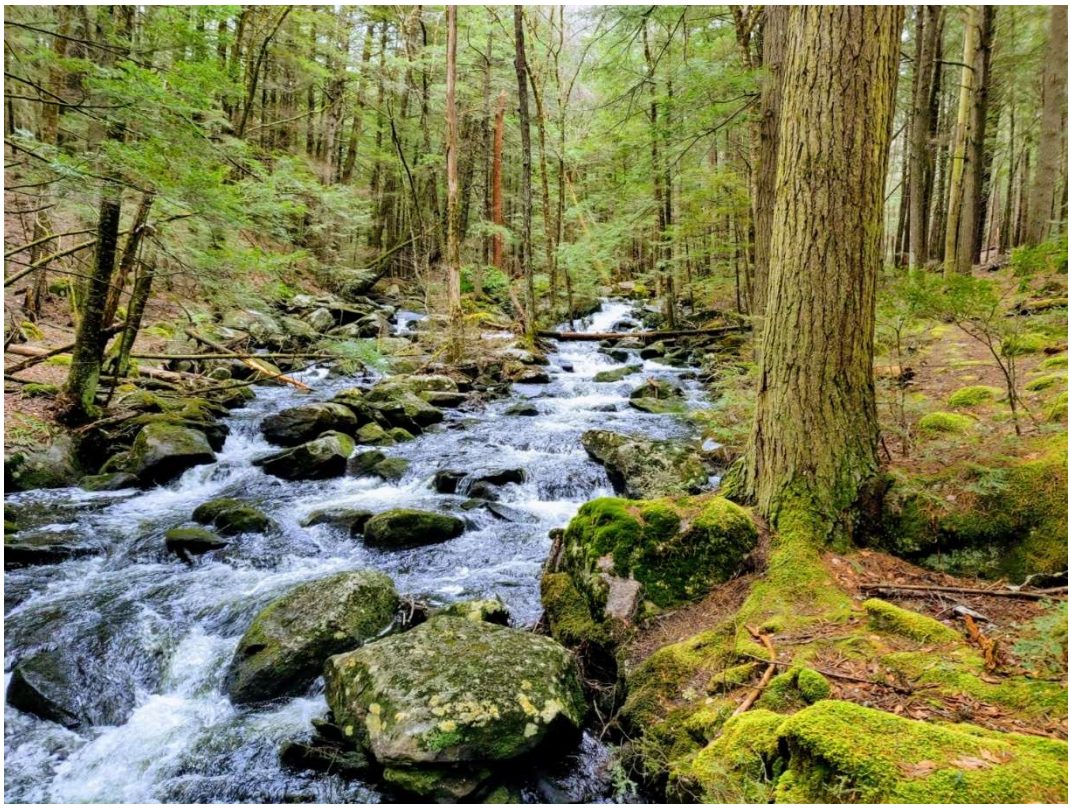
Establishing Numeric Biological Condition Thresholds



***Prepared for*:**

Massachusetts Department of Environmental Protection
James Meek, Work Assignment Manager

***Prepared by*:**

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

August 6, 2020

Executive Summary

Indices of Biotic Integrity (IBI) were calibrated for Massachusetts Department of Environmental Protection (MassDEP) benthic macroinvertebrate kick net samples in two naturally distinct regions: Western Highlands and Central Hills. A separate report (Jessup and Stamp 2020) describes the development of the two IBIs. The IBIs improve MassDEP's diagnostic ability to identify degradation in biological integrity and associated stressors. The Massachusetts Surface Water Quality Standards (SWQS) (314 CMR 4.00; MassDEP 2013) currently has narrative biological criteria that define biological integrity as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having species composition, diversity, and functional organization comparable to that of the natural habitat of the region." In addition, the SWQS designate specific uses for surface water classes. For inland waters, Class A must sustain excellent habitat, while Class B waters must sustain habitat for aquatic life and wildlife. Waters supporting Aquatic Life Use should be suitable for "sustaining a native, naturally diverse, community of aquatic flora and fauna. This use includes reproduction, migration, growth and other critical functions" (MassDEP 2013).

In addition to having narrative biocriteria, some state biomonitoring programs have integrated numeric biocriteria into their SWQS. With numeric biocriteria, management actions can be triggered or prioritized based on assessments relative to a threshold (or thresholds). States like Maine and Minnesota use numeric biocriteria to evaluate Aquatic Life Use Attainment decisions and to designate different categories of biological condition. MassDEP has begun to explore potential IBI thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded). In the future, if MassDEP decides to try and integrate numeric biocriteria into their SWQS, it will warrant additional analyses as well as a rule-making process that includes a period for public review and comment. Any proposed amendments to use numeric biocriteria as the basis for water quality management actions under the Clean Water Act (CWA) would need to be approved by the U.S. Environmental Protection Agency (EPA) following promulgation.

In this document, we provide results from several different analyses that were performed to explore potential IBI thresholds for the four biological condition categories. In addition to presenting results from each analysis, we describe rationale that could potentially be used to justify selection of thresholds, as well as implications of selecting the different thresholds. The analyses derive potential thresholds based on multiple lines of evidence, including distribution statistics, Type I and II error, standard deviation from reference, interpolation with stressor variables, proportional odds logistic regression and models that predict taxa loss. Of these, the distribution statistics and balancing Type I and II error are the most established approaches. The distribution statistics of the least disturbed reference sites focused on potential thresholds in the range of the 10^{th} – 25^{th} percentiles. Thresholds at these percentiles would necessarily misidentify some reference sites as biologically degraded, which would be a Type I error. Considerations for balancing this error with Type II error (recognizing stressed sites as biologically unimpacted) lead to options for appropriate percentiles to suggest as thresholds. The other lines of evidence are less established and thus were regarded as secondary, supporting lines of evidence. Which thresholds are ultimately deemed most appropriate will vary based on factors such as how data are distributed within the datasets and policy decisions.

Acknowledgments

The threshold exploration exercise was performed in collaboration with a team from MassDEP that included James Meek, Arthur Johnson, Allyson Yarra, Anna Mayor, Robert Nuzzo and Joan Beskenis. We are very grateful for their helpful feedback throughout the process, as well as for the review comments by Kalman Bugica.

Table of Contents

Attachments

A Distribution statistics worksheets (Excel)
B Interpolation with stressors worksheets (Excel)
C O/E worksheet – Central Hills (Excel)
D O/E worksheet – Western Highlands (Excel)
E Combined threshold worksheets (Excel)

List of Tables

List of Figures

1 Background

The Massachusetts Department of Environmental Protection (MassDEP) is responsible for sampling and assessing Massachusetts's surface water quality pursuant to the Clean Water Act (CWA) Section 305(b). One of the purposes of the CWA is the restoration and maintenance of the chemical, physical, and biological integrity of the Nation's waters. The Massachusetts Surface Water Quality Standards (SWQS) (314 CMR 4.00; MassDEP 2013) has narrative biological criteria that define biological integrity as "the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having species composition, diversity, and functional organization comparable to that of the natural habitat of the region." In addition, the SWQS designate specific uses for surface water classes. For inland waters, Class A must sustain excellent habitat, while Class B waters must sustain habitat for aquatic life and wildlife. Waters supporting Aquatic Life Use should be suitable for "sustaining a native, naturally diverse, community of aquatic flora and fauna. This use includes reproduction, migration, growth and other critical functions" (MassDEP 2013).

To determine whether the macroinvertebrate communities in Massachusetts' freshwater wadeable streams exhibit biological integrity, Indices of Biotic Integrity (IBI) were calibrated for freshwater wadeable streams in all but the southeastern portion of the state (Narragansett/Bristol Lowlands (NBL), Cape Cod (CC), and the Islands)[1] (Jessup and Stamp 2020). The IBIs were calibrated for MassDEP benthic macroinvertebrate kick net samples in two naturally distinct regions: Western Highlands (WH) and Central Hills (CH). The IBIs for each region were comprised of biological metrics that were found to be responsive to a general stressor gradient. The new IBIs improve MassDEP's diagnostic ability to identify degradation in biological integrity and water quality.

In addition to having narrative biocriteria, some state biomonitoring programs have integrated numeric biocriteria into their SWQS. With numeric biocriteria, management actions can be triggered or prioritized based on assessments relative to a threshold (or thresholds). States like Maine and Minnesota use numeric biocriteria to evaluate Aquatic Life Use (ALU) Attainment decisions and to designate different categories of biological condition[2]. MassDEP has begun to explore potential IBI thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded, and Severely Degraded). In this document, we provide results from several different analyses that were performed to explore potential IBI thresholds for the four categories. The analyses derive potential thresholds by calculating distribution statistics, examining Type I and II error, evaluating standard deviation from reference, interpolating with stressor variables, performing proportional odds logistic regression and evaluating models that predict taxa loss. In addition to presenting results from each analysis, we describe rationale that could potentially be used to justify the selection of the thresholds. We also discuss implications of selecting the different thresholds. If at some point MassDEP decides to take the additional step of integrating numeric biocriteria into their SWQS, it will warrant additional analyses as well as a rule-making process that includes a period for public review and comment, as well as approval by the U.S. Environmental Protection Agency (EPA).

[1]NBL/CC/Islands is a third distinct region with insufficient data at this time to develop an IBI.

[2]More information on numeric biocriteria in Maine and Minnesota can be found at the following links: Maine - https://www.maine.gov/dep/water/monitoring/biomonitoring/retro/pt1ch1pref.pdf; Minnesota - https://www.pca.state.mn.us/sites/default/files/wq-bsm4-02.pdf.

2 Dataset

The threshold analyses were performed on the MassDEP benthic macroinvertebrate kick net IBI calibration and verification dataset. Section 2 of the kick net IBI report (Jessup and Stamp 2020) describes the dataset in detail. We used two IBI scoring schemes in the threshold analyses: 'minimum floor' (*Indx_MinFloor*)[3] and a normalized 'minimum floor' scheme (*IndxMF_MSD*)[4]. Figure 1 shows the locations of the sites that were used in the threshold analyses. Sites were assigned to disturbance categories using the process outlined in Figure 2. Seven disturbance variables were considered[5]: Index of Catchment Integrity (ICI), Index of Watershed Integrity (IWI), 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 (Table 1). Sites were initially assigned to seven disturbance categories, ranging from Best Reference to High Stress (Table 2), using the thresholds in Table 1, combination rules described in the kick net IBI report and local knowledge of MassDEP staff (Section 3; Jessup and Stamp 2020). Sites were then collapsed into three broader disturbance categories (reference, stressed, other) for the analyses.

In regard to overall watershed condition (as measured by the ICI and IWI) and percent urban land cover, the CH sites have higher levels of disturbance than WH sites (Appendix A, Figures A1-A3). 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 in the CH were derived from the High Stress category, while in the WH, the High Stress and Stress categories were combined (Table 2). 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 2).

[3] The *Indx_MinFloor* is based on a 100-point scale; it uses the minimum possible value (zero) and the 95th percentile for all decreasing metrics. Metrics that increase with stress use the 5th and 95th percentiles. For more information, see the kick IBI report (Jessup and Stamp 2020).

[4]The *IndxMF_MSD* is based on the *Indx_MinFloor* scheme, normalized to a scale centered around 0 and frames scores standardized to the mean and standard deviation of the index calibration reference distribution, such that a score of -1 means it is 1 standard deviation from the mean. For more information, see Jessup and Stamp 2020.

[5]Variables were selected based on the process described in Section 3 of the kick IBI report (Jessup and Stamp 2020)

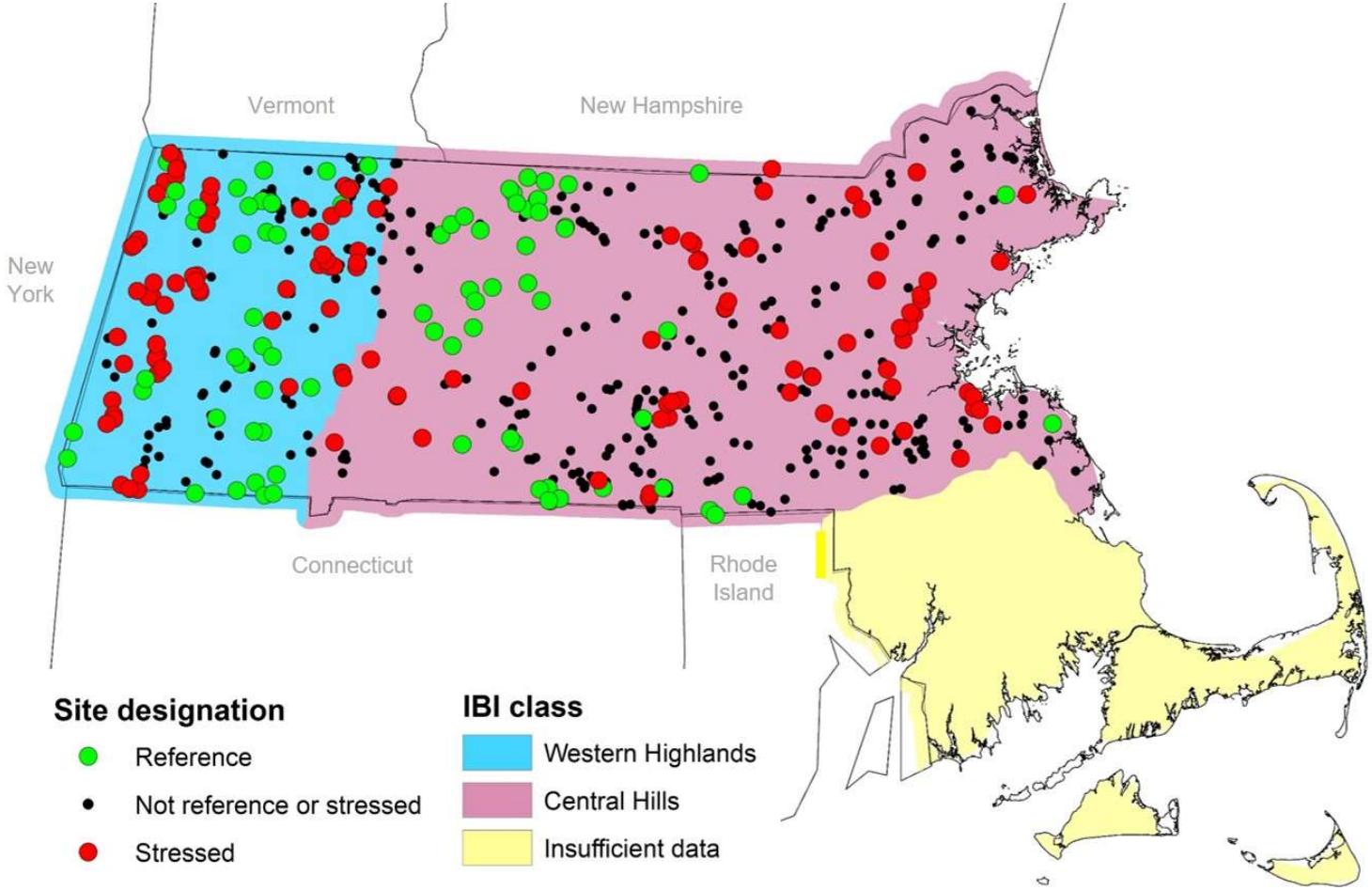


Figure 1. Locations of sites that were used in the threshold analyses. Sites are color-coded by broad disturbance category. Several sites are located just across the Massachusetts border. These samples were included in the analyses because they were sampled by MassDEP field crews using MassDEP kick RBP methods.

Disturbance variables (landscape-scale, GIS-based)

1. Index of watershed integrity (IWI)
2. Index of catchment integrity (ICI)
3. % Urban land cover
4. % Hay + Row Crop land cover
5. Ag application rates (kg N/ha/yr)
6. Road density (km/square km)
7. Dam storage volume (cubic meters/square km)

↓

Score each metric

+3 (best) to -3 (worst) based on the disturbance level thresholds (see Table 1).

↓

Assign sites to preliminary disturbance categories

7 categories: Best Reference to High Stress (see Table 2) based on the combination rules described in the kick IBI report (Section 3; Jessup and Stamp 2020)

↓

Finalize disturbance category assignments

Review by MassDEP staff; change designations as needed based on local knowledge or other information not available in the GIS-based data

↓

Collapse to broader disturbance categories for analyses

Reference:

- WH = Best Reference + Reference
- CH = Best Reference + Reference + Sub Reference

Stressed

- WH = High Stress + Stress
- CH = High Stress

Figure 2. Process that was used to assign sites to disturbance categories. More detailed information on development of the disturbance gradient can be found in the kick IBI report (Jessup and Stamp 2020).

Table 1. Variables and thresholds that were used to define the disturbance gradient. For more information, see Section 3 in Jessup and Stamp (2020).

