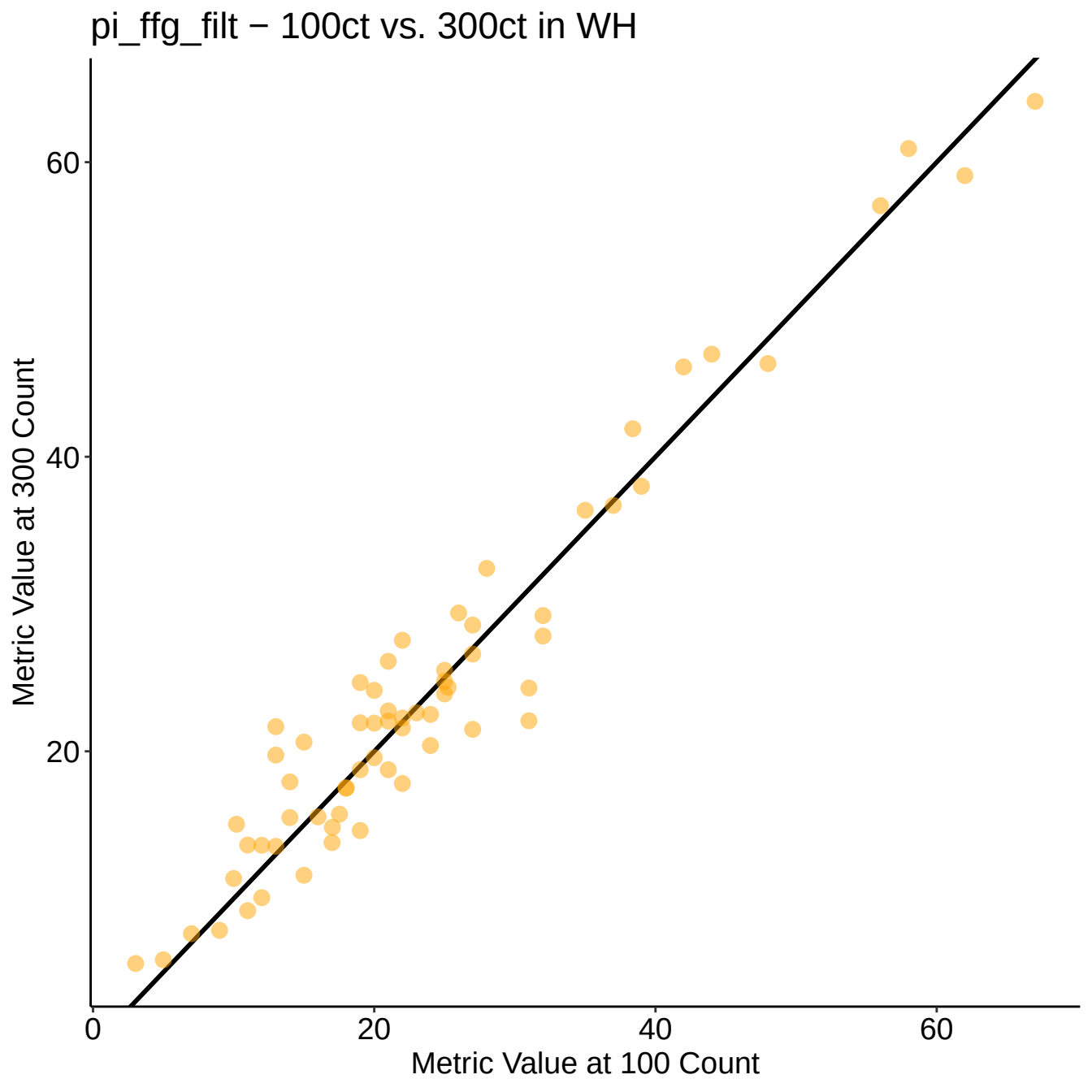

pi_ffg_filt − 100ct vs. 300ct in CH
Metric Value at 300 Count
0
25
50
75
0
25
50
75
Metric Value at 100 Count

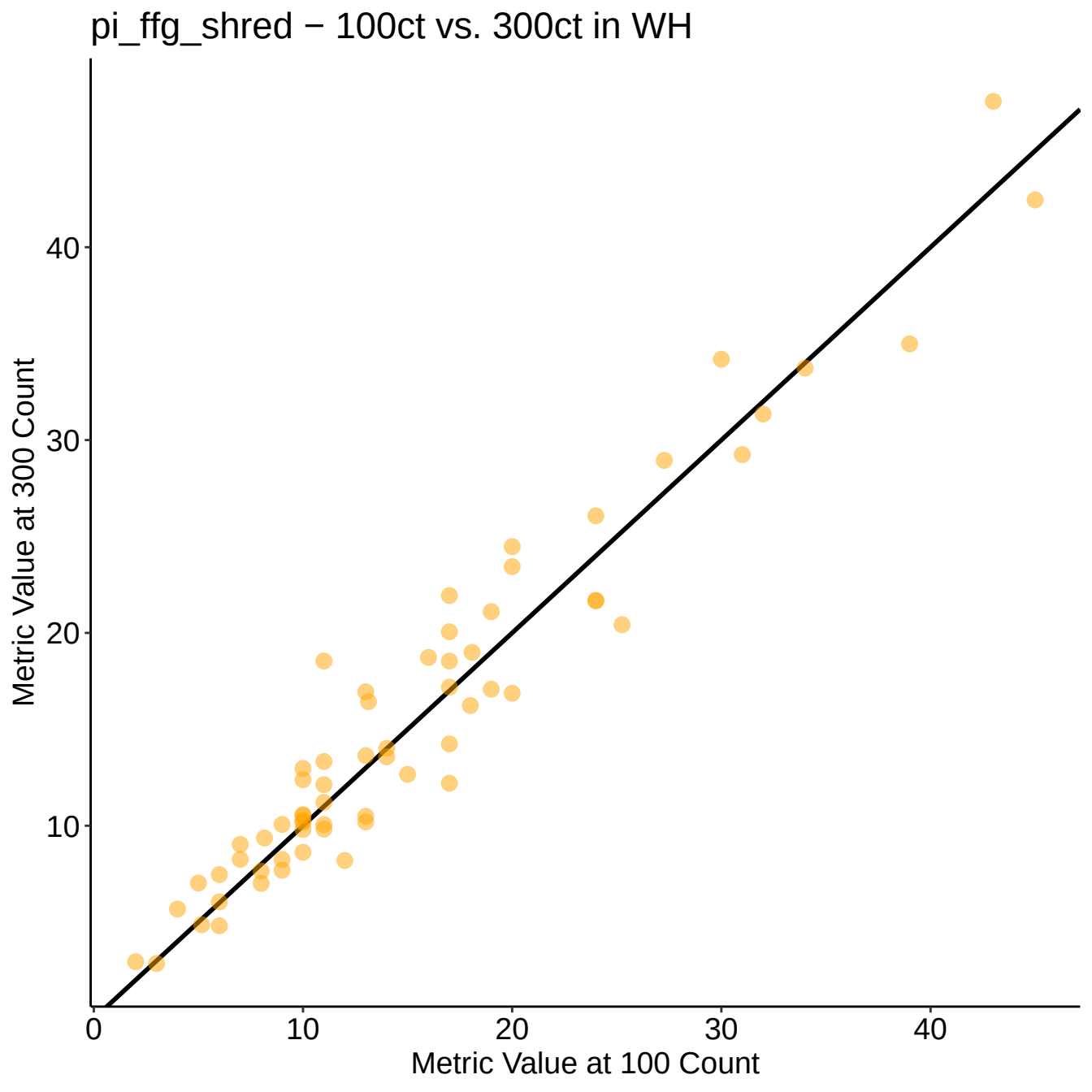

pt_EPT − 100ct vs. 300ct in CH
Metric Value at 300 Count
Metric Value at 100 Count
20
40
60
20
40
60

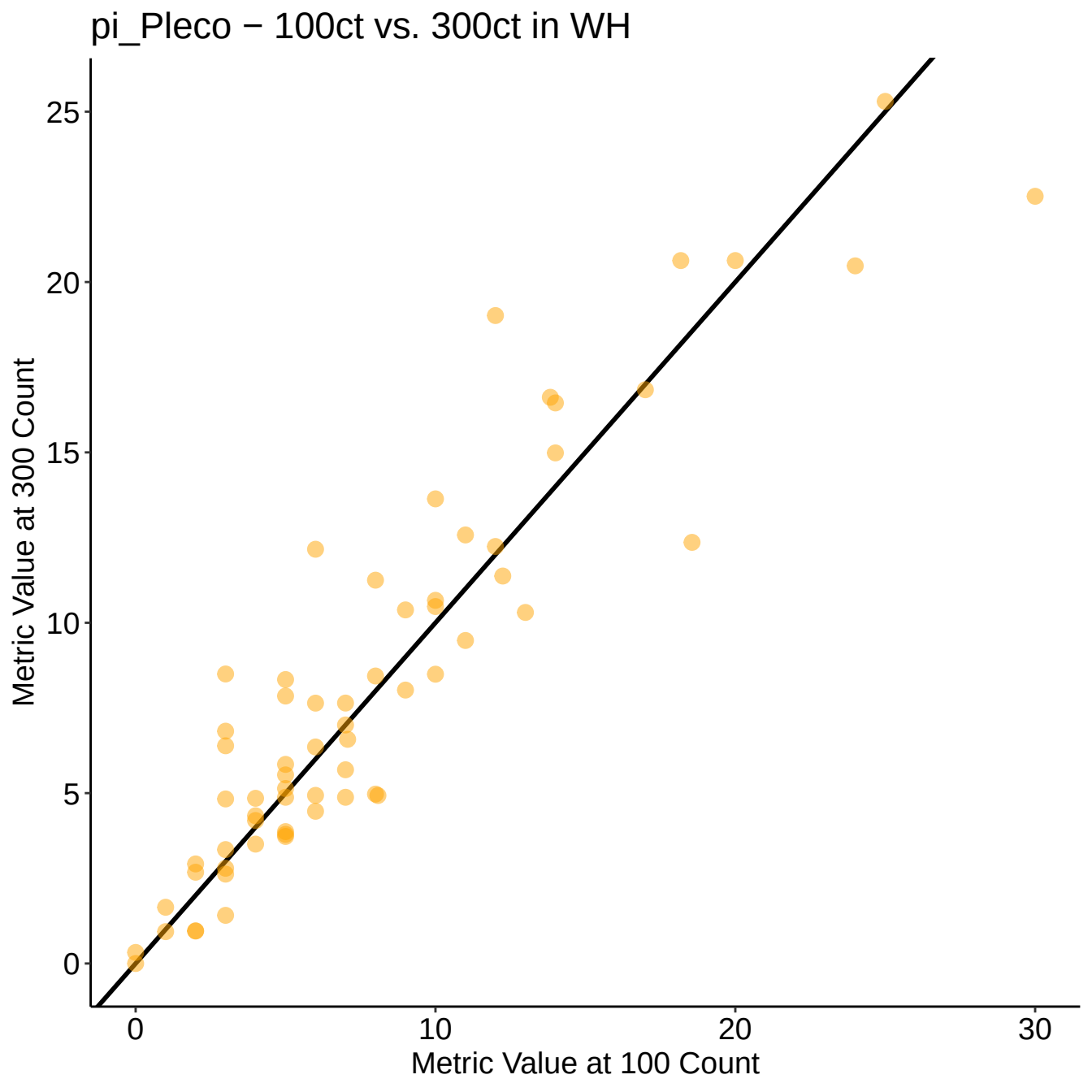

pt_ffg_pred − 100ct vs. 300ct in CH
Metric Value at 300 Count
30
20
10
0
10
20
30
Metric Value at 100 Count

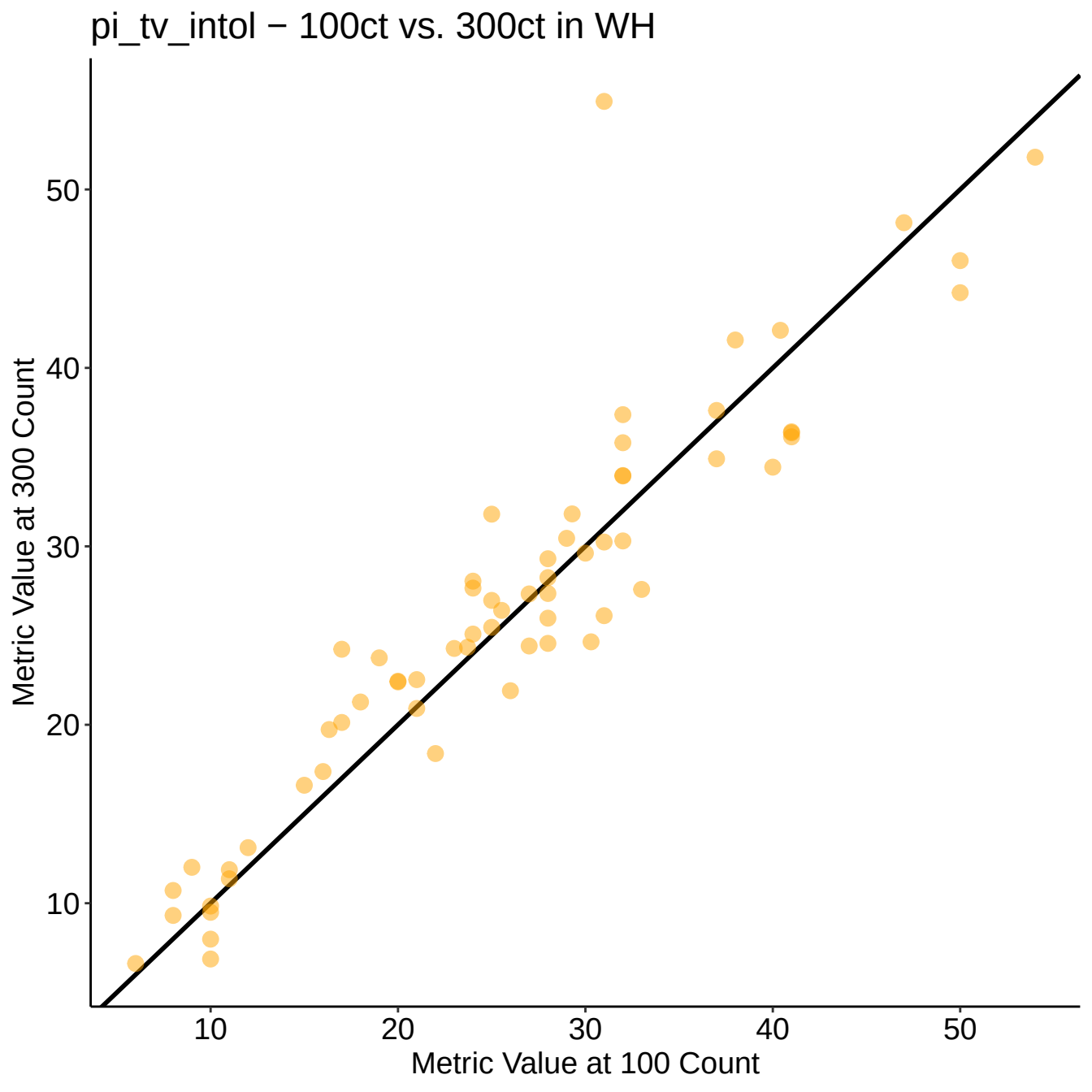

pt_tv_intol − 100ct vs. 300ct in CH
Metric Value at 300 Count
Metric Value at 100 Count
0
10
20
30
40
50