Metric disturbance levels (scores)	Metric thresholds						
	Index of Watershed Integrity[a] (IWI)	**Index of Catchment Integrity (ICI)[a]**	**% Urban[b]**	**% Hay + Row Crop[c]**	**Ag application rates (kg N/ha/yr)[d]**	**Road density (km/square km)[e]**	**Dam storage volume (cubic meters/square km)[f]**
Disturb Level 1 (least disturbed) (score +3)	≥0.875	≥0.875	≤1%	≤1%	≤0.5	≤1.5	≤0.1
Disturb Level 2 (score +2)	≥0.85	≥0.85	≤2%	≤2%	≤1	≤2	≤1,000
Disturb Level 3 (score +1)	≥0.80	≥0.80	≤5%	≤5%	≤2.5	≤3	≤10,000
Disturb Level 4 (score 0)	>0.80 and <0.75	>0.80 and <0.75	>5 and <10%	>5 and <10%	>2.5 and <5	>3 and <5	>10,000 and <50,000
Disturb Level 5 (score -1)	≤0.75	≤0.75	≥10%	≥10%	≥5	≥5	≥50,000
Disturb Level 6 (score -2)	≤0.60	≤0.60	≥40%	≥15%	≥7.5	≥7.5	≥100,000
Disturb Level 7 (most disturbed) (score -3)	≤0.50	≤0.50	≥60%	≥20%	≥10	≥10	≥200,000

[a]Scoring scale ranges from 0 (worst) to 1 (best); based on version 1 (Thornbrugh et al. 2018)

[b]Percent of watershed area classified as developed, high + medium + low-intensity land use (NLCD 2011 class 24+23+22)

[c]Percent of local catchment area classified as hay and crop land use (NLCD 2011 class 82+81)

[d][CBNFWs]+[FertWs]+[ManureWs]

- CBNFWs = Mean rate of biological nitrogen fixation from the cultivation of crops in kg N/ha/yr, within watershed
- FertWs = Mean rate of synthetic nitrogen fertilizer application to agricultural land in kg N/ha/yr, within watershed
- ManureWs = Mean rate of manure application to agricultural land from confined animal feeding operations in kg N/ha/yr, within watershed

[e]Density of roads (2010 Census Tiger Lines) within catchment (km/square km)

[f]Volume all reservoirs (NID_STORA in NID) per unit area of watershed (cubic meters/square km)

Table 2. Number of samples in each of the seven disturbance categories (Best Reference to High Stress, as described in Section 3 of Jessup and Stamp 2020). Sites were later collapsed into three broader categories (reference, stressed, other) 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	Number of samples	
	Western Highlands	Central Hills
Best Reference	7	4
Reference	34	13
Sub Reference	25	28
Other	15	26
Some Stress	48	89
Stress	58	135
High Stress	12	63
Reference	41	45
Other	88	250
Stress	70	63
Total	***199***	***358***

3 Thresholds

We explored potential threshold options for four biological categories (Figure 3):

- Exceptional Condition
    - Denotes samples that demonstrate an exceptional condition as determined by measurement of the biology, and that support biological assemblages with community structure and ecosystem functions that represent the best observable biological conditions.
- Satisfactory Condition
- Moderately Degraded Condition
- Severely Degraded Condition
    - Denotes samples that are in the worst biological condition, where severe changes have occurred in the structure and function of the biological assemblage as compared to the reference condition.

Having multiple thresholds provides an opportunity to document incremental change. For example, if restoration activities are performed at a site with IBI scores in the Severely Degraded range, and after restoration the IBI score improves to Moderately Degraded, this shows that the site is on the desired trajectory. On the other end of the disturbance spectrum, if a high quality reference site that has

consistently scored at or above the Exceptional Condition threshold starts to slip below that threshold, but still meets the Satisfactory Condition threshold, it is important to detect this change early and address the issues causing the degradation.

If, in the future, MassDEP takes the additional step of integrating numeric biocriteria into their SWQS, management actions would be triggered or prioritized based on assessments relative to a threshold (or thresholds). For example, there would be a threshold that denotes whether samples were supporting vs. not supporting ALU Goals and Clean Water Act (CWA) objectives. If the sample did not meet the threshold for attaining ALU, it would potentially be listed on the 303(d) list of impaired waters following further evaluation. Whether or not a segment supports ALU is based on a variety of weights of evidence (beyond just macroinvertebrate biological integrity). Integration of numeric biocriteria into SWQS requires a rule-making process that includes a period for public review and comment and would need to be approved by the EPA following promulgation.

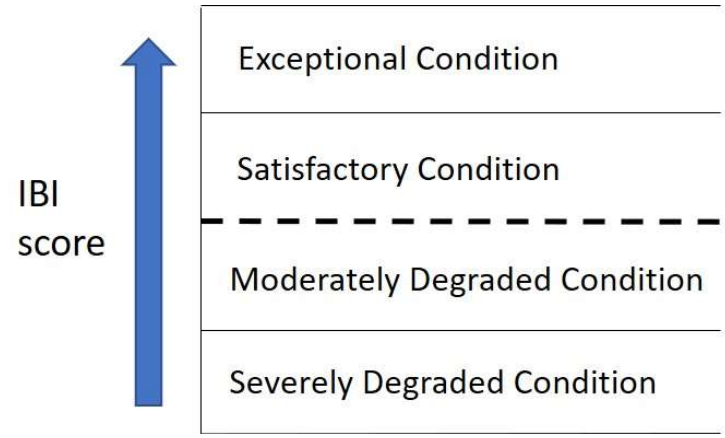


Figure 3. We performed several different analyses to explore potential thresholds for four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded Condition, and Severely Degraded Condition).

4 Individual lines of evidence

To help inform potential thresholds, we used several different, independent exploratory approaches, referred to as 'lines of evidence'. They included:

- Distribution statistics and balancing error
- Standard deviation from reference
- Interpolation with stressors
- Proportional odds logistic regression
- Taxa loss

4.1 Distribution statistics & balancing error

The distribution statistics, or percentile-based, approach (in particular, the reference condition (RC) approach) is commonly used by states for setting numeric biocriteria thresholds (Hughes et al. 1986, Gibson et al. 1996). With the RC approach, IBI scores are calculated from a least-disturbed reference site

dataset, and then a percentile of the IBI scores is chosen to represent the RC. Typically, the 25th or 10th percentile is used for the satisfactory condition threshold (e.g., Yoder and Rankin 1995, DeShon 1995, Barbour et al. 1996, Roth et al. 1997), but this varies across datasets. Having a sound, well-documented, reference dataset is critical to this approach.

When selecting percentiles, it is important to consider:

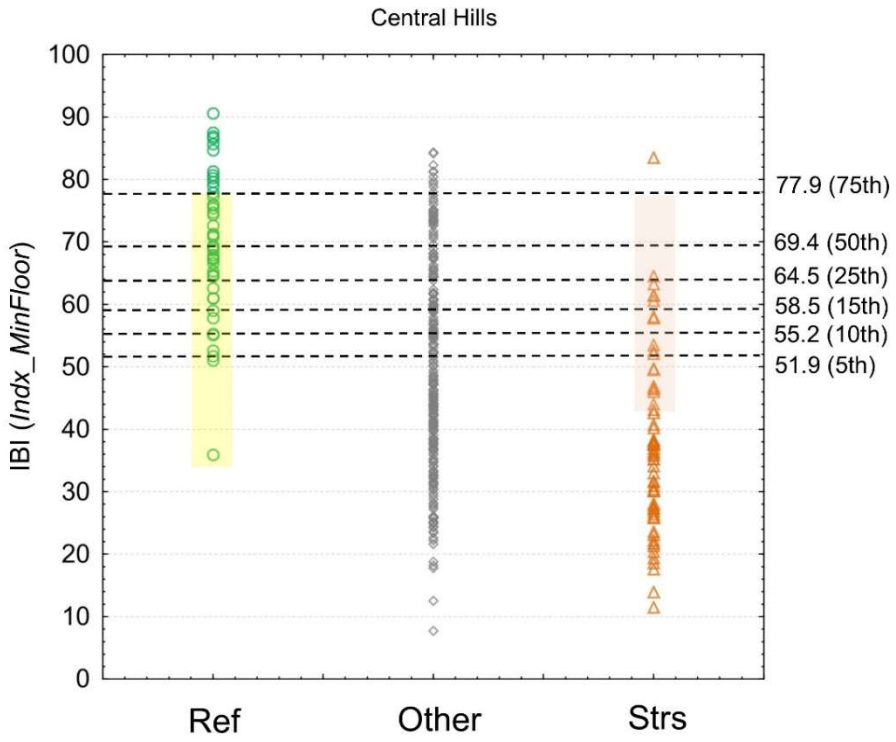
- **Level of disturbance**. Using a higher percentile in situations where the reference dataset has higher levels of disturbance (and likely includes some sites that are not truly of reference quality) provides a degree of safety. The 10th percentile of RC is generally used where there is greater confidence that the reference sites are of high quality (for example, with datasets that have many "minimally disturbed" sites). In reference datasets with only "least disturbed" sites, the 25th percentile is typically used (Stoddard et al. 2006). The 5th percentile of reference datasets is not commonly used since it is prone to the effect of outliers as well as variability and potential error in reference designations. Even when reference sites are carefully selected from a robust pool of minimally disturbed locations, natural variability and sampling error precludes the assumption that every reference sample is representative of biological integrity goals, so using percentiles less than 10% would likely underestimate impairment (or the departure from the desired reference condition). In heavily disturbed areas or regions where a stream class has overall poor condition (i.e., poorer than least disturbed), thresholds based on the 25th or 10th percentile are likely to be under-protective so an alternative or modified approach (such as the 75th or 90th percentile of all sites) is sometimes used. For example, in the Huron/Erie Lake Plains (HELP) ecoregion, Ohio based their threshold on the 90th percentile of all sites (Ohio EPA 1987, 1989).
- **Ratios of Type I and Type II error.** 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.

For this exercise, we calculated distribution statistics for two datasets: 1) reference sites only; and 2) all samples. Results varied across the two regions, with the reference-based percentiles being approximately 5 points higher on average in the CH dataset (Table 3). The values in the 'all' dataset showed the opposite pattern for the 5th through 50th percentiles, with higher values in the WH dataset; values for the 75th through 95th percentiles were comparable (Table 3). Attachment A contains the Excel worksheets that were used to calculate the distribution statistics. Additional statistics are included in Appendix B (e.g., the mean and standard deviation of IBI scores in each of the seven disturbance categories, from Best Reference to High Stress).

Table 3. Comparison of IBI (Indx_MinFloor) scores for multiple percentiles in the Western Highlands vs Central Hills based on two datasets: reference only (highlighted in green); and all samples.

Dataset	Percentiles	IBI score (*Indx_MinFloor*)	
		Western Highlands	Central Hills
Reference sites	5th	45.9	51.9
	10th	47	55.2
	15th	55	58.5
	20th	57.8	61
	25th	59.6	64.5
	50th	62.7	69.4
	75th	73.9	77.9
	90th	81.5	85.3
	95th	87.7	86.8
All sites	5th	27.4	23.4
	10th	32	27.1
	25th	42.1	36.6
	50th	53.1	48.8
	75th	64.7	64.5
	90th	75.8	74.5
	95th	79.9	79.3

We measured Type I error as the percentage of reference sites that fell below the various percentiles/potential thresholds and Type II error as the percentage of stressed sites that had scores greater than or equal to the thresholds. Type I and II error rates and the number and percentages of samples that fell above or below the various thresholds are summarized in Figures 4 (CH) and 5 (WH). In the CH dataset, the 10th percentile of reference sites (which corresponds with an IBI score of 55.2) had the smallest difference between Type I and II errors (Type I = 11.1%; Type II = 14.3%; difference 3.2%). If 55.2 was used as the Satisfactory Biological Condition threshold in the CH, 147 samples (41.1%) would score greater than or equal to the threshold, and 211 samples (58.9%) would score below the threshold. In the WH dataset, the 15th percentile of reference sites (which corresponds with an IBI score of 55.0) had the smallest difference (Type I = 14.6%; Type II = 17.1%; difference 2.5%). If 55.0 was used as the Satisfactory Biological Condition threshold in the Western Highlands, 90 samples (45.2%) would score greater than or equal to the threshold, and 109 samples (54.8%) would score below the threshold.



Type I error

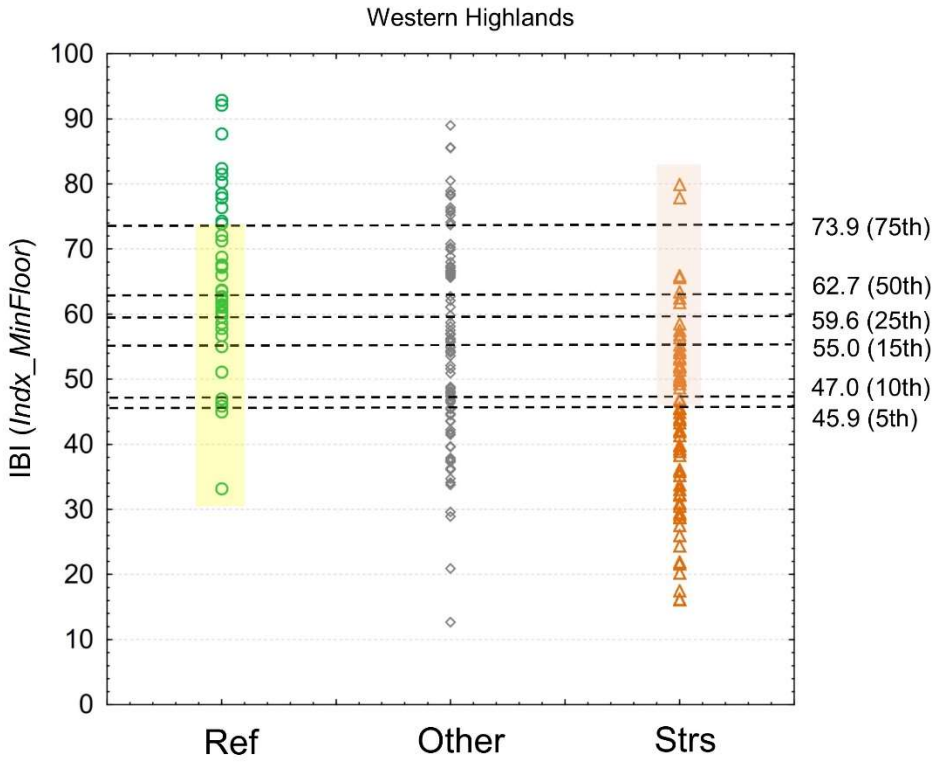
- Erroneously calling a site disturbed
- Calculated as the % of Ref sites with IBI scores < the threshold. For example, if 55.2 is the threshold, 5 out of 45 ref sites (or 11.1%) have IBI scores < 55.2.
- Type I error rates increase as the threshold increases

Type II error

- Erroneously calling a disturbed site undisturbed
- Calculated as the % of stressed sites with IBI scores ≥ the threshold. For example, if 55.2 is the threshold, 9 out of 63 strs sites (or 14.3%) have IBI scores ≥ 55.2.
- Type II error rates decrease as the threshold increases

Percentile (ref only)	IBI score (*Indx_MinFloor*)	Number (%) samples in each disturbance group						Type I error	Type II error	Difce
		≥ Threshold			< Threshold					
		Ref (n= 45)	Other (n=250)	Strs (n=63)	Ref (n= 45)	Other (n=250)	Strs (n=63)			
5th	51.9	42 (93.3%)	112 (44.8%)	12 (19.0%)	3 (6.7%)	138 (55.2%)	51 (81.0%)	6.7%	19.0%	12.4%
10th	55.2	40 (88.9%)	98 (39.2%)	9 (14.3%)	5 (11.1%)	152 (60.8%)	54 (85.7%)	11.1%	14.3%	3.2%
15th	58.5	38 (84.4%)	82 (32.8%)	6 (9.5%)	7 (15.6%)	168 (67.2%)	57 (90.5%)	15.6%	9.5%	6.0%
20th	61.0	37 (82.2%)	65 (26.0%)	5 (7.9%)	8 (17.8%)	185 (74.0%)	58 (92.1%)	17.8%	7.9%	9.8%
25th	64.5	35 (77.8%)	55 (22.0%)	2 (3.2%)	10 (22.2%)	195 (78.0%)	61 (96.8%)	22.2%	3.2%	19.0%
50th	69.4	23 (51.1%)	35 (14.0%)	1 (1.6%)	22 (48.9%)	215 (86.0%)	62 (98.4%)	48.9%	1.6%	47.3%
75th	77.9	12 (26.7%)	10 (4.0%)	1 (1.6%)	33 (73.3%)	240 (96.0%)	62 (98.4%)	73.3%	1.6%	71.7%

Figure 4. Distribution of IBI scores (Indx_MinFloor) across the three broad disturbance categories (reference (Ref), stressed (Strs) and other (not reference or stressed) in the Central Hills dataset. 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 the various thresholds. 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 sites with IBI scores < the threshold. For example, if 55 is the threshold, 6 out of 41 ref sites (or 14.6%) have IBI scores < 55
- Type I error rates increase as the threshold increases

Type II error

- Erroneously calling a disturbed site undisturbed
- Calculated as the % of stressed sites with IBI scores ≥ the threshold. For example, if 55 is the threshold, 12 out of 70 strs sites (or 17.1%) have IBI scores ≥ 55
- Type II error rates decrease as the threshold increases