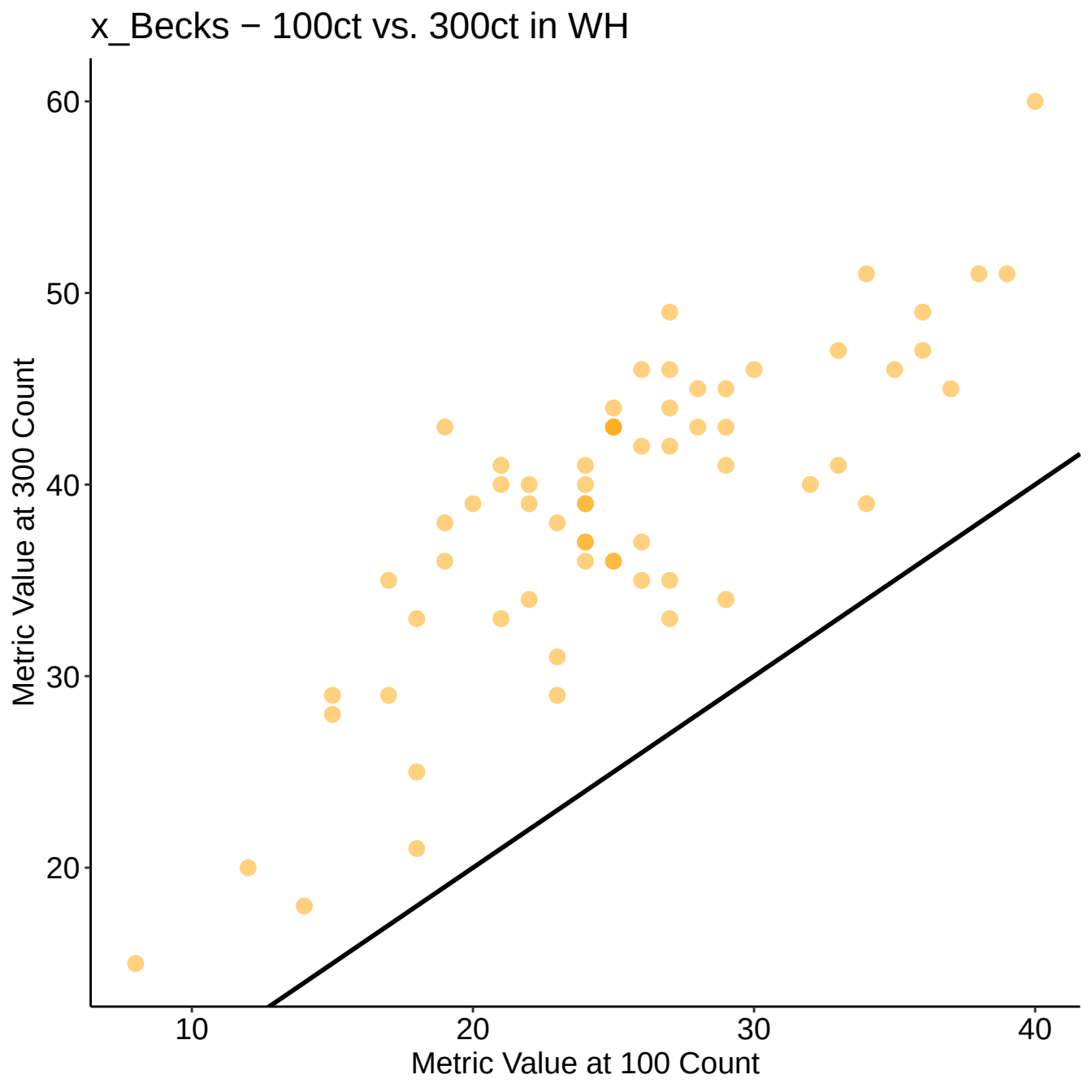

nt_total − 100ct vs. 300ct in WH
Metric Value at 300 Count
60
50
40
30
20
25
30
35
40
Metric Value at 100 Count


pi_ffg_filt − 100ct vs. 300ct in WH
Metric Value at 300 Count
Metric Value at 100 Count
0
20
40
60
20
40
60


pi_ffg_shred − 100ct vs. 300ct in WH
Metric Value at 300 Count
Metric Value at 100 Count
0
10
20
30
40
10
20
30
40


pi_Pleco − 100ct vs. 300ct in WH
Metric Value at 300 Count
Metric Value at 100 Count
0
5
10
15
20
25
0
10
20
30


pi_tv_intol − 100ct vs. 300ct in WH
Metric Value at 300 Count
Metric Value at 100 Count
10
20
30
40
50