Percentile (ref only)	IBI score (*Indx_MinFloor*)	Number (%) samples in each disturbance group						Type I error	Type II error	Difce
		≥ Threshold			< Threshold					
		Ref (n= 41)	Other (n=88)	Strs (n=70)	Ref (n= 41)	Other (n=88)	Strs (n=70)			
5th	45.9	39 (95.1%)	64 (72.7%)	30 (42.9%)	2 (4.9%)	24 (27.3%)	40 (57.1%)	4.9%	42.9%	38.0%
10th	47.0	37 (90.2%)	60 (68.2%)	29 (41.4%)	4 (9.8%)	28 (31.8%)	41 (58.6%)	9.8%	41.4%	31.7%
15th	55.0	35 (85.4%)	43 (48.9%)	12 (17.1%)	6 (14.6%)	45 (51.1%)	58 (82.9%)	14.6%	17.1%	2.5%
20th	57.8	33 (80.5%)	35 (39.8%)	8 (11.4%)	8 (19.5%)	53 (60.2%)	62 (88.6%)	19.5%	11.4%	8.1%
25th	59.6	31 (75.6%)	33 (37.5%)	7 (10.0%)	10 (24.4%)	55 (62.5%)	63 (90.0%)	24.4%	10.0%	14.4%
50th	62.7	21 (51.2%)	29 (33.0%)	5 (7.1%)	20 (48.8%)	59 (67.0%)	65 (92.9%)	48.8%	7.1%	41.6%
75th	73.9	11 (26.8%)	12 (13.6%)	2 (2.9%)	30 (73.2%)	76 (86.4%)	68 (97.1%)	73.2%	2.9%	70.3%

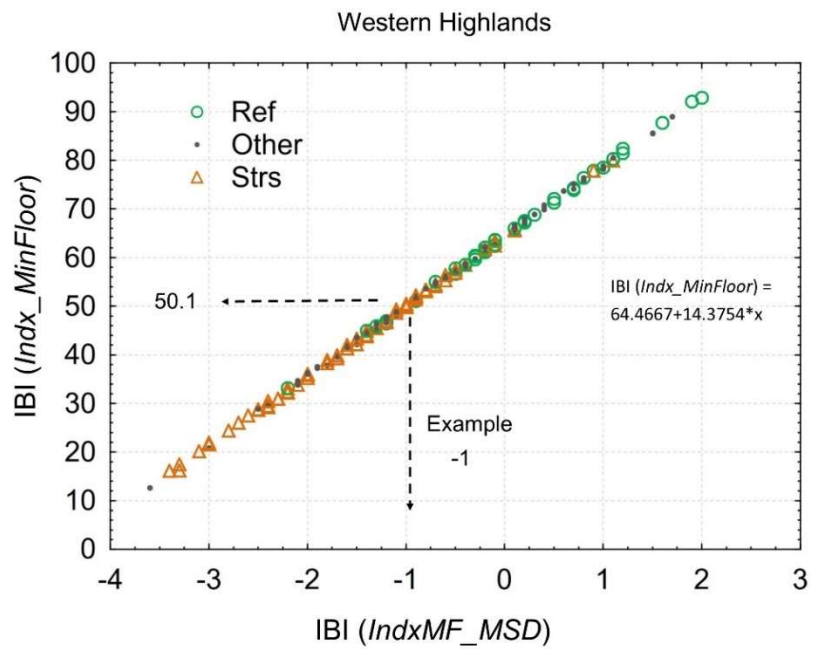
Figure 5. Distribution of IBI scores (Indx_MinFloor) across the three broad disturbance categories (reference (Ref), stressed (Strs) and other (not reference or stressed) in the Western Highlands dataset. The table summarizes Type I and II error rates and the number and percentages of samples in each disturbance category that fell above or below the various thresholds. Cells highlighted in yellow show Type I error rates; light orange cells show Type II error rates.

4.2 Standard deviation from reference

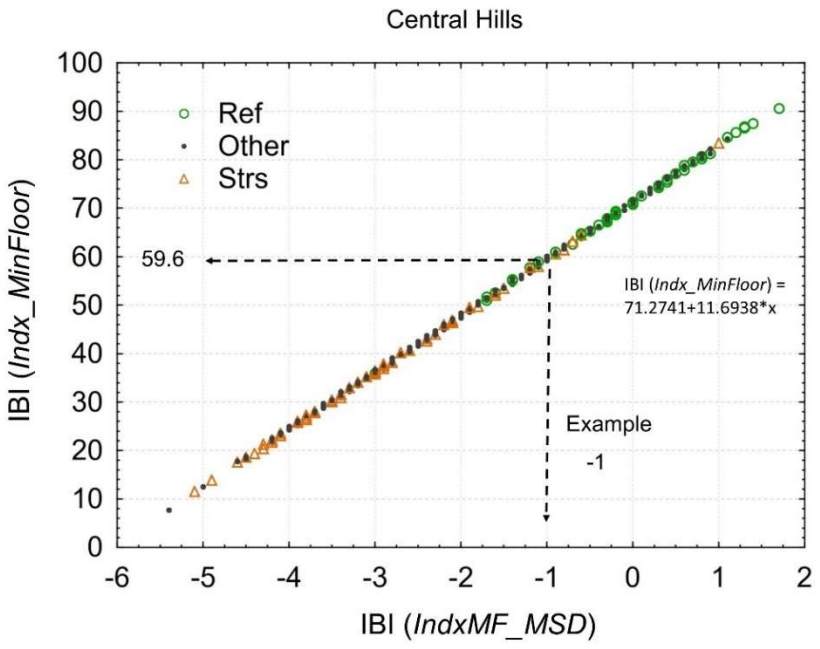
Another threshold derivation approach involved using the *IndxMF_MSD* IBI scoring scheme to frame results in terms of divergence from the mean reference. IBI (*IndxMF_MSD*) scores were standardized to the mean and standard deviation of the index calibration reference distribution, such that a score of -1 equals 1 standard deviation below the mean. Figure 6 shows the relationship between the IBI (*IndxMF_MSD*) scores versus IBI (*Indx_MinFloor*) scores, which are based on the more traditional 100-point scoring scale.

In the MassDEP dataset, differences in IBI (*IndxMF_MSD*) scores were evident across the two regions. An IBI (*IndxMF_MSD*) score of 0 equals 71.3 in the CH versus 64.5 in the WH dataset, and a value of 1 standard deviation in the WH dataset equals 14.4, compared to 11.7 in the CH dataset. If, as a hypothetical example, -1 standard deviation was used as a threshold for an acceptable divergence from mean reference, that would correspond with an IBI (*Indx_MinFloor*) score of 50.1 in the WH dataset and a score of 59.6 in the CH dataset. The tables in Figure 6 show how standard deviations ranging from -5 to +1 correspond with IBI (*Indx_MinFloor*) scores in both regions.

Unlike the RC approach where the 10^{th} or 25^{th} percentile of reference sites are typically used, there is not a universal, well-accepted rule-of-thumb for how many standard deviations from the mean reference to use when setting thresholds. Results will vary based on factors such as how data are distributed within the dataset and policy decisions (e.g., how much deviation from reference management feels is acceptable). Nevertheless, we felt this was a useful line of evidence to consider in combination with the other approaches.



IBI (*IndxMF_MSD*)	IBI (*Indx_MinFloor*)
-5	--
-4	7.0
-3	21.3
-2	35.7
-1.5	42.9
-1	50.1
-0.5	57.3
0	64.5
+0.5	71.7
+1	78.8



IBI (*IndxMF_MSD*)	IBI (*Indx_MinFloor*)
-5	12.8
-4	24.5
-3	36.2
-2	47.9
-1.5	53.7
-1	59.6
-0.5	65.4
0	71.3
+0.5	77.1
+1	83.0

Figure 6. Relationships between the IBI (IndxMF_MSD) scores, which are framed in terms of divergence from the mean reference, versus IBI (Indx_MinFloor) scores, which are based on the more traditional 100-point scoring scale, in the Western Highlands and Central Hills datasets. The dotted lines show a hypothetical example in which an IBI threshold based on a standard deviation of -1 corresponds with IBI (Indx_MinFloor) scores of 50.1 in the Western Highlands and 59.6 in the Central Hills. There is not a universal, well-accepted rule-of-thumb for how many standard deviations from the mean reference to use when setting thresholds.

4.3 Interpolation with stressors

Interpolation with disturbance variables is another approach that can help inform potential thresholds. IBI values can be regressed on disturbance variables and regression equations can then be used to determine IBI values associated with disturbance thresholds. We do not recommend this approach as a primary line of evidence (more of a post hoc illustration than a criterion) so the analyses we performed were 'quick and dirty.' However, results could potentially be used as a line of evidence in selecting index impairment thresholds and may be useful in describing the stressors associated with impairment thresholds established by other methods.

For this exercise, we examined relationships between IBI (*Indx_MinFloor*) scores and the seven disturbance variables that were used to develop the disturbance gradient (ICI, IWI, 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) (Table 1). Development of the disturbance gradient is described in detail in the kick net IBI report (Section 3; Jessup and Stamp 2020). Each of the seven variables were scored (+3 to -3) based on the thresholds shown in Table 1, which distinguish between seven disturbance levels. Before running the regression analyses for each variable, we ran a Spearman Rank Order correlation analysis to evaluate how strongly the IBI scores were associated with the different disturbance variables. As shown in Table 4, some variables were weakly correlated to the IBI scores. To simplify the analysis, we made an arbitrary decision to only include variables that had $|r_s|$ values ≥ 0.30. This limited the CH analysis to four variables: ICI, IWI, percent urban land cover, and road density. In the WH dataset, the same four variables plus percent Agricultural land cover[6] were included (Table 4).

After the IBI scores were regressed on the selected variables, the regression equations were used to determine the IBI scores associated with each of the disturbance level thresholds. We evaluated each variable/threshold combination individually. In addition, we calculated the average IBI scores associated with each criterion level (based on the selected variables only) to give a synthesis of the possible IBI values associated with the disturbance level thresholds (Table 5). As an example of how this information could be used, one could potentially examine the mean IBI scores that correspond with the reference threshold levels used in each region. In the CH, which has higher disturbance levels than the WH (see Section 2 and Appendix A), reference sites had to meet the criteria for disturbance levels #3/4 (which corresponds with a mean IBI score of 56.8), whereas in the WH, the reference site threshold corresponded with disturbance levels #2/3 (which corresponds with mean IBI score of 56) (Table 5). Scatterplots that allow for better visualization of the relationships between the IBI scores and the disturbance variables can be found in Figure 7 (IWI/CH example) and Appendix C (all seven variables/both regions). Attachment B contains the Excel worksheets that were used to calculate the regression equations and IBI scores associated with each disturbance level threshold.

[6]In addition to having a weak correlation in the Central Hills, the two agricultural metrics were also inconsistent with expectations: they showed a positive relationship with IBI scores, such that higher IBI scores corresponded with higher agricultural land cover and fertilizer, etc. application rates (note: the agricultural metrics do not distinguish between different management practices (e.g., organic farms versus conventional); for more information on the StreamCat dataset, see Hill et al. 2016. The dam storage volume variable was not used because it was weakly correlated and also because the regression fit could have been either linear or logarithmic, and neither fit gave credible results.

One caution about this approach is that the disturbance criteria were established through professional judgment to identify a reasonable number of sites for the kick net IBI calibration, and they were not scrutinized for effects or responses. Yet, they were based on stressor variables that were approved by the technical workgroup and were documented in the kick net IBI report (Jessup and Stamp 2020). Therefore, this line of evidence has merit for associating disturbances with the indices, but the relationship is relative to the observed distribution of disturbance values, not to explicit harmful effects.

Table 4. Spearman correlation coefficients (r_s) and p-values showing the strength of associations between IBI (Indx_MinFloor) scores and the seven disturbance variables. To simplify the analysis, we made an arbitrary decision to only include variables that had $|r_s|$ values ≥ 0.30, which are shown in bold text. For more detailed information on the disturbance variables, see Table 1.

Disturbance variables	Spearman correlation coefficients (r_s)	
	Western Highlands IBI scores	Central Hills IBI scores
Index of catchment integrity (ICI)	**0.51**; $p < .001$	**0.50**; $p < .001$
Index of watershed integrity (IWI)	**0.45**; $p < .001$	**0.69**; $p < .001$
% Urban	**-0.45**; $p < .001$	**-0.71**; $p < .001$
% Hay + row crop	**-0.37**; $p < .001$	0.27; $p < .001$
Ag application rates	-0.16; p = .03	0.20; $p < .001$
Road density	**-0.49**; $p < .001$	**-0.49**; $p < .001$
Dam storage volume	-0.29**;** $p < .001$	-0.11; $p = .04$

Table 5. IBI scores were regressed on the disturbance variables, and the regression equations were used to calculate the IBI scores associated with each of the disturbance level thresholds shown in Table 1. Variables with entries in gray text had $|r_s|$ values < 0.30 and were not included in the mean IBI calculations.

Central Hills								
Metric disturbance levels		**IBI score (*Indx_MinFloor*)**						
	Mean	**ICI**	**IWI**	**Urban**	**Road density**	Hay + row crop	Ag application rates	Dam storage volume
Level #1/#2 threshold	61.9	61.3	66.5	60.1	59.8	48.1	48.3	51.1
Level #2/#3 threshold	60.3	59.5	63.5	59.5	58.6	48.7	48.6	51.1
Level #3/#4 threshold	56.8	56.1	57.4	57.7	56.0	50.2	49.7	51.0
Level #4/#5 threshold	52.4	52.7	51.4	54.7	50.9	52.8	51.4	50.4
Level #5/#6 threshold	39.2	42.4	33.2	36.6	44.5	55.3	53.2	49.6
Level #6/#7 threshold	29.8	35.5	21.1	24.6	38.1	57.9	54.9	48.2

Table 5 continued…

Western Highlands								
Metric disturbance levels		**IBI score (*Indx_MinFloor*)**						
	Mean	**ICI**	**IWI**	**Urban**	**Road density**	**Hay + row crop**	**Ag application rates**	**Dam storage volume**
Level #1/#2 threshold	58.5	61.5	60.7	55.2	56.6	58.5	53.9	53.8
Level #2/#3 threshold	56.0	59.1	57.1	51.3	54.5	57.8	53.8	53.8
Level #3/#4 threshold	50.0	54.3	50.0	39.4	50.4	55.8	53.5	53.6
Level #4/#5 threshold	41.3	49.5	42.8	19.7	42.1	52.4	53.0	52.6
Level #5/#6 threshold	27.5	35.1	21.4	0.0	31.8	49.0	52.5	51.4
Level #6/#7 threshold	20.0	25.5	7.1	0.0	21.5	45.6	52.0	49.1

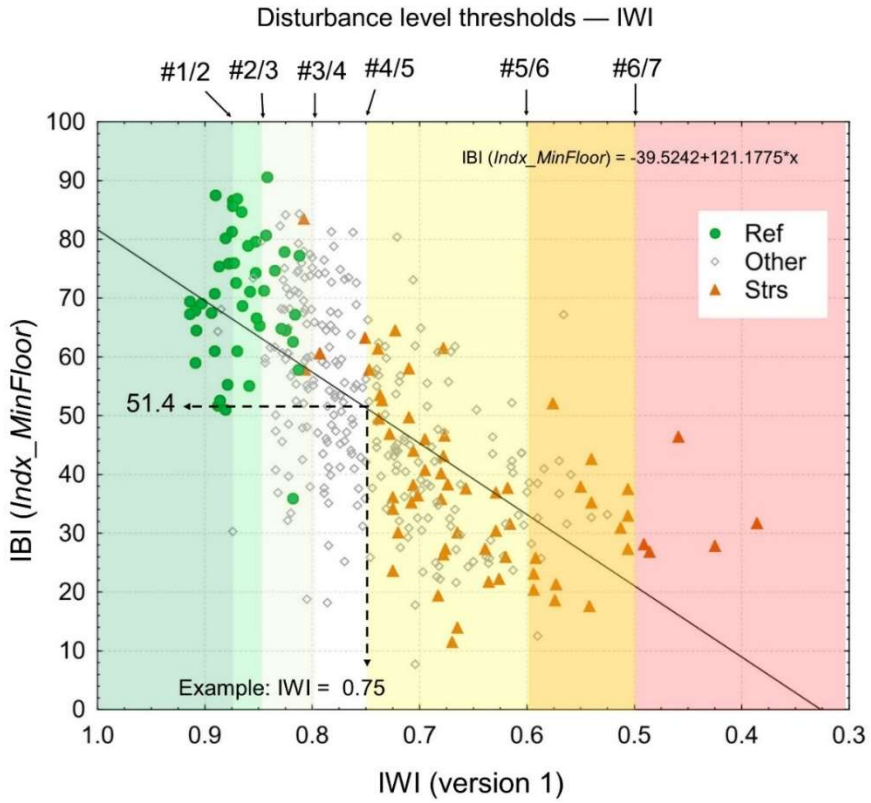


Figure 7. Scatterplot of Index of Watershed Integrity (IWI) scores (Thornbrugh et al. 2018) vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. The plot is color-coded based on the disturbance thresholds shown in Table 1. The IWI is scaled from 1 (best condition) to 0 (worst condition). The dotted lines show a hypothetical example in which an IBI threshold based on an IWI score of 0.75 corresponds with an IBI score of 51.4.

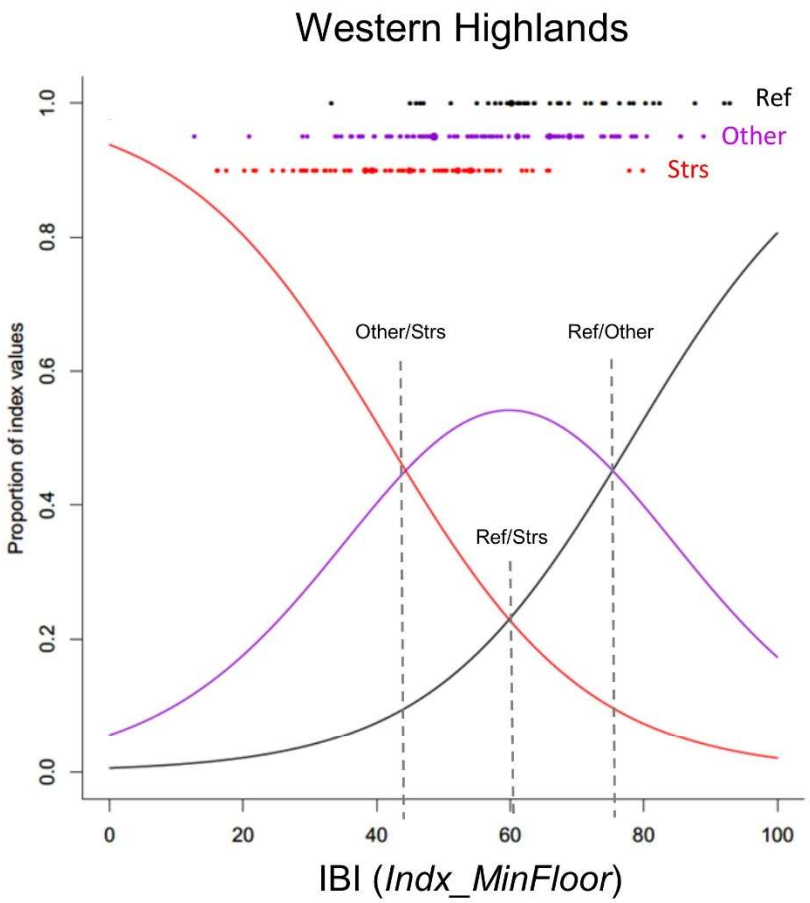
4.4 Proportional odds logistic regression

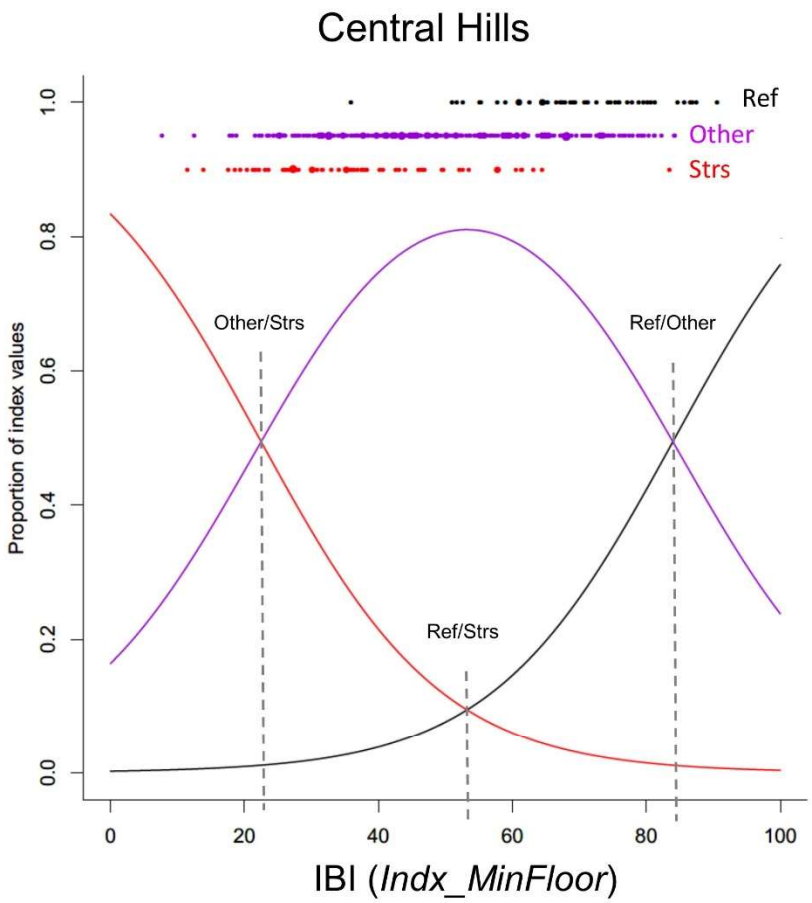
Another approach that we considered for threshold derivation was proportional odds logistic regression (POLR). POLR is useful to show points along the IBI scale at which there are equal probabilities of being in comparable disturbance categories (reference, other, stressed), thus suggesting points at which the IBI flips between categories. This can also be derived from the percentiles of the distributions (as discussed in Section 4.1), but the POLR can consider multiple categories simultaneously and can smooth the curves instead of depending on a relatively smooth and unskewed set of data in each disturbance category.

For this exercise, we used POLR to derive three points at which equal probabilities between points could be identified:

- Reference and Other
- Reference and Stressed
- Other and Stressed

Results are shown in Figure 8. This method is intended as a secondary line of evidence. More established methods, like the distribution statistics/RC approach, should be given more weight.





Intersections of Curves	IBI score (*Indx_MinFloor*)	IBI score (*IndxMF_MSD*)
Reference and Other	76	0.8
Reference and Stressed	61	-0.3
Other and Stressed	45.5	-1.4

Intersections of Curves	IBI score (*Indx_MinFloor*)	IBI score (*IndxMF_MSD*)
Reference and Other	85	1.1
Reference and Stressed	54	-1.5
Other and Stressed	23.5	-4.2

Figure 8. Proportional odds logistic regression (POLR) shows points along the IBI scale at which there are equal probabilities of being in comparable disturbance categories (reference, other, stressed).

4.5 Taxa loss

A fifth line of evidence that we considered was taxa loss, as quantified by observed over expected (O/E) ratios. This method is intended as a secondary line of evidence. More established methods, like the distribution statistics/RC approach, should be given more weight. There is not a universal, well-accepted rule-of-thumb for how much taxa loss is acceptable when setting thresholds. Results will vary based on factors such as how data are distributed within the dataset, as well as policy decisions (e.g., how much taxa loss management feels is acceptable). The O/E analyses that we performed were 'quick and dirty.' However, results could potentially be used to help inform selection of IBI thresholds and to make the thresholds more ecologically meaningful, which could make communication easier.

O/E ratios are derived from empirical models that compare the taxa expected (E) at a site without anthropogenic degradation to the taxa that are actually observed (O) (Hawkins et al. 2000). The basis for the comparison is that any differences between O and E communities reflect biological responses to the range of environmental pollutants or alterations that are intended to be evaluated (Wright 2000). O/E is easily interpreted because it simply represents the extent to which expected taxa are missing. The mean O/E ratio at reference sites is 1. An O/E ratio of 0.40 implies that, on average, 60% of the taxa are missing as a result of environmental stresses to the system. A number of states utilize O/E models for bioassessments (e.g., Utah (UT DWQ 2016), Colorado (Paul et al., 2005), Montana (Feldman, 2006; Jessup et al., 2006), and Wyoming (Hargett et al., 2007, 2012)).

For this exercise, we developed 'null' O/E models[7] for the WH and CH regions. Attachments C and D contain the O/E worksheets for the CH and WH datasets, respectively. The list of E taxa is derived based on reference datasets, which are assumed to encompass the range of ecological variability observed among streams in each region. Lists of E taxa are derived by calculating the proportion of reference samples that each taxon occurs in, and then selecting a probability of capture (Pc) limit for determining which taxa are included on the E list. The user can select whatever Pc limit they feel is most appropriate for their dataset. Selecting the Pc is a balance between being too inclusive (e.g., Pc ≥ 0 has the greatest number of E taxa, which can add variability and reduce model precision) versus being too restrictive (e.g., Pc ≥ 0.90 will result in a very small, limited list of E taxa). Several studies have found $P_C \geq 0.50$ to be a good compromise that results in a more precise index than Pc ≥ 0 (Hawkins et al. 2006 and 2000, Ostermiller and Hawkins 2004, Van Sickle et al. 2005). Pc ≥ 0.50 means that E taxa have to occur in at least 50% of the reference samples to be included on the E list.

We initially tried models based on three Pc limits (≥ 0.1, 0.25, 0.50) before deciding to focus on Pc 0.25 and 0.50. Those models were more precise (as measured by reference site O/E standard deviations) and also had better correspondence between their E taxa lists vs. the lists of tolerant/intolerant taxa that were used for kick net IBI development. The lists of E taxa are included in Attachments C & D, as are comparisons with the taxa tolerance designations used for the kick IBI calibration. Based on reference site O/E standard deviations, the WH models performed better than the CH models. The standard deviations of the two WH models were 0.19 (Pc ≥ 0.50) and 0.20 (Pc ≥ 0.25) versus the CH models, which had standard deviations of 0.38 (Pc ≥ 0.50) and 0.33 (Pc ≥ 0.25). Hawkins (2006), Hawkins et al. (2000) and Van Sickle et al. (2005) found that the best performing O/E models generally have reference site O/E standard deviations <0.20.

[7] 'null' means that the models were calibrated without a classification component; instead we used the two existing IBI regions (WH and CH).

After calculating O/E ratios for each sample, we calculated percent taxa loss based on the following formula: (1 - O/E)*100. Positive numbers mean that there is taxa loss (fewer than expected taxa are present) and negative numbers mean the number of E taxa in the sample exceed expectations. For each region, we regressed percent taxa loss against IBI scores and used the regression equations to calculate percent taxa loss associated with various IBI thresholds. As an illustration, Figure 9 shows the relationship between IBI scores and percent taxa loss in the WH based on the Pc ≥ 0.25 model. Table 6 quantifies the relationships for both IBI scoring schemes (*Indx_MinFloor* and *IndxMF_MSD*) and describes implications of decreasing IBI scores in terms of taxa loss (e.g., in the CH for the Pc ≥ 0.50 model, for each decrease of 10 IBI (*Indx_MinFloor*) points, 11.9% of expected taxa are lost). Table 7 shows the correspondence between IBI scores (*Indx_MinFloor*) ranging from 85 to 15 (in increments of 10) to percent taxa loss. To simplify results, we averaged results across the two models (Pc ≥ 0.25 and Pc ≥ 0.50). In both regions, IBI scores of 75 were roughly equal to 0% taxa loss (where Observed taxa roughly equaled Expected taxa). Full results for each individual model are provided in Attachments C and D.

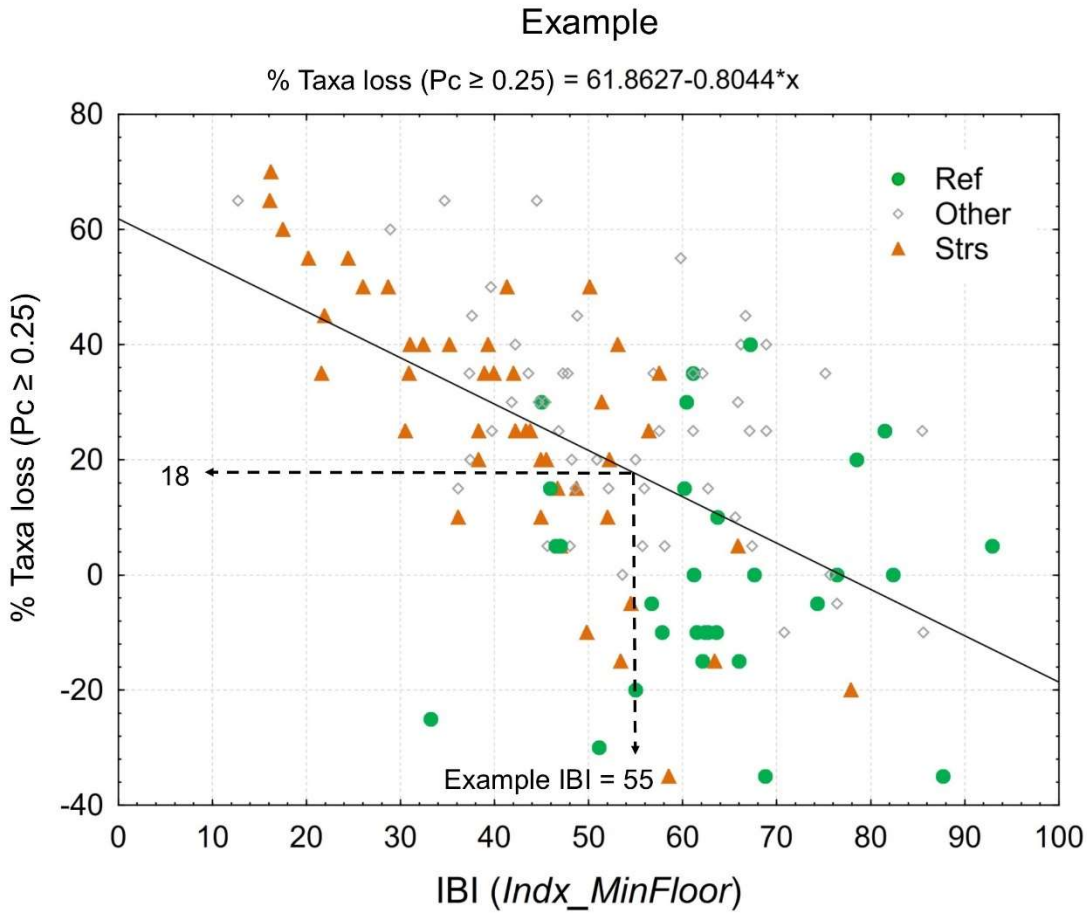


Figure 9. Example of a scatterplot of IBI (Indx_MinFloor) scores vs. % expected taxa loss. This is based on the Pc ≥ 0.25 model in the Western Highlands. Positive percentages indicate taxa loss (fewer than the expected number of taxa), while negative numbers indicate gains (more than the expected number of taxa). The dotted lines show a hypothetical example in which an IBI score of 55 corresponds with 18% taxa loss. Plots for each model and region are provided in Attachments C & D.

Table 6. Regression equations used to calculate % taxa loss for various IBI scores. Implications were characterized based on 10-point reductions in IBI scores (Indx_MinFloor), or standard deviations below mean reference value (IndxMF_MSD).

Central Hills		
IBI scoring scheme	**Taxa loss vs. IBI equation**	**Implication**
Indx_MinFloor	% taxa loss (Pc 050) = 89.6482-1.1928*x	For each 10 lower index points, 11.9% of expected taxa are lost (which equates to 1.1 taxa)
	% taxa loss (Pc 025) = 80.3846-1.0723*x	For each 10 lower index points, 10.7% of expected taxa are lost (which equates to 1.6 taxa)
IndxMF_MSD	% taxa loss (Pc 050) = 4.6457-13.9477*x	For each SD away from the mean reference index value, 13.9% of expected taxa are lost (which equates to 1.3 taxa)
	% taxa loss (Pc 025) = 3.9589-12.5455*x	For each SD away from the mean reference index value, 12.5% of expected taxa are lost (which equates to 1.8 taxa)

Western Highlands		
IBI scoring scheme	**Taxa loss vs. IBI equation**	**Implication**
Indx_MinFloor	% Taxa loss (Pc 050) = 69.0452-0.9316*IBI score	For each 10 lower index points, 9.3% of expected taxa are lost (which equates to 1.1 taxa)
	% Taxa loss (Pc 025) = 61.8627-0.8044*IBI score	For each 10 lower index points, 8.0% of expected taxa are lost (which equates to 1.6 taxa)
IndxMF_MSD	% Taxa loss (Pc 050) = 9.018-13.3345*IBI score	For each SD away from the mean reference index value, 13.3% of expected taxa are lost (which equates to 1.5 taxa)
	% Taxa loss (Pc 025) = 10.0283-11.5134*IBI score	For each SD away from the mean reference index value, 11.5% of expected taxa are lost (which equates to 2.3 taxa)

Table 7. Mean % taxa loss ± standard deviation (st dev) associated with IBI scores (Indx_MinFloor) ranging from 85 to 15 (decreasing in 10). Positive numbers indicate loss (fewer than the expected number of taxa), while negative numbers indicate gains (more than the expected number of taxa). Calculations are based on the mean loss from the Pc 0.25 and 0.50 models. Results for each individual model are provided in Attachments C & D.

IBI score (*Indx_MinFloor*)	Mean % taxa loss ± st dev	
	Western Highlands	Central Hills
85	-8.3 ± 2.6	-11.3 ± 0.7
75	0.4 ± 1.7	0.1 ± 0.2
65	9.0 ± 0.8	11.4 ± 1.0
55	17.7 ± 0.1	22.7 ± 1.9
45	26.4 ± 1.0	34.1 ± 2.7
35	35.1 ± 1.9	45.4 ± 3.6
25	43.8 ± 2.8	56.7 ± 4.4
15	52.4 ± 3.7	68 ± 5.3

5 Combining multiple lines of evidence

In Section 4, we laid out multiple lines of evidence that can be used to help inform derivation of thresholds. Here we explore different ways of combining the various lines of evidence to see if they converge on particular IBI scores for the four biological condition categories (Exceptional Condition, Satisfactory Condition, Moderately Degraded Condition, and Severely Degraded Condition). As a starting point, we created Cumulative Distribution Plots (CDFs) to see how the various thresholds (derived from the methods described in Section 4) are apportioned across the IBI (*Indx_MinFloor*) scale in the CH and WH datasets. Results, which are shown in Figure 10, show a convergence of potential Satisfactory Condition thresholds in the 50-60 scoring range and a potential breakpoint for Exceptional Condition around a score of 75 in both classes. For the Severely Degraded Condition threshold, an IBI score in the 30-40 range appears to capture a potential breakpoint in both classes. Figure 11 shows an example plot in which thresholds of 75/55/35 were applied to the CH and WH datasets. The figure contains summary tables showing Type I and II error rates and the number and percentages of samples that fall above or below a threshold of 55.