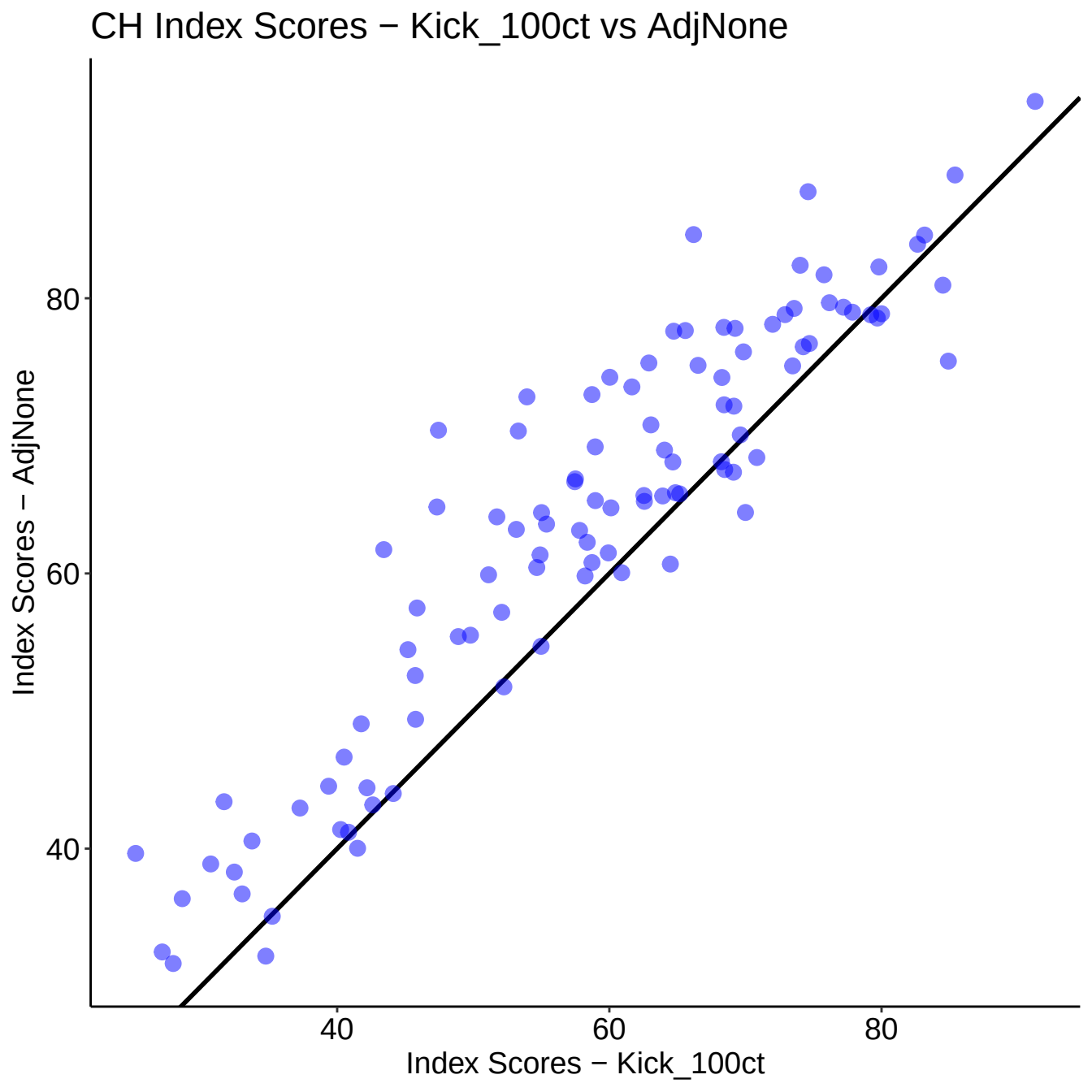

x_Becks − 100ct vs. 300ct in WH
Metric Value at 300 Count
Metric Value at 100 Count
60
50
40
30
20
10
20
30
40