Table 8 summarizes potential ways to use the various lines of evidence to rationalize and frame decisions on setting thresholds. For example, of the different lines of evidence described in Section 4, the distribution statistics/balancing error approach is most commonly used to set biocriteria thresholds, so one may decide to weight that method more heavily than the other approaches, which tend to be more exploratory (but which nevertheless still provide valuable supplemental/secondary information). The best threshold derivation method will depend on the goals and priorities of the organization, as well as ecosystem considerations, such as acceptable deviations from reference conditions and non-actionable taxa loss rates from stream assemblages.

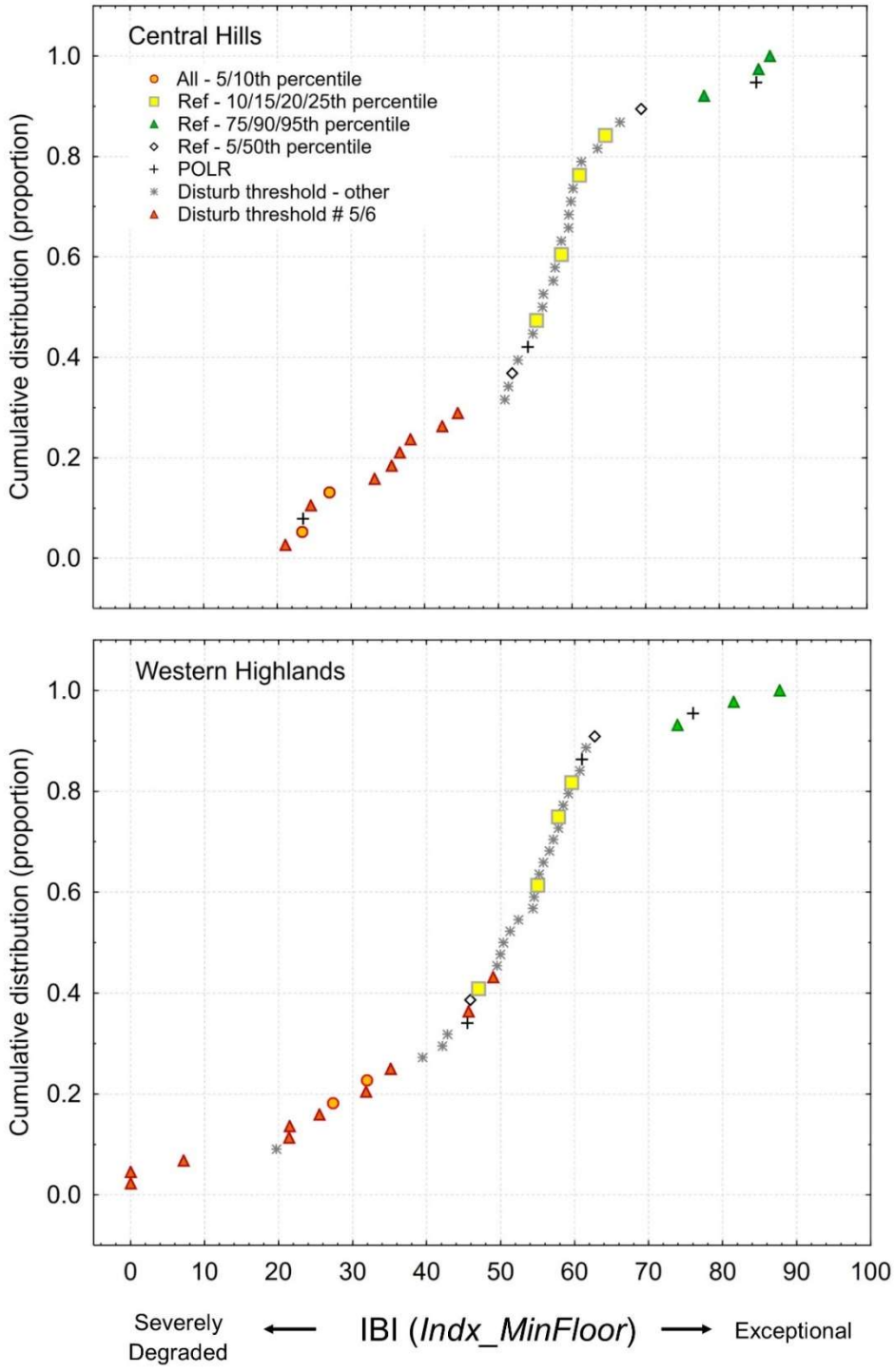


Figure 10. Cumulative distribution function (CDF) plots showing how the various thresholds that were generated are apportioned across the IBI (Indx_MinFloor) scale in the Central Hills and Western Highlands. To illustrate how to interpret the plot, in the CH, 50% of the thresholds are at or below an IBI score of 55; in the WH, 61% of the thresholds are at or below an IBI score of 55.

Table 8. Potential ways to use the various lines of evidence to rationalize and frame decisions on where to set thresholds. CH = Central Hills; WH = Western Highlands; CDF = Cumulative Distribution Function.

Method	Possible Thresholds		
	Severely Degraded	Moderately Degraded/Satisfactory	Exceptional
Distribution statistics (Section 4.1)	10th percentile of all sites in combination with interpolation with stressors	10th or 25th of reference are most commonly used (see Section 4.1).	The 75th percentile of reference appears to capture the CDF breakpoint fairly well in both classes (Fig 10).
Balancing Type I and II error (Section 4.1)	Not relevant	Choose the threshold that provides the best balance (has the smallest difference between Type I and II errors). How confident are you in your reference and stressed datasets? Are you willing to accept more of one type of error than the other (e.g., you could choose to accept higher Type II error in the WH if you are less confidence in the stressed dataset)?	Not relevant
Standard deviation from reference (Section 4.2)	CH – use a threshold that is ~ -3 standard deviations below reference	CH – use a threshold that falls between -1 and -1.5 standard deviations from reference	Use a threshold that is greater than 0 (the mean of reference)
	WH – use a threshold that ~ -2 standard deviations below reference	WH – use a threshold that falls between -0.5 and -1 standard deviations from reference	
Interpolation with stressors (Section 4.3)	CH – use a threshold that roughly corresponds to the mean of disturbance thresholds #5 & 6	CH – use a threshold that roughly corresponds to the mean of disturbance thresholds #3 & 4	> the mean disturbance threshold #1
	WH – use a threshold that roughly corresponds to the mean of disturbance thresholds #4 & 5	WH – use a threshold that roughly corresponds to the mean of disturbance thresholds #2 & 3 (stricter to account for the lower levels of disturbance in the WH dataset).	
Proportional odds logistic regression (Section 4.4)	Strs/Other intersection	Ref/Strs intersection	Ref/Other intersection
Taxa loss (Section 4.5)	CH – 45% loss or greater	How many taxa are you willing to lose from a stream assemblage? Could consider the variability in taxa loss in the reference dataset (as measured by the standard deviation) to help inform this decision.	≥ 0% taxa loss (number of taxa exceed expectations)
	WH - 35% or greater (stricter to account for the lower levels of disturbance in the WH dataset).		

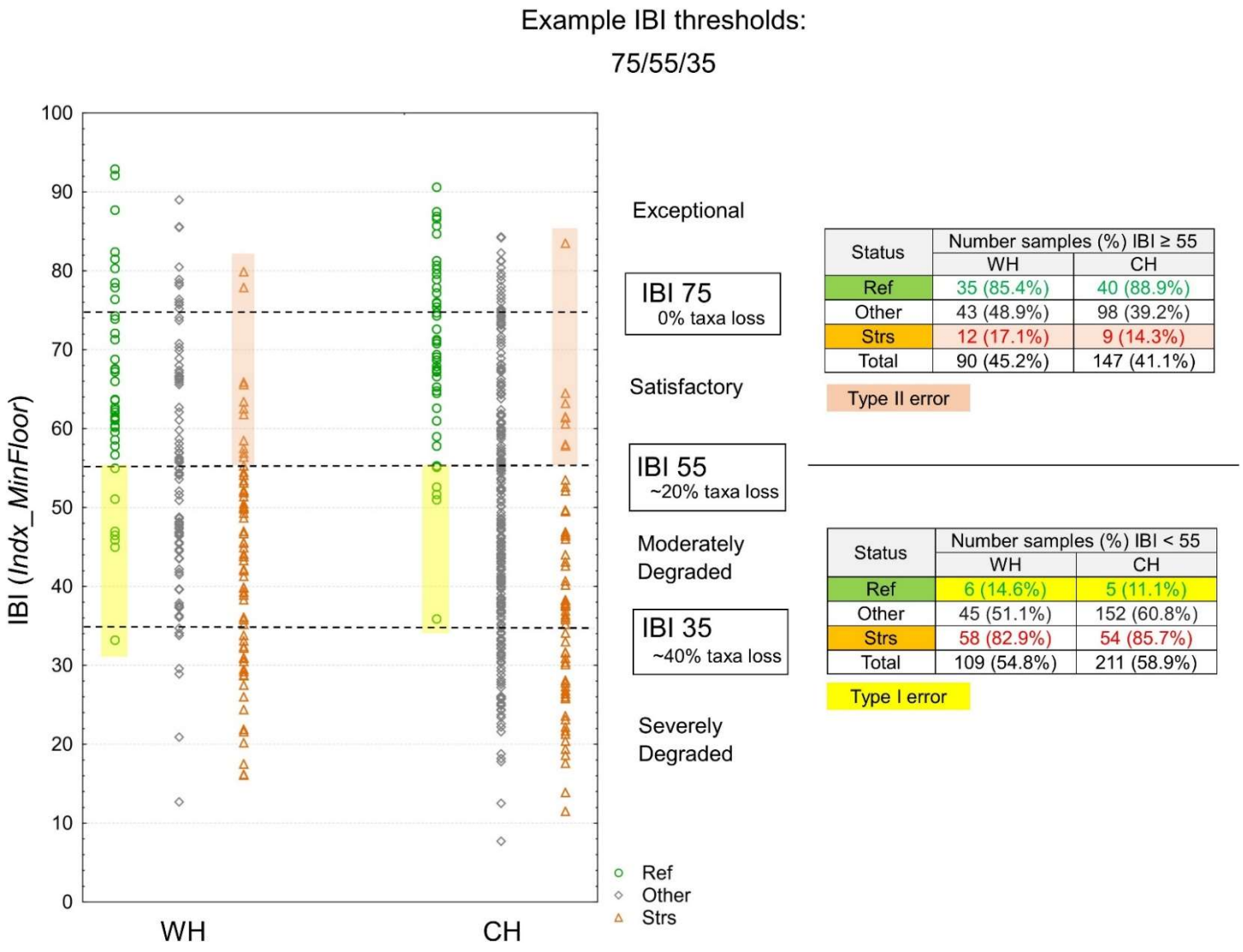


Figure 11. An example scenario in which thresholds of 75 (Exceptional Condition), 55 (Satisfactory/Moderately Degraded Condition) and 35 (Severely Degraded Condition) were selected.

6 References

Barbour, M. T., J. Gerritsen, G.E. Griffith, R. Frydenborg, E. McCarron, J.S. White, and M.L. Bastian. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society 15(2):185-211.

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.

DeShon, J.E. 1995. Development and Application of the Invertebrate Community Index (ICI). In: Davis, W.S. and Simon, T.P., Eds., Biological Assessment and Criteria—Tools for Water Resource Planning and Decision Making, Lewis Publ., Boca Raton, 217-244.

Feldman, D. 2006. A Report to the DEQ Water Quality Planning Bureau on the Proper Interpretation of Two Recently Developed Bioassessment Models. Helena, Montana: Montana Department of Environmental Quality.

Gibson G. R., M. Barbour, J. B. Stribling, J. Gerritsen & J. R. Karr. 1996. Biological criteria: technical guidance for streams and rivers - revised edition. EPA 822-B-96-001. U.S. Environmental Protection Agency, Washington, D.C.

Hargett, E.G., J.R. ZumBerge, C.P. Hawkins, and J.R. Olson. 2007 Development of a RIVPACS-type predictive model for bioassessment of wadeable streams in Wyoming. Ecological Indicators 7:807-826.

Hargett, E.G. 2012. Assessment of aquatic biological condition using WY RIVPACS with comparisons to the Wyoming Stream Integrity Index (WSII). Document #12-0151. Available online: http://deq.wyoming.gov/media/attachments/Water%20Quality/Surface%20Water%20Monitoring/Publications/wqd-wpp-monitoring_Assessment-of-Aquatic-Biological-Condition-Using-WY-RIVPACS-With-Comparisons-to-the-Wyoming-Stream-Integrity-Index.pdf

Hawkins, C.P., R.H. Norris, J.N. Hogue, and J.W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10:1456-1477.

Hawkins, C.P., 2006. Quantifying biological integrity by taxonomic completeness: evaluation of a potential indicator for use in regional and global-scale assessments. Ecological Applications 16:1277-1294.

Hughes R. M., D. P. Larsen & J. M. Omernik. 1986. Regional reference sites: a method for assessing stream potentials. Environmental Management 10: 629-635.

Jessup, B., C.P. Hawkins, and J. Stribling. 2006. Biological Indicators of Stream Condition in Montana Using Benthic Macroinvertebrates. Tetra Tech. Technical report prepared for the Montana Department of Environmental Quality, Helena, Montana

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.

Massachusetts Division of Water Pollution Control. 2013. 314 CMR 4.00: Massachusetts Surface Water Quality Standards. Available online: https://www.mass.gov/files/documents/2016/11/nv/314cmr04.pdf

Ohio EPA. 1987. Biological criteria for the protection of aquatic life: Volume I: The role of biological data in water quality assessment. Ohio EPA division of Water Quality Planning and Assessment, Columbus, OH.

Ohio EPA. 1989. Biological Criteria for the Protection of Aquatic Life: Vol. III. Standardized Biological Field Sampling and Laboratory Methods for Assessing Fish and Macroinvertebrate Communities. Ohio EPA division of Water Quality Planning and Assessment, Columbus, OH.

Ostermiller, J.D. and C.P. Hawkins. 2004. Effects of sampling error on bioassessments of stream ecosystems: application of RIVPACS-type models. Journal of the North American Benthological Society 23:363-382.

Paul, M.J., J. Gerritsen, C.P. Hawkins, and E. Leppo. 2005. Development of biological assessment tools for Colorado. Tetra Tech, Inc. 400 Red Brook Boulevard, Suite 200, Owings Mills, MD, 21117

Roth, N.E., M.T. Southerland, J.C. Chaillou, J.H. Vølstad, S.B. Weisberg, H.T. Wilson, D.G. Heimbuch, J.C. Seibel. 1997. Maryland Biological Stream Survey: Ecological status of non-tidal streams in six basins sampled in 1995. Maryland Department of Natural Resources, Chesapeake Bay and Watershed Programs, Monitoring and Non-tidal Assessment, Annapolis, Maryland. CBWP-MANTA-EA-97-2.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting expectations for the ecological condition of running waters: the concept of reference condition. Ecological Applications 16:1267–1276.

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.

Utah Division of Water Quality (UT DWQ). 2019. Utah's final 2018/2020 303(d) assessment methods. Salt Lake City, Utah. Utah Department of Environmental Quality. Available online: https://documents.deq.utah.gov/water-quality/monitoring-reporting/integrated-report/DWQ-2019-005601.pdf

Van Sickle, J., C.P. Hawkins, D.P. Larsen, and A.T. Herlihy. 2005. A null model for the expected macroinvertebrate assemblage in streams. Journal of the North American Benthological Society 24:178-191.

Wright, J.F., D.W. Sutcliffe, and M.T. Furse. 2000. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association. Ambleside Cumbria, UK.

Yoder, C. O., & Rankin, E. T. 1995. Biological criteria program development and implementation in Ohio. In W. S. Davis & T. P. Simon (Eds.), Biological assessment and criteria: tools for water resource planning and decision making (pp. 109–144). Boca Raton: Lewis Publishers.

Appendix A. Comparison of disturbance levels in reference and stressed sites in the Central Hills vs Western Highlands

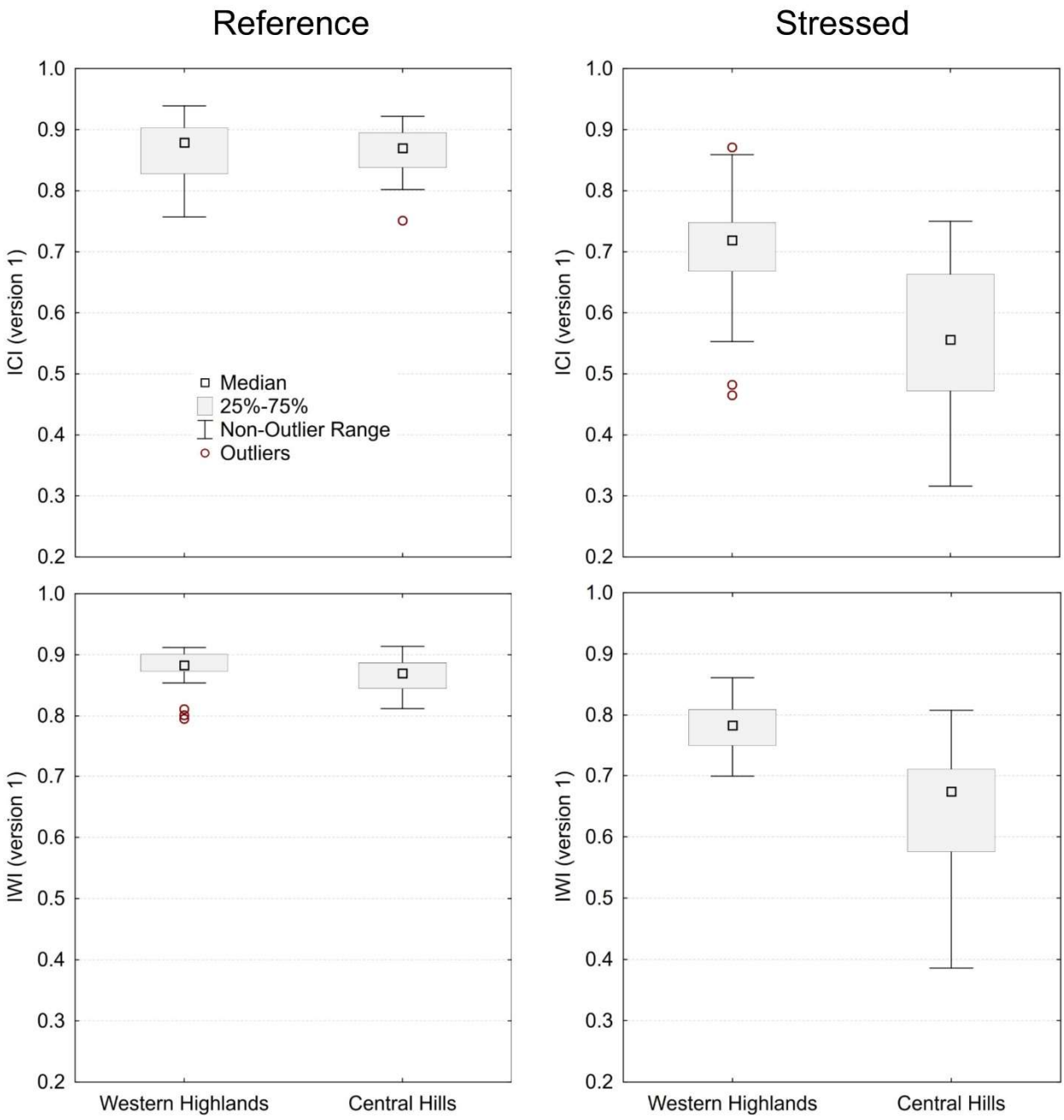


Figure A1. Box plots showing distributions of ICI and IWI scores in the reference and stressed datasets in the Central Hills and Western Highlands.

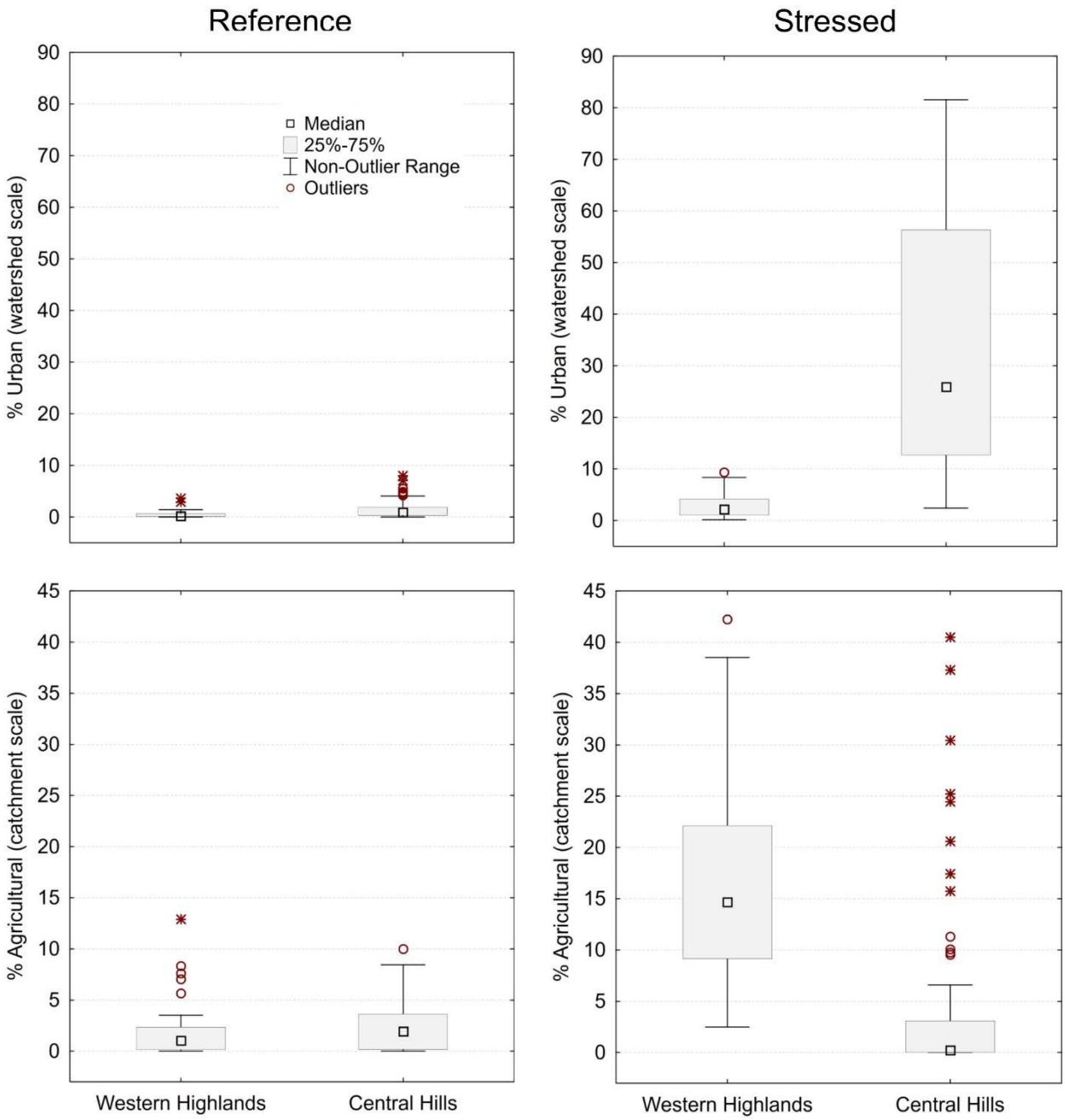


Figure A2. Box plots showing distributions of % urban (watershed scale; NLCD 2011) and % agricultural lands use (catchment scale; NLCD 2011) in the reference and stressed datasets in the Central Hills and Western Highlands.

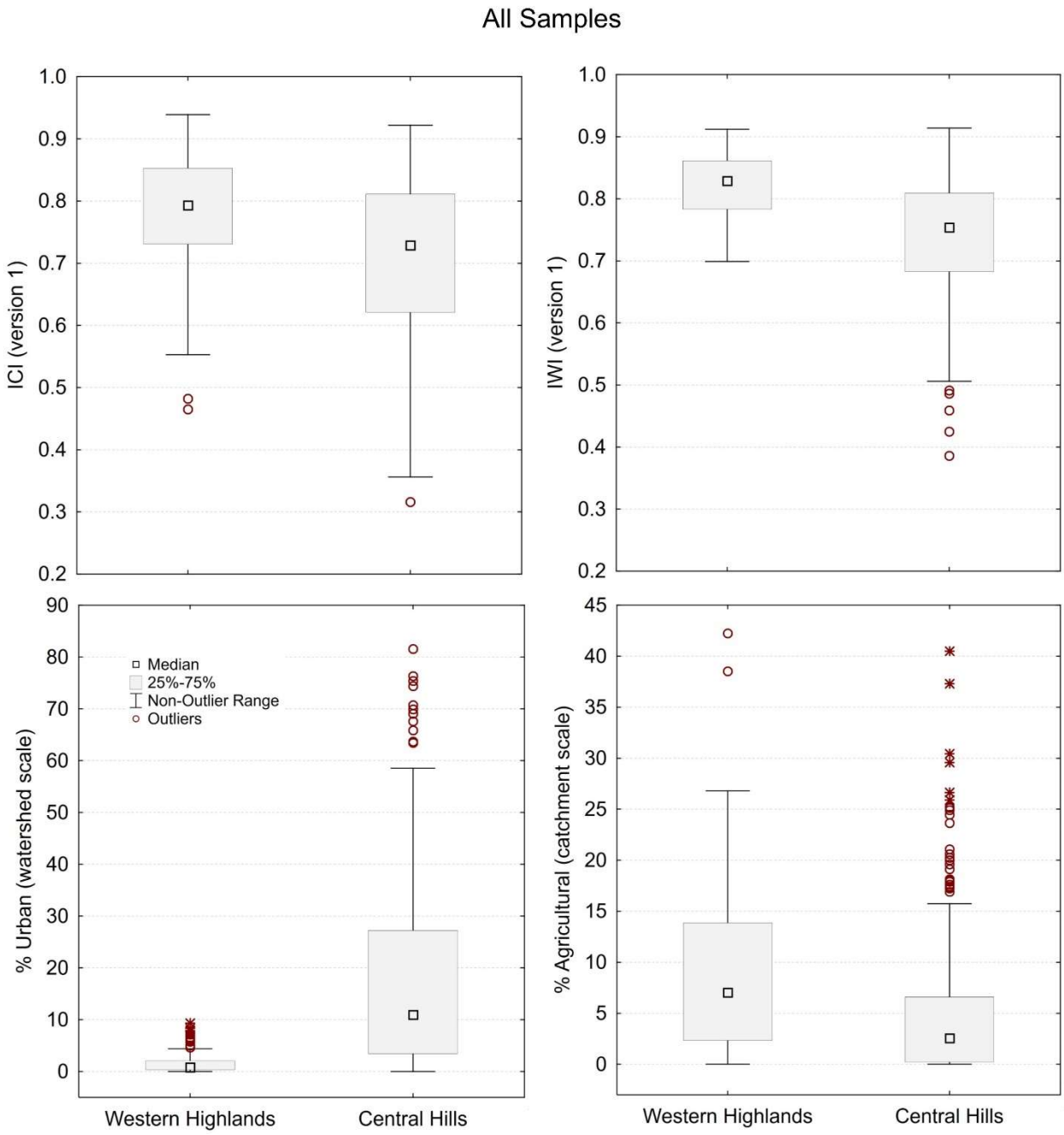
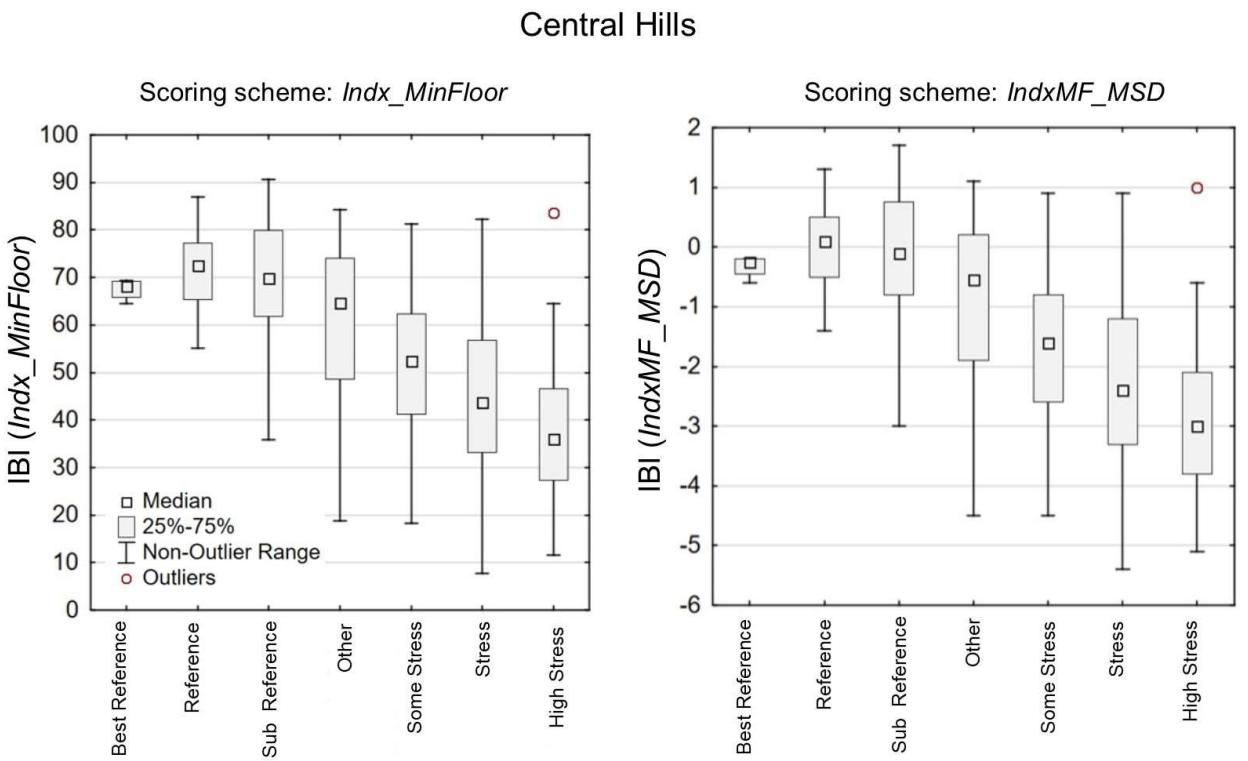


Figure A3. Box plots showing distributions of ICI and IWI scores, percent urban and percent agricultural land cover in all samples in the Central Hills and Western Highlands datasets.

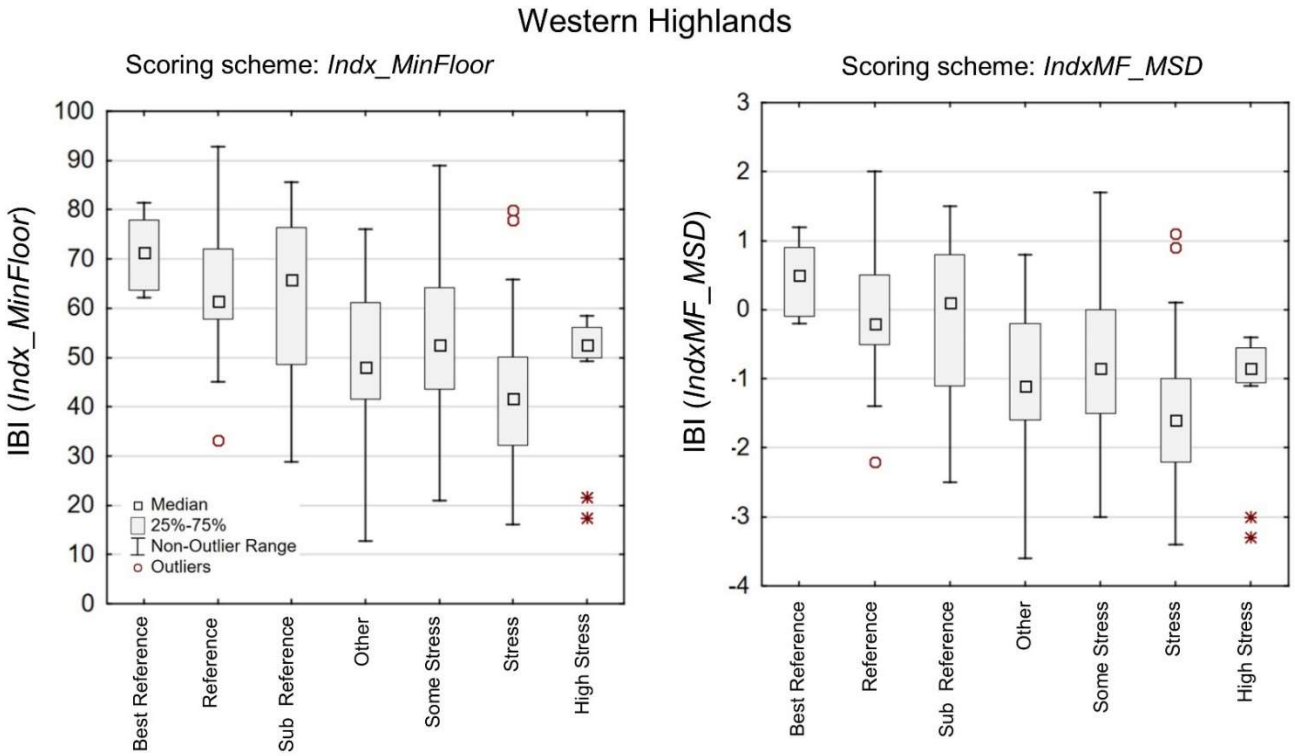
Appendix B. Additional distribution statistics



Disturb category	Count	IBI (*Indx_MinFloor*)	
		Mean	StDev
Best Reference	4	67.6	2.3
Reference	13	71.5	9.6
Sub Reference	28	69.5	13.0
Other	26	61.3	16.4
Some Stress	89	52.3	14.5
Stress	135	45.9	16.1
High Stress	63	37.7	14.4

Disturb category	Count	IBI (*IndxMF_MSD*)	
		Mean	StDev
Best Reference	4	-0.33	0.19
Reference	13	0.02	0.82
Sub Reference	28	-0.15	1.12
Other	26	-0.85	1.41
Some Stress	89	-1.62	1.25
Stress	135	-2.17	1.37
High Stress	63	-2.87	1.23

Figure B1. Summary statistics (mean and standard deviation) of IBI (Indx_MinFloor) scores and IBI (IndxMF_MSD) scores in each disturbance category in the Central Hills. Stressed sites (highlighted in orange) were derived from the High Stress category. Reference sites (highlighted in green) were comprised of sites in the Best Reference, Reference, and Sub Reference categories. For more information on disturbance categories, see Section 3 of Jessup and Stamp (2020).