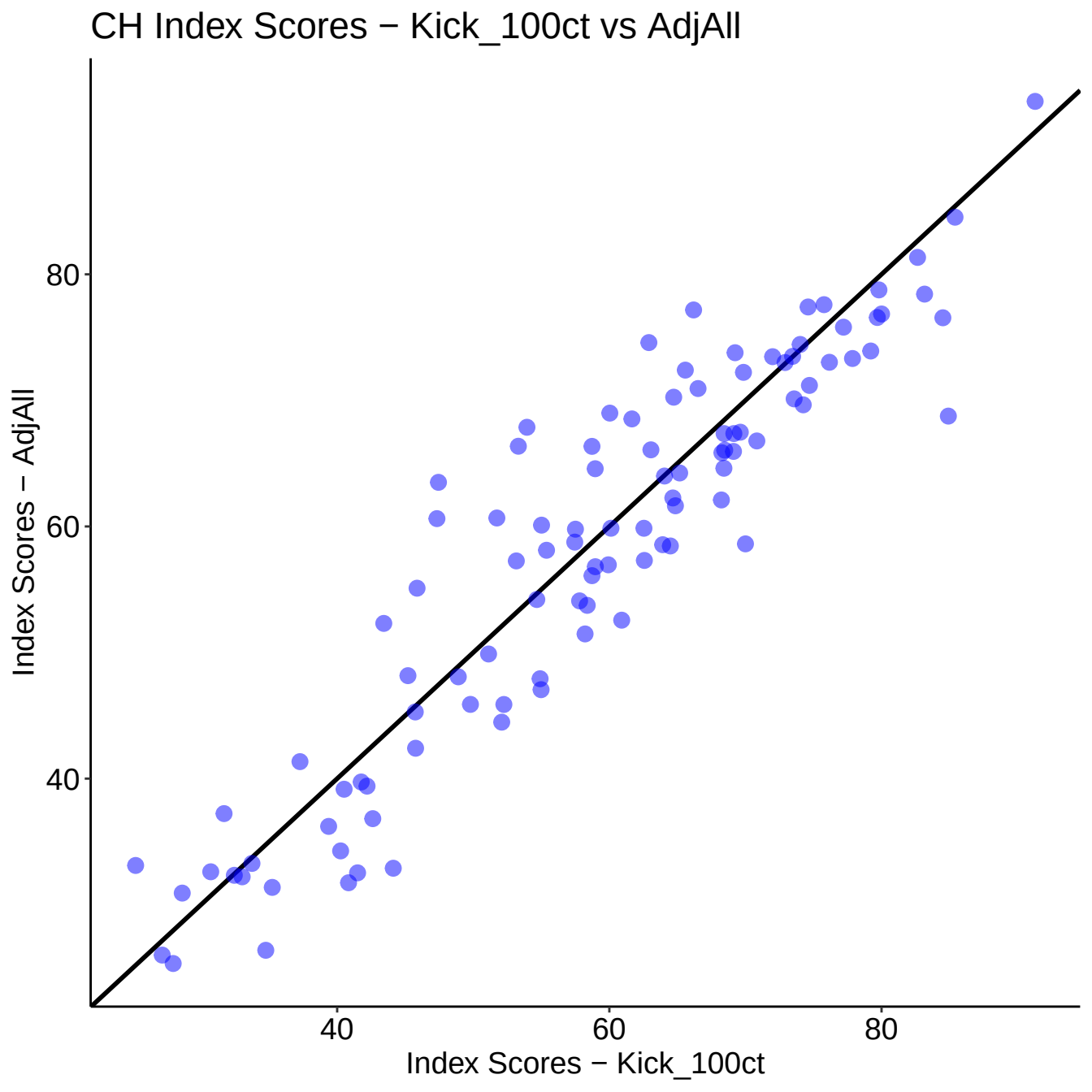
Appendix D

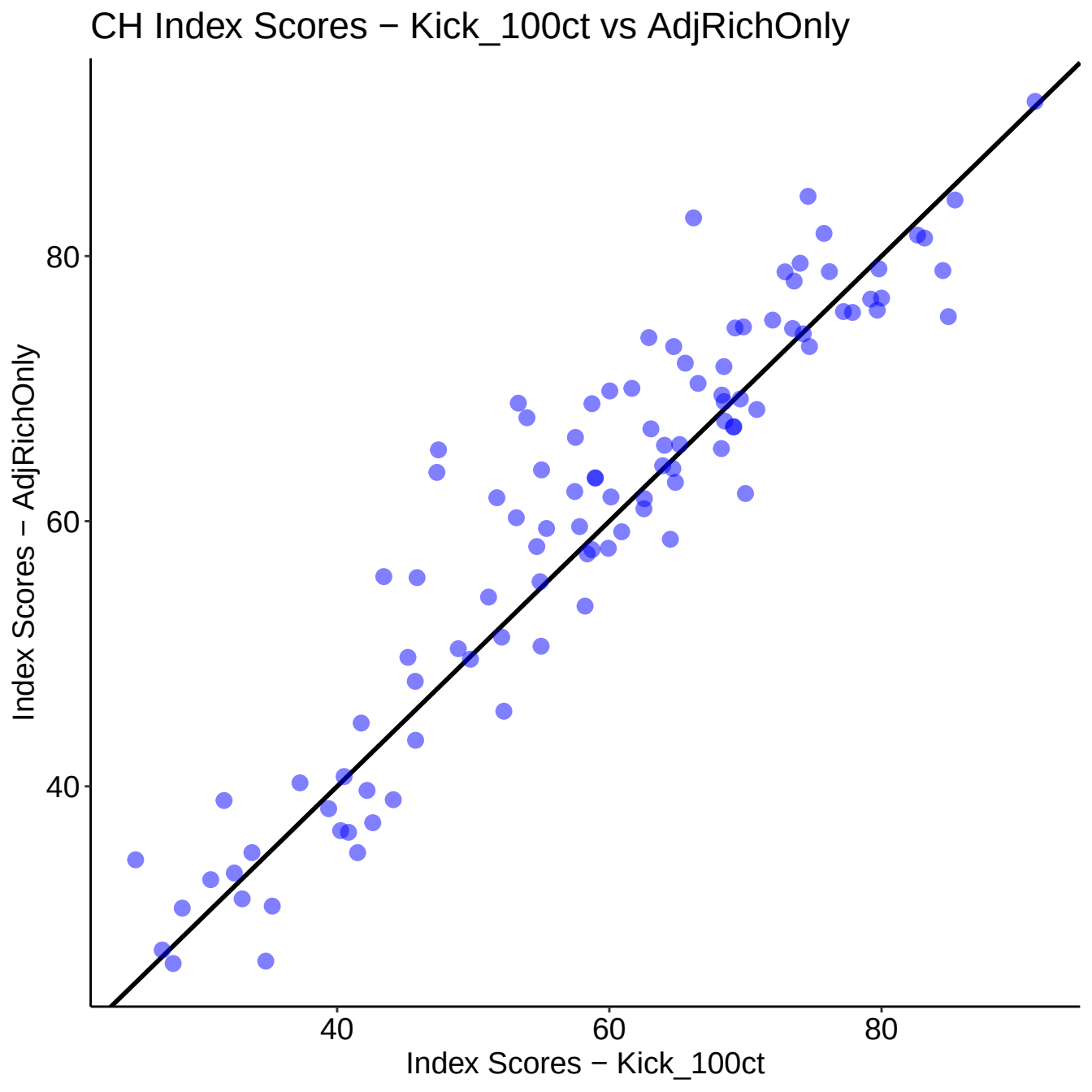
Comparison scatterplots of IBI scores in the Central Hills and Western Highlands. IBI scores were calculated for 100- and 300-count paired samples. The black line indicates a slope of 1 which would mean that IBI scores would be equal between subsample sizes on a sample-by-sample basis.

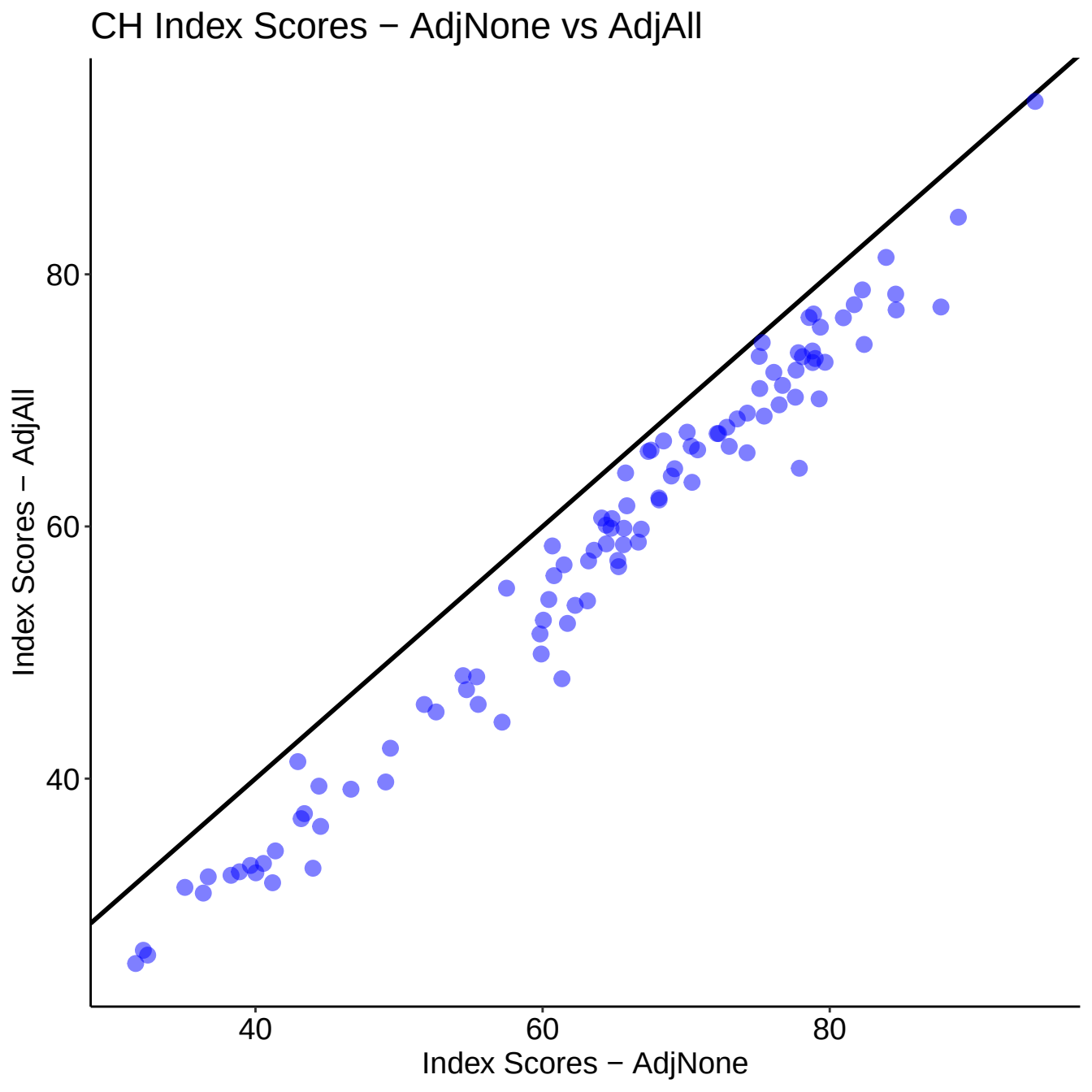
Table 1. Input metrics and scoring formulas for the three IBI alternatives (AdjNone, AdjRichOnly and AdjAll) in the Central Hills and Western Highlands.