Disturb category	Count	IBI (*Indx_MinFloor*)	
		Mean	StDev
Best Reference	7	71.5	7.4
Reference	34	64.0	13.5
Sub Reference	25	61.9	16.0
Other	15	50.4	15.5
Some Stress	48	53.1	13.9
Stress	58	42.3	13.9
High Stress	12	48.2	13.7

Disturb category	Count	IBI (*IndxMF_MSD*)	
		Mean	StDev
Best Reference	7	0.47	0.53
Reference	34	-0.03	0.94
SubRef	25	-0.17	1.12
Other	15	-0.99	1.08
Some Stress	48	-0.79	0.97
Stress	58	-1.54	0.96
High Stress	12	-1.14	0.96

Figure B2. Summary statistics (mean and standard deviation) of IBI (Indx_MinFloor) scores (left) and IBI (IndxMF_MSD) scores (right) in each disturbance category in the Western Highlands. Stressed sites (highlighted in orange) were derived from the High Stress and Stress categories. Reference sites (highlighted in green) were comprised of sites in the Best Reference and Reference categories. For more information on disturbance categories, see Section 3 of Jessup and Stamp (2020).

Appendix C. Interpolation with stressors

Central Hills

Scatterplots of IBI scores vs. disturbance variables

1. ICI
2. IWI
3. % Urban
4. % Hay and row crop
5. AllAgN (ag application rates)
6. Road density
7. Dam volume storage

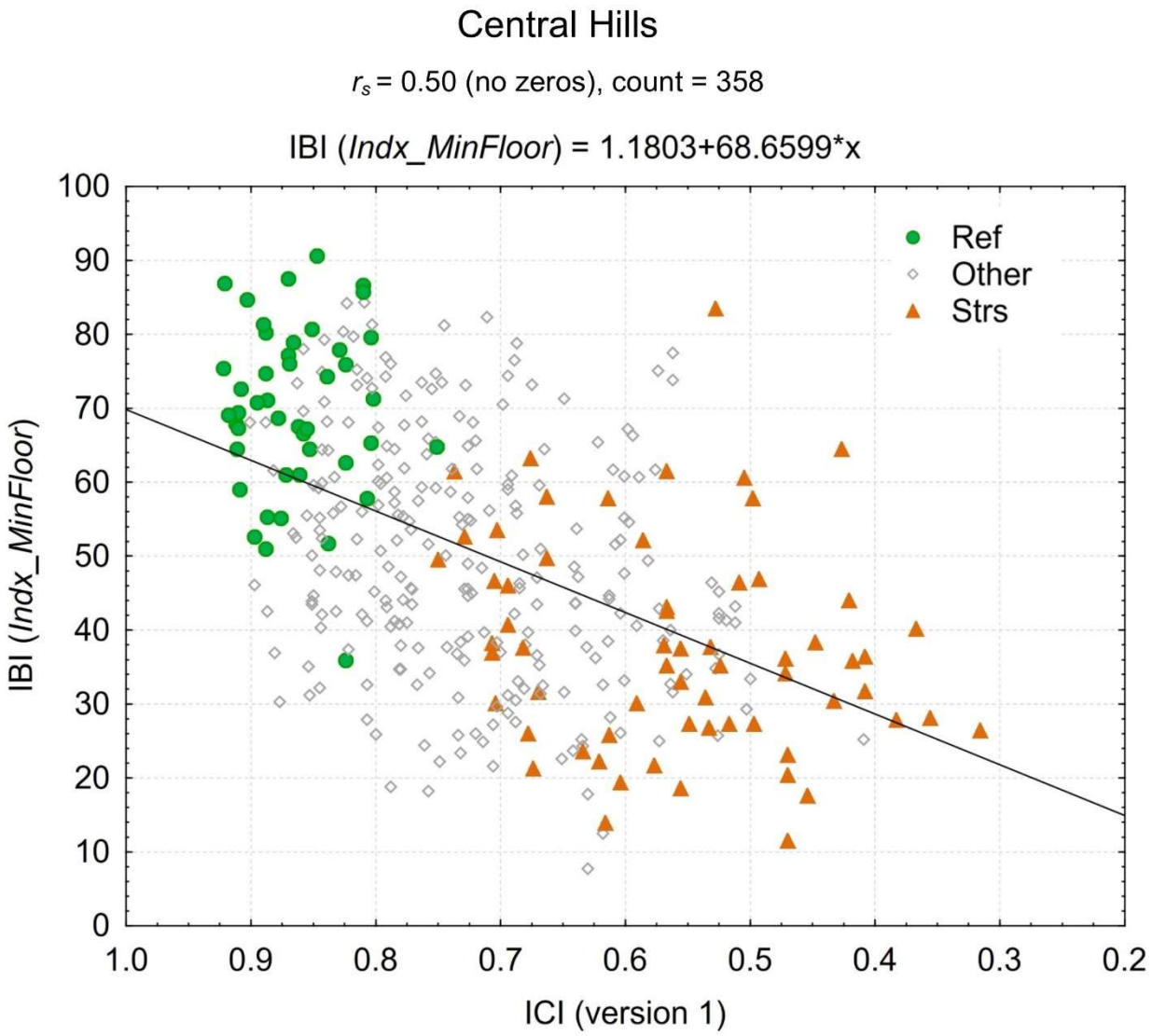


Figure C1. Scatterplot of ICI scores (Thornbrugh et al. 2018) vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. The ICI is scaled from 1 (best condition) to 0 (worst condition). All samples had ICI scores > 0. Also included (above the plot) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

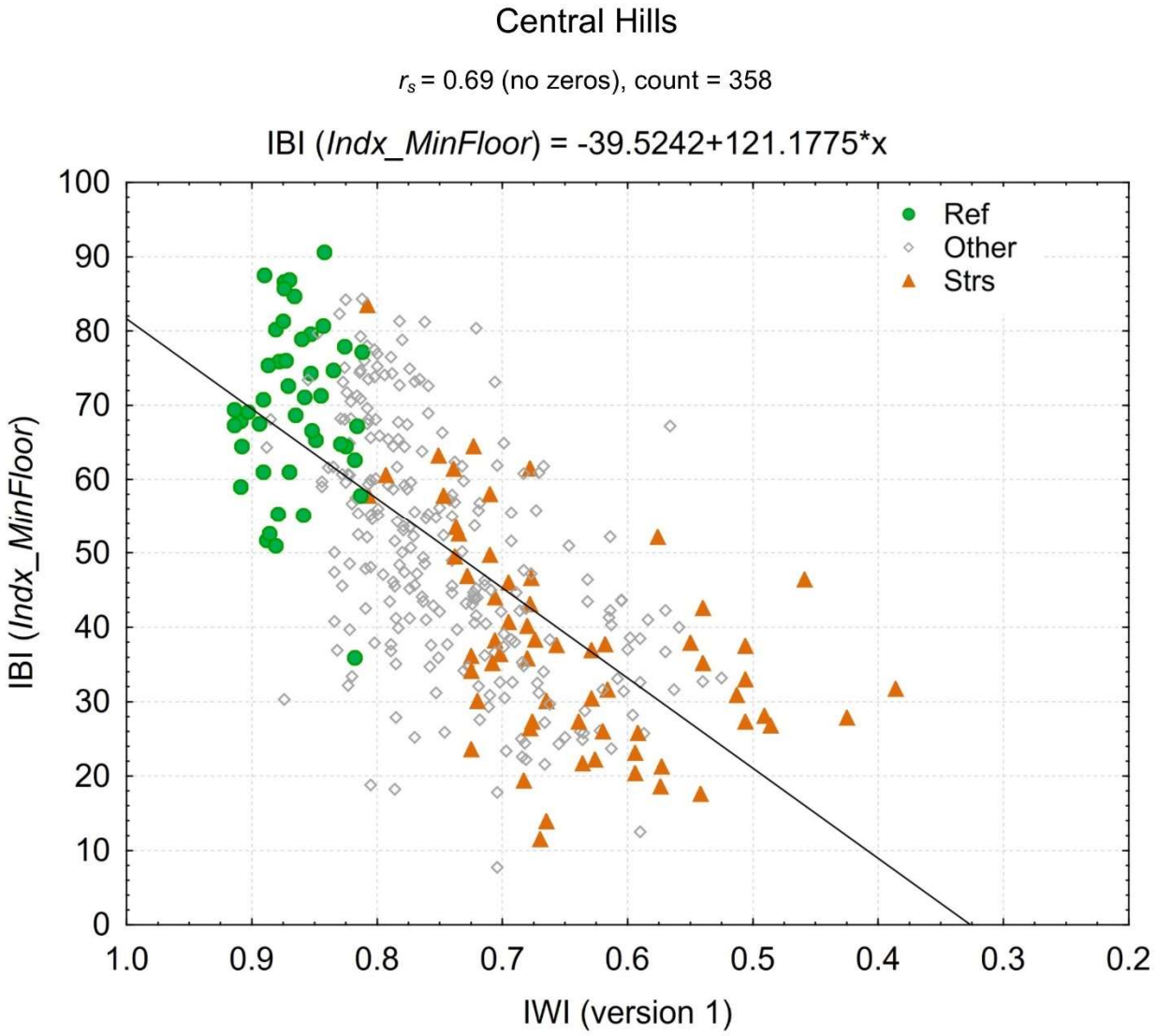


Figure C2. Scatterplot of IWI scores (Thornbrugh et al. 2018) vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. The ICI is scaled from 1 (best condition) to 0 (worst condition). All samples had IWI scores > 0 (no zeros). Also included (above the plot) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

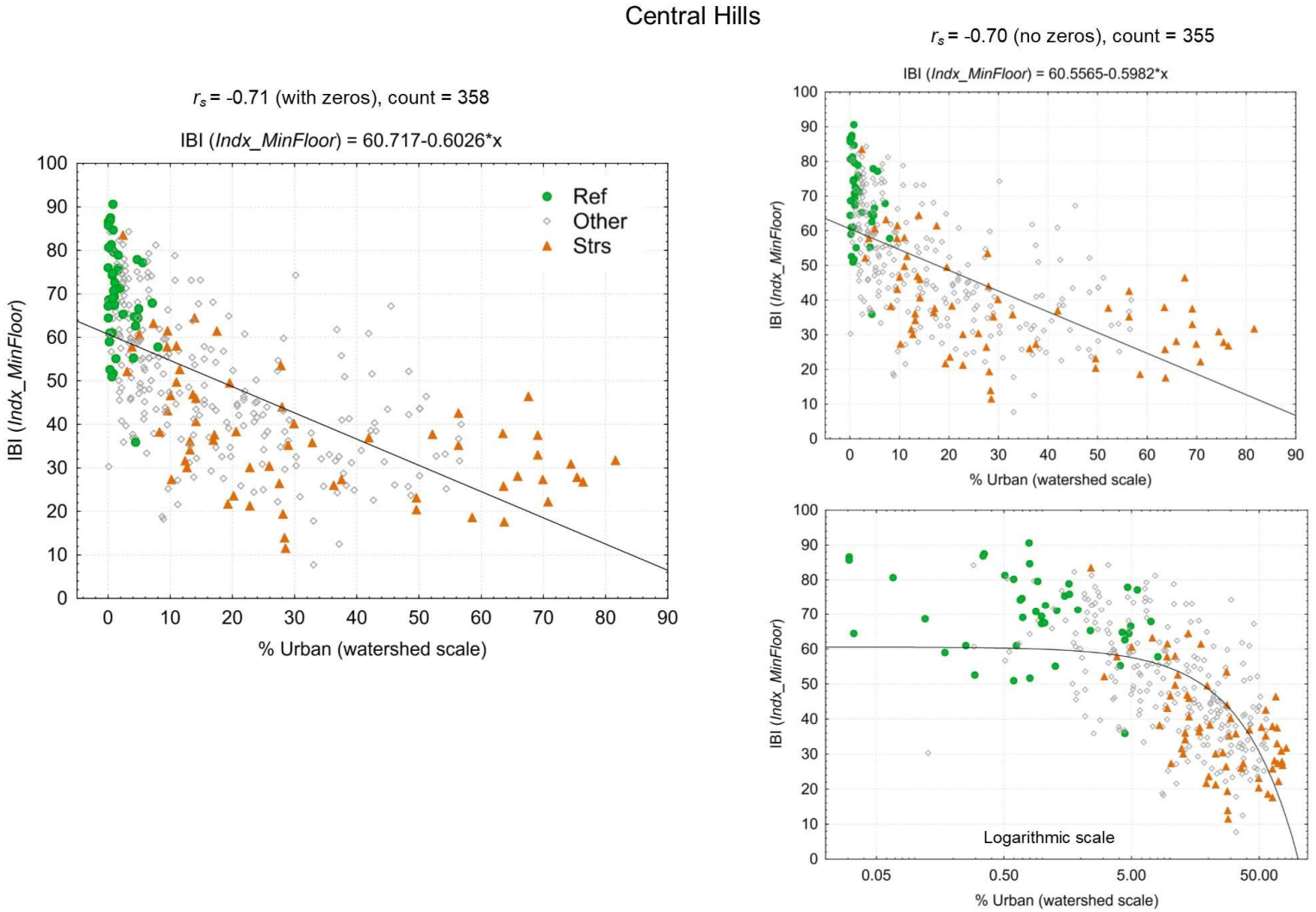


Figure C3. Scatterplot of % urban land cover (2011 NLCD, watershed-scale) vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. The plot on the left includes all samples (including the 3 samples with % urban values of 0). The plots on the right are limited to samples that have % urban values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

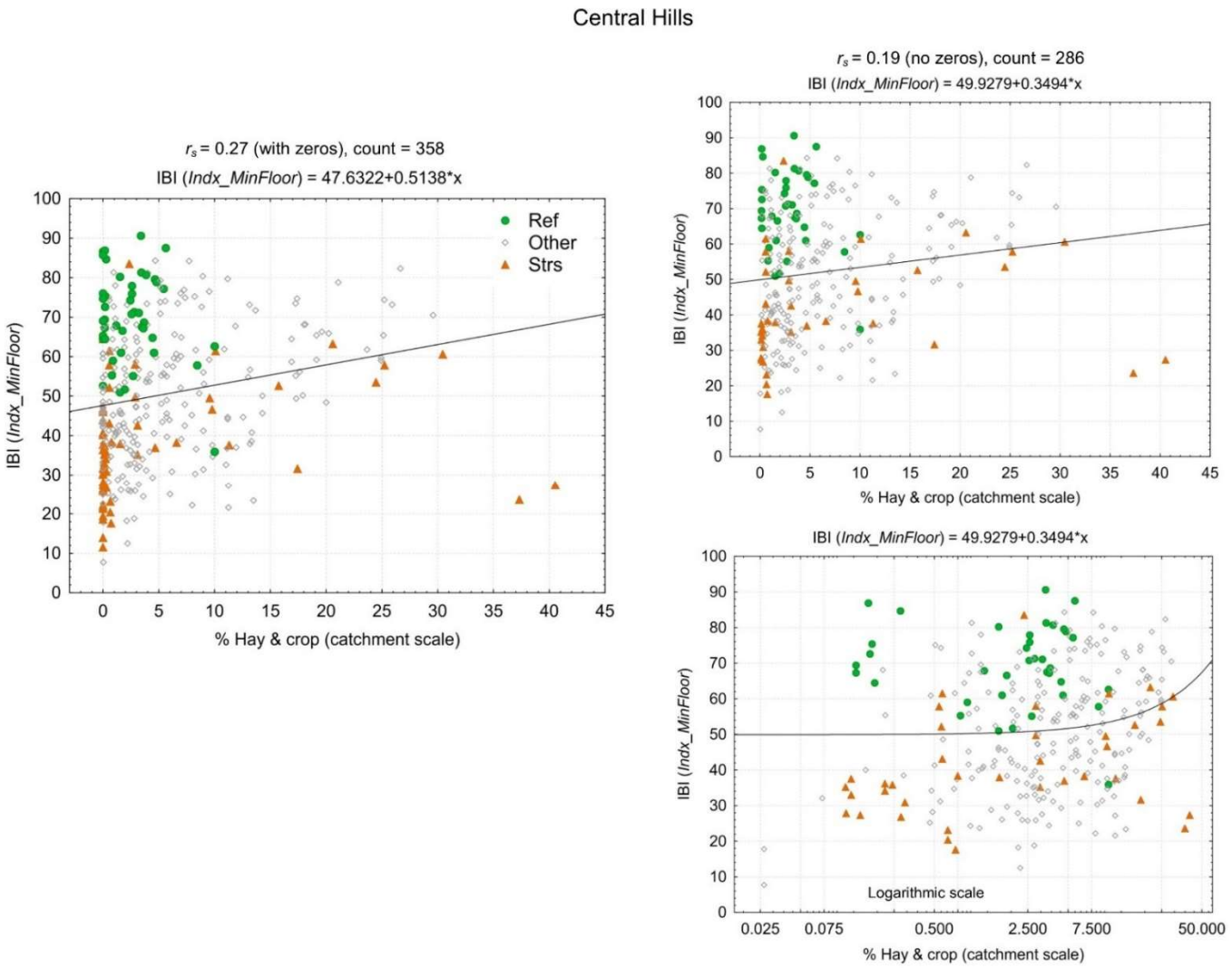


Figure C4. Scatterplot of % hay + row crop land cover (2011 NLCD, local catchment-scale) vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. The plot on the left includes all samples (including the 72 samples with % hay + crop values of 0). The plots on the right are limited to samples that have % hay + crop values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