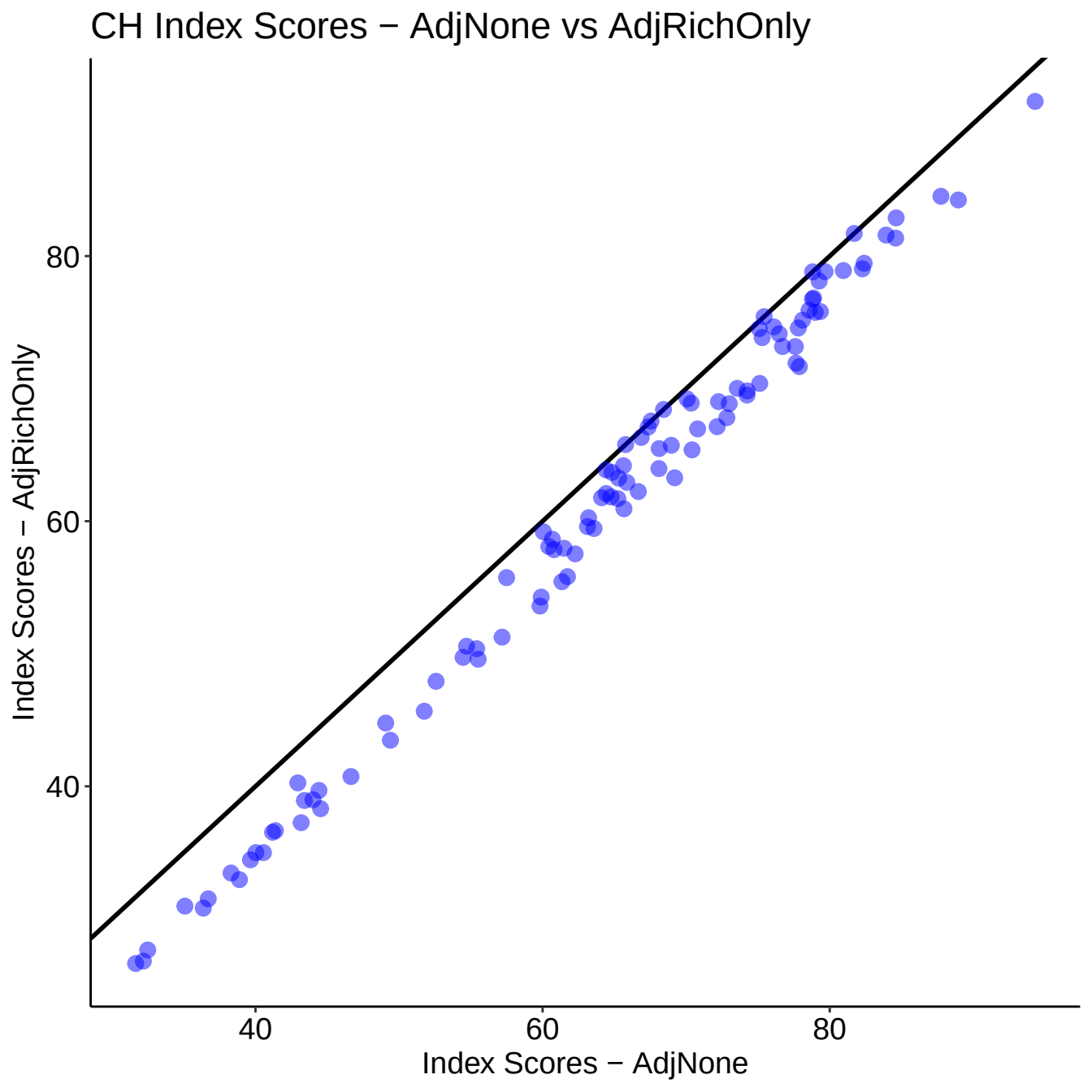
Central Hills 300-count kick net IBI alternatives				
Metric Name (abbrev)	**Response to stress**	**AdjNone**	**AdjRichOnly**	**AdjAll**
Number of total taxa (nt_total)	Dec.	100*(metric)/ 34.9	100*(metric)/ 55.8	100*(metric)/ 55.8
% EPT taxa (pt_EPT)	Dec.	100*(metric)/54.5	100*(metric)/54.5	100*(metric)/ 49.8
% Ephemeroptera individuals, excluding Caenidae and Baetidae (pi_Ephem NoCaeBae)	Dec.	100*(metric)/13.9	100*(metric)/13.9	100*(metric)/ 28.5
% Collector-filterer individuals (pi_ffg_filt)	Inc.	100*(79.9-metric)/66.9	100*(79.9-metric)/66.9	100*(68.7-metric)/ 52.4
% Predator taxa (pt_ffg_pred)	Dec.	100*(metric)/28.5	100*(metric)/28.5	100*(metric)/ 28.5
% Intolerant taxa (pt_tv_intol)	Dec.	100*(metric)/39.1	100*(metric)/39.1	100*(metric)/ 40
Western Highlands 300-count kick net IBI alternatives				
Metric Name	**Response to stress**	**AdjNone**	**AdjRichOnly**	**AdjAll**
Number of total taxa (nt_total)	Dec.	100*(metric)/38.8	100*(metric)/ 61.8	100*(metric)/ 61.8
% Plecoptera individuals (pi_Pleco)	Dec.	100*(metric)/18.3	100*(metric)/18.3	100*(metric)/ 20.6
% Collector-filterer individuals (pi_ffg_filt)	Inc.	100*(50.5-metric)/40.7	100*(50.5-metric)/40.7	100*(55.03-metric)/ 46.9
% Shredder individuals (pi_ffg_shred)	Dec.	100*(metric)/23	100*(metric)/23	100*(metric)/ 34.1
% Intolerant individuals (pi_tv_intol)	Dec.	100*(metric)/51.5	100*(metric)/51.5	100*(metric)/ 45.6
Becks Biotic Index (x_Becks)	Dec.	100*(metric)/36.8	100*(metric)/ 50.6	100*(metric)/ 50.6

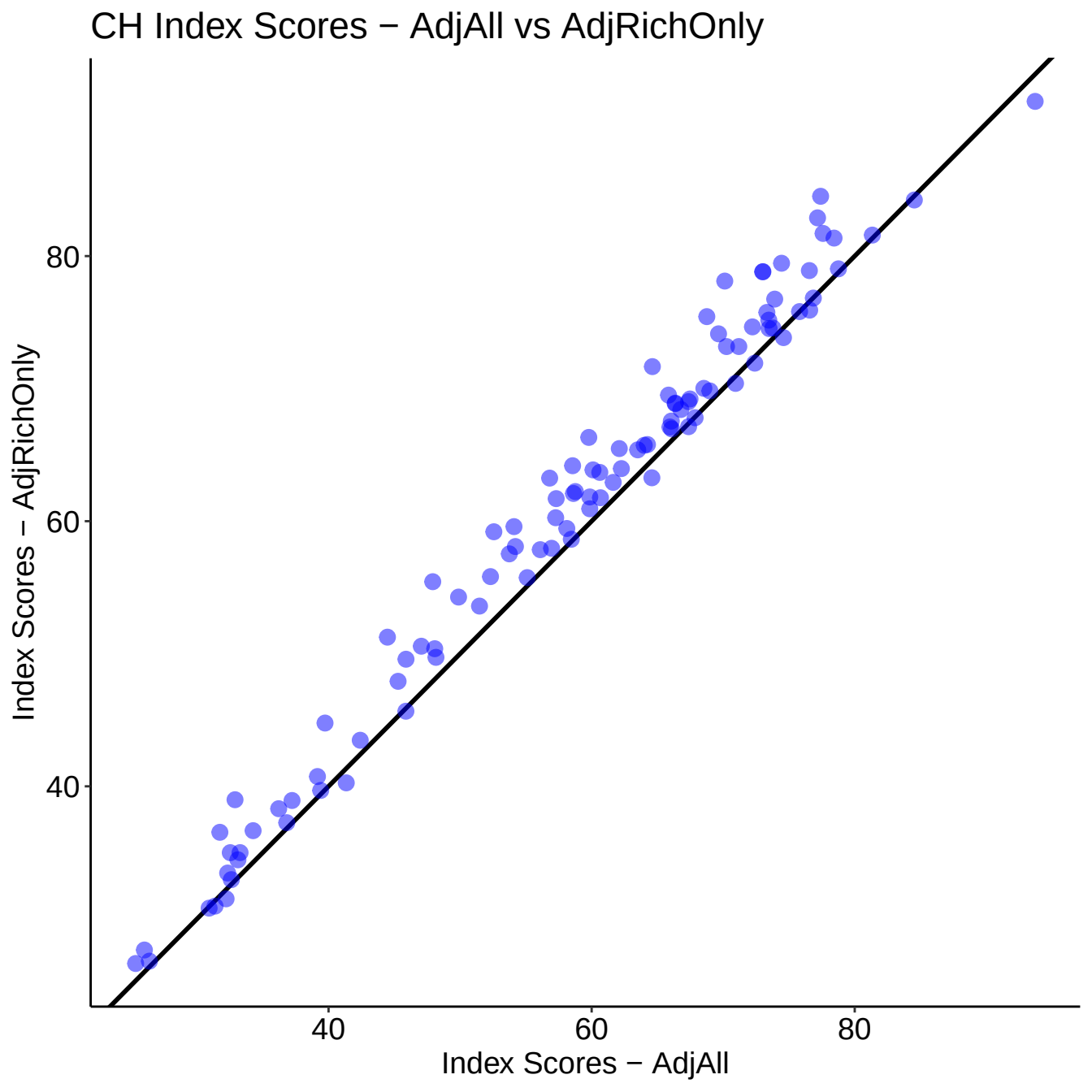

CH Index Scores − Kick_100ct vs AdjNone
Index Scores − AdjNone
80
60
40
40
60
80
Index Scores − Kick_100ct


CH Index Scores − Kick_100ct vs AdjAll
Index Scores − AdjAll
80
60
40
40
60
80
Index Scores − Kick_100ct


CH Index Scores − Kick_100ct vs AdjRichOnly
Index Scores − AdjRichOnly
80
60
40
40
60
80
Index Scores − Kick_100ct

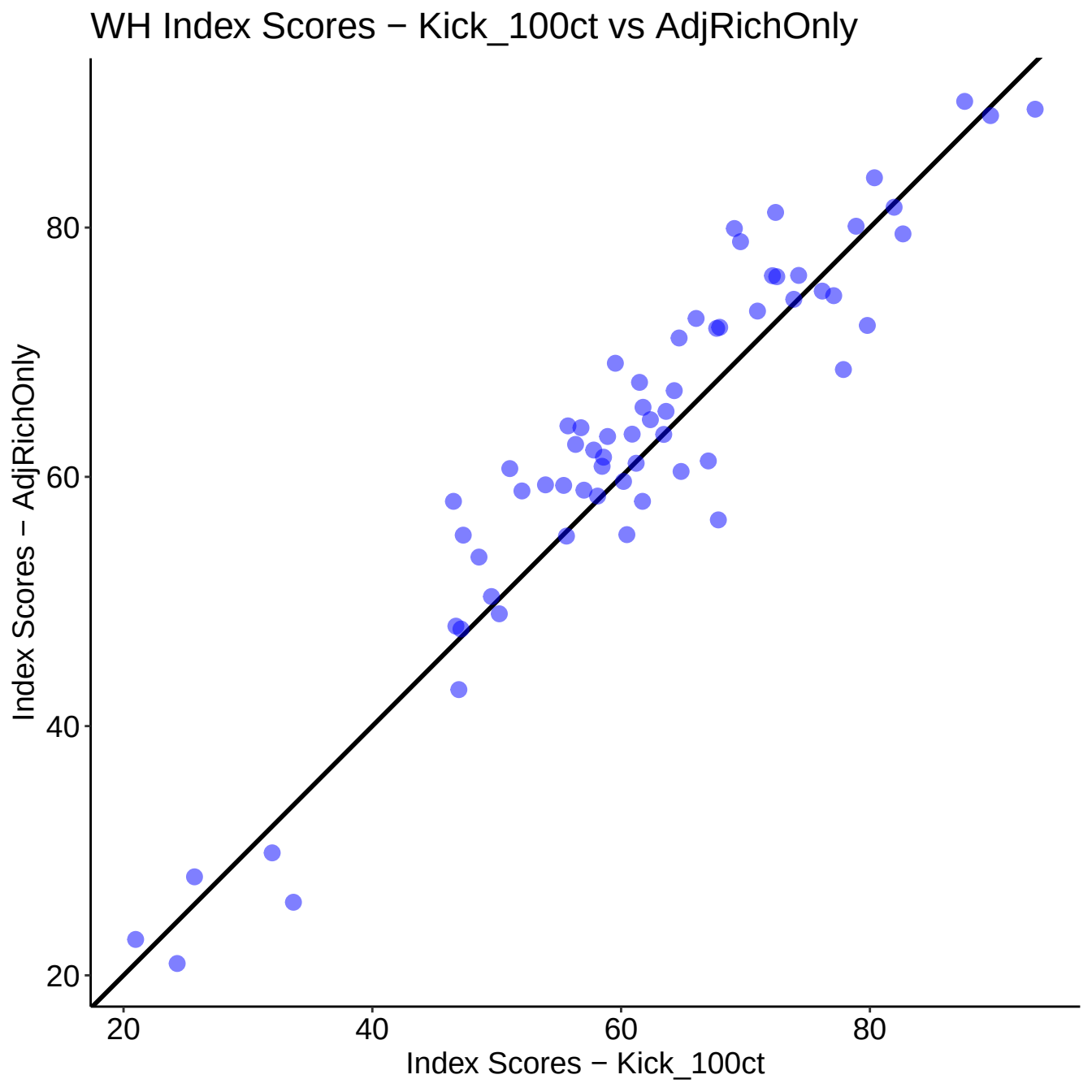

CH Index Scores − AdjNone vs AdjAll
Index Scores − AdjAll
80
60
40
40
60
80
Index Scores − AdjNone

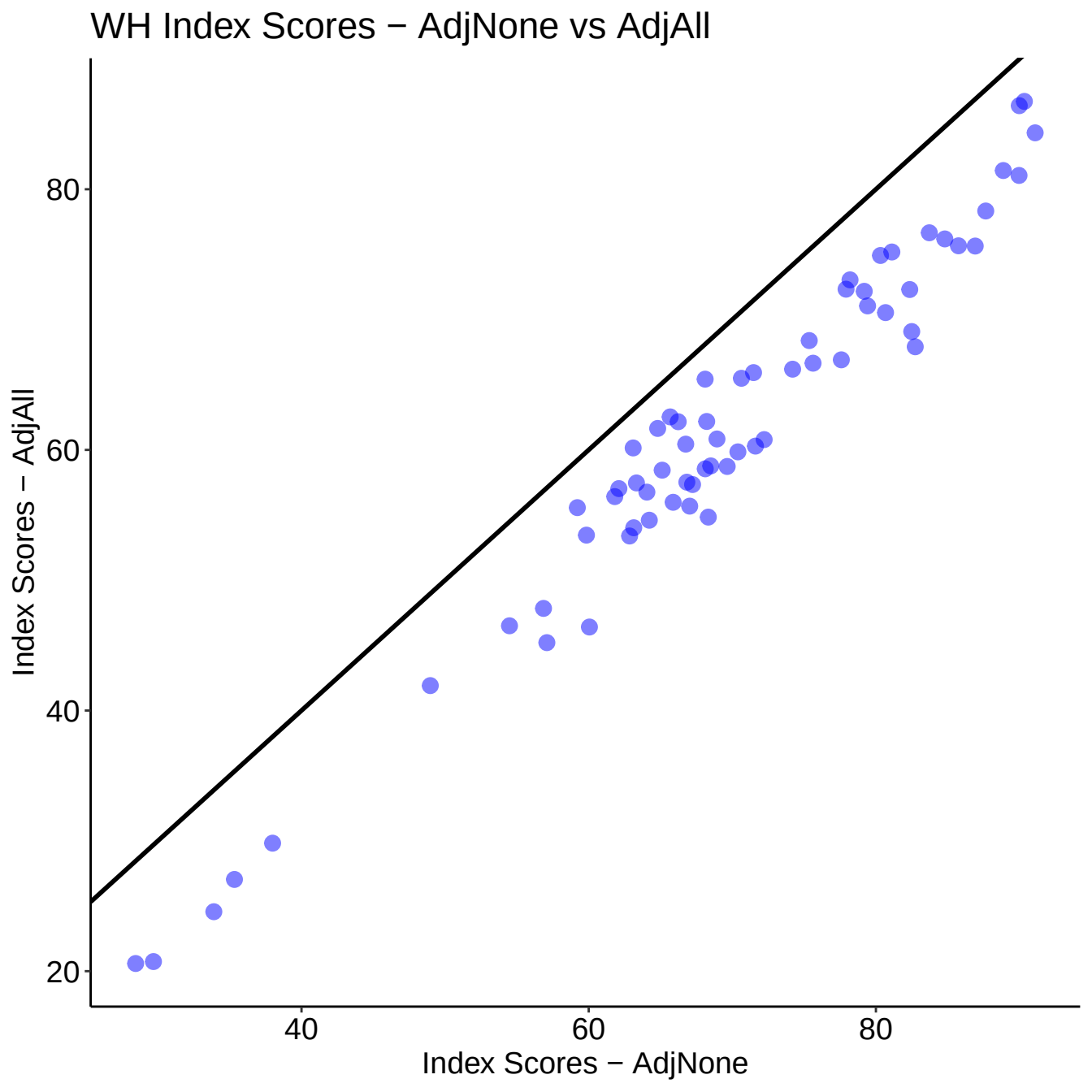

CH Index Scores − AdjNone vs AdjRichOnly
Index Scores − AdjRichOnly
80
60
40
40
60
80
Index Scores − AdjNone

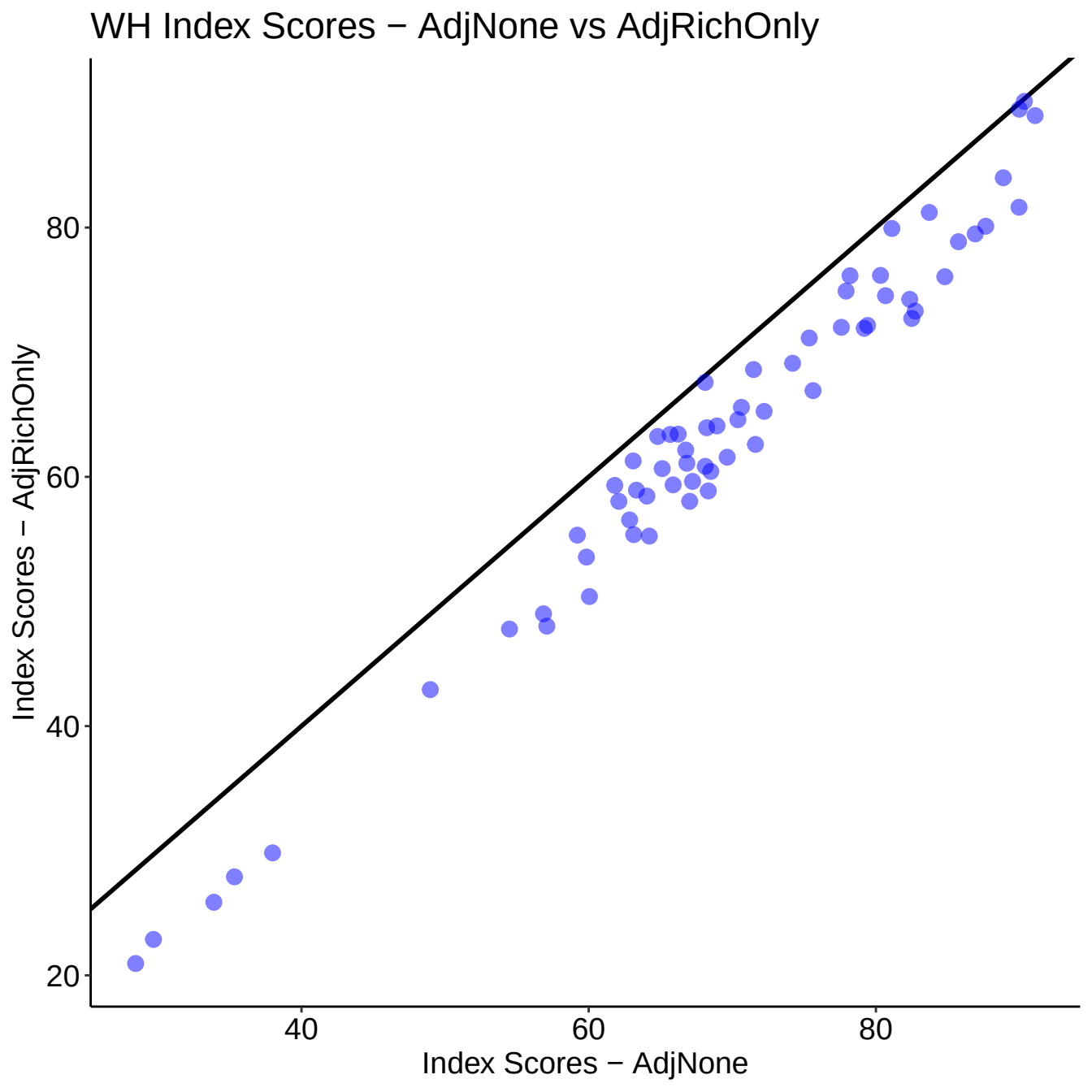

CH Index Scores − AdjAll vs AdjRichOnly
Index Scores − AdjRichOnly
80
60
40
40
60
80
Index Scores − AdjAll


WH Index Scores − Kick_100ct vs AdjNone
Index Scores − AdjNone
Index Scores − Kick_100ct
80
60
40
20
40
60
80


WH Index Scores − Kick_100ct vs AdjAll
Index Scores − AdjAll
Index Scores − Kick_100ct
20
40
60
80
20
40
60
80


WH Index Scores − Kick_100ct vs AdjRichOnly
Index Scores − AdjRichOnly
Index Scores − Kick_100ct
20
40
60
80


WH Index Scores − AdjNone vs AdjAll
Index Scores − AdjAll
Index Scores − AdjNone
20
40
60
80
40
60
80


WH Index Scores − AdjNone vs AdjRichOnly
Index Scores − AdjRichOnly
Index Scores − AdjNone
20
40
60
80
40
60
80


WH Index Scores − AdjAll vs AdjRichOnly
Index Scores − AdjRichOnly
Index Scores − AdjAll
20
40
60
80
20
40
60
80