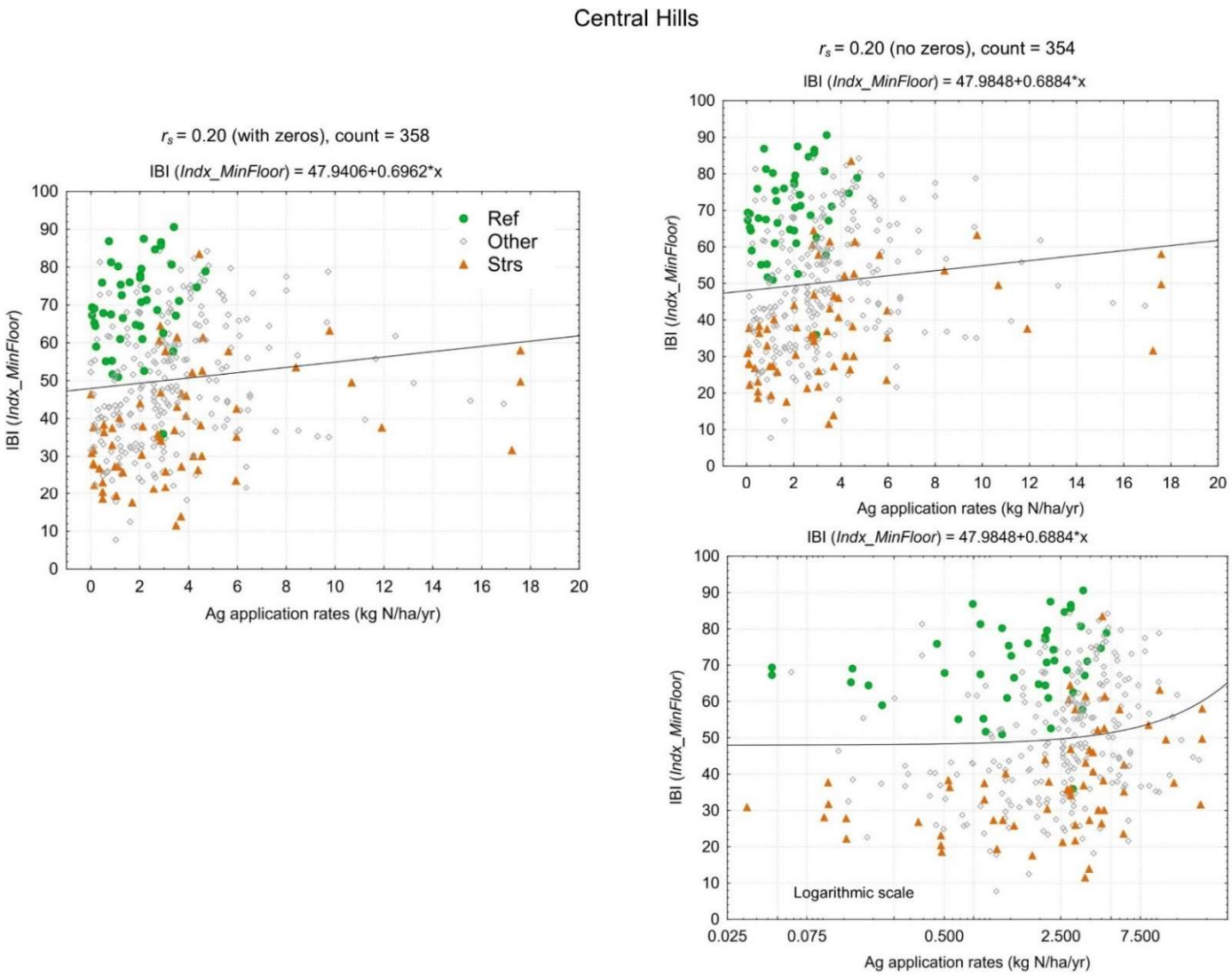


Figure C5. Scatterplot of agricultural (ag) application rates vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. The plot on the left includes all samples (including the 4 samples with ag application rate values of 0). The plots on the right are limited to samples that have % ag application rate values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

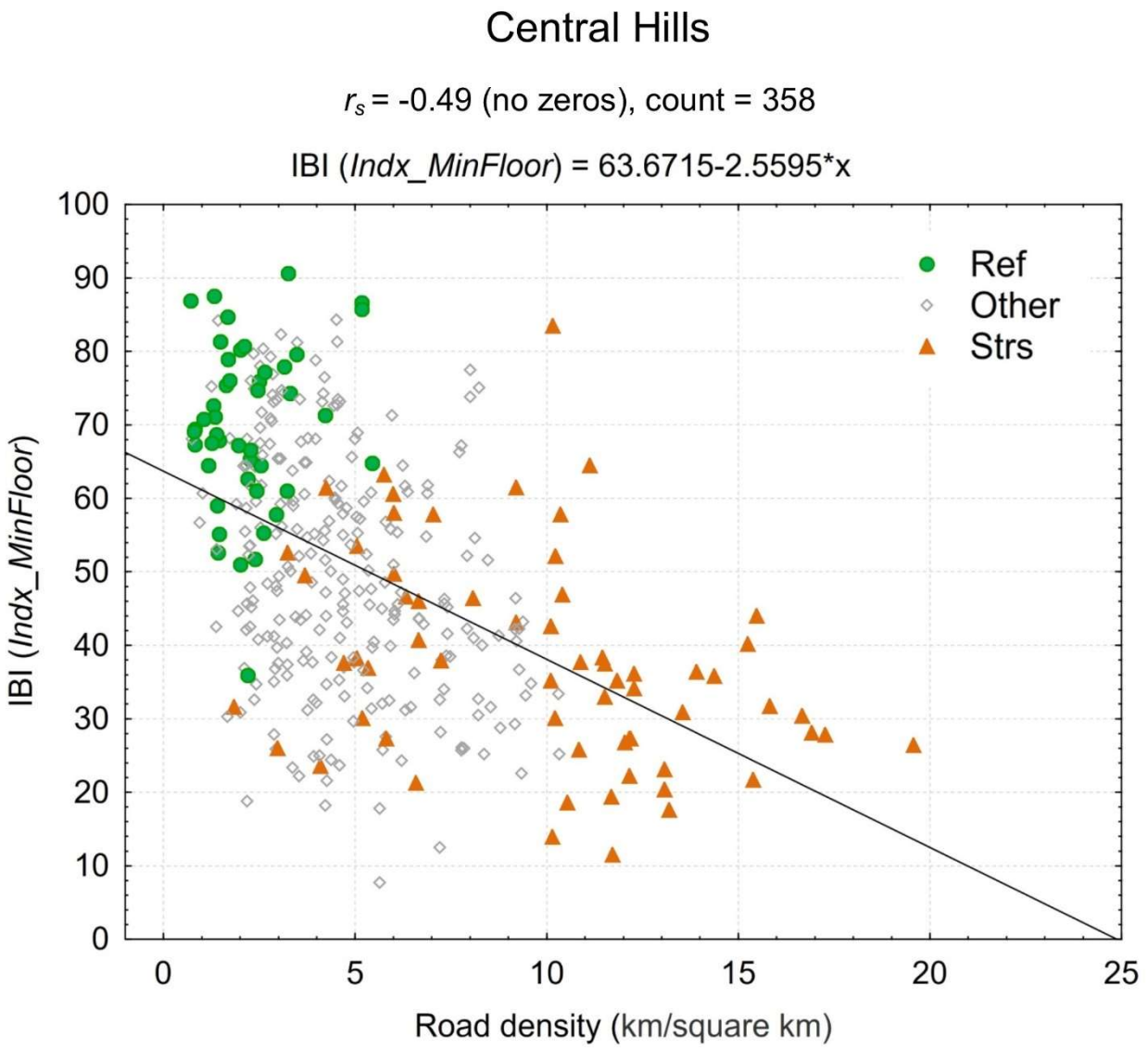


Figure C6. Scatterplot of road density vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. All samples had road density values > 0 (no zeros). Also included (above the plot) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

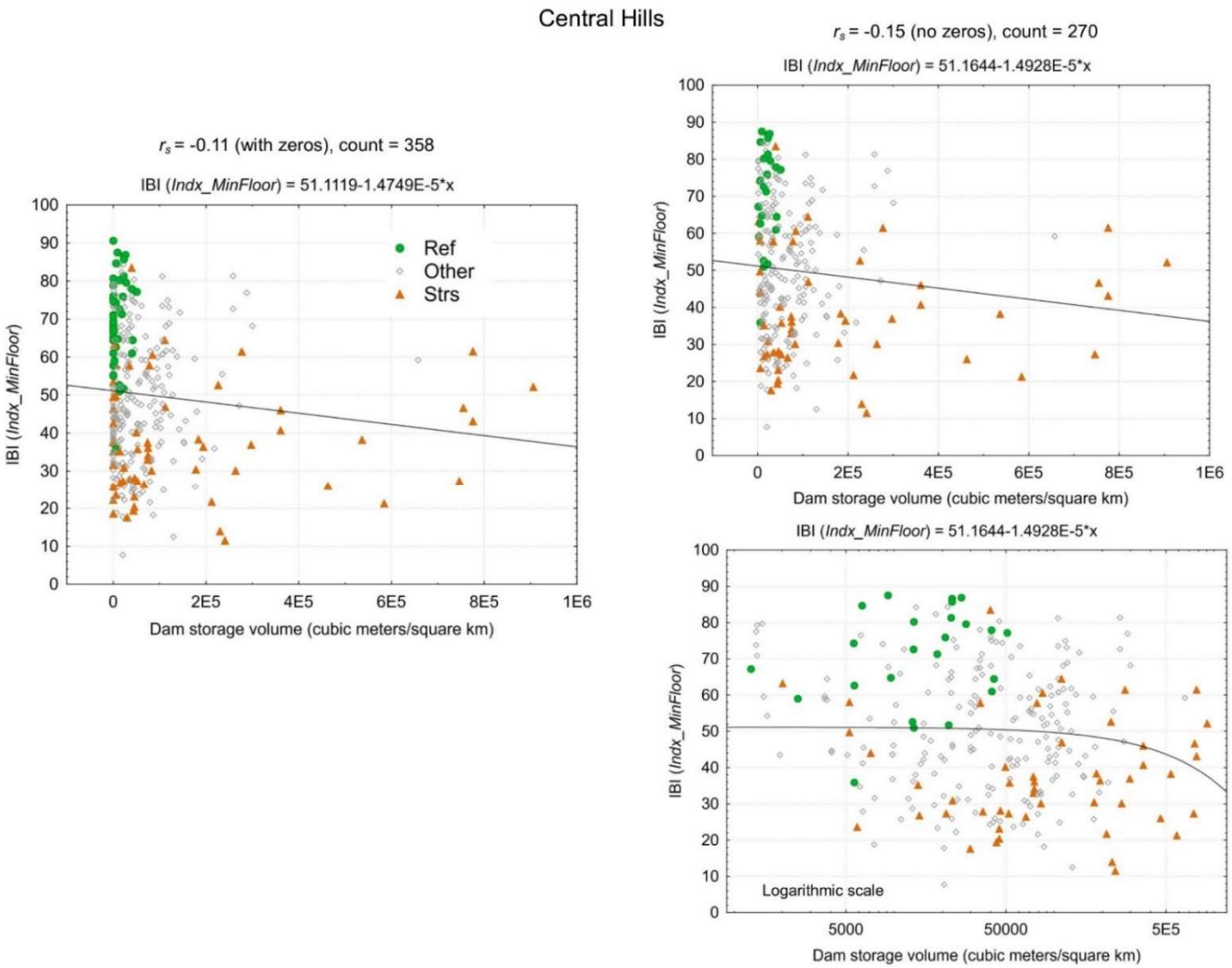


Figure C7. Scatterplot of dam storage volume vs. IBI scores (Indx_MinFloor) in the Central Hills dataset. The plot on the left includes all samples (including the 88 samples with dam storage volume values of 0). The plots on the right are limited to samples that have dam storage volume values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

Western Highlands

Scatterplots of IBI scores vs. disturbance variables

1. ICI
2. IWI
3. % Urban
4. % Hay and row crop
5. AllAgN (ag application rates)
6. Road density
7. Dam volume storage

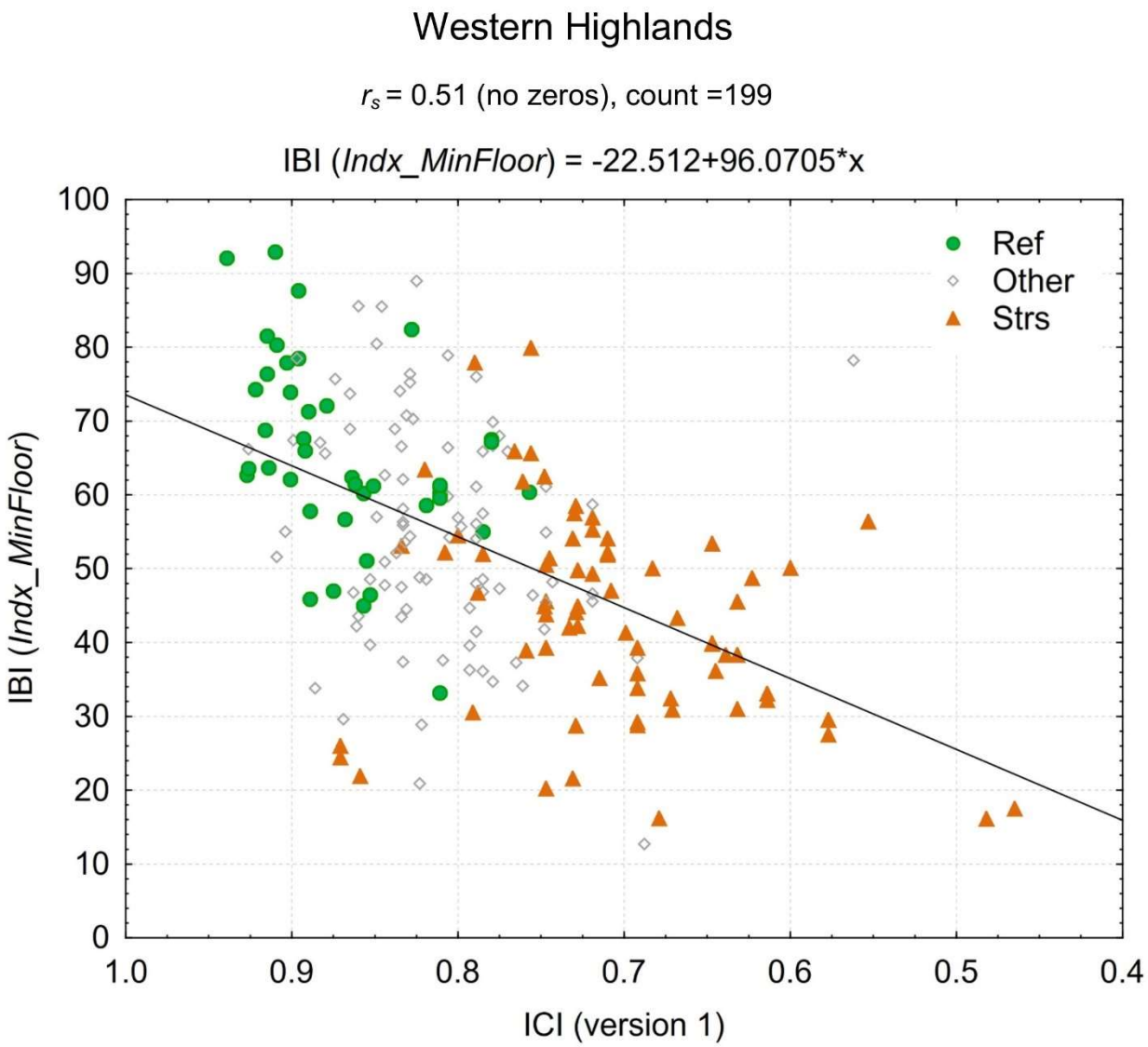


Figure C8. Scatterplot of ICI scores (Thornbrugh et al. 2018) vs. IBI scores (Indx_MinFloor) in the Western Highlands dataset. The ICI is scaled from 1 (best condition) to 0 (worst condition). All samples had ICI scores > 0. Also included (above the plot) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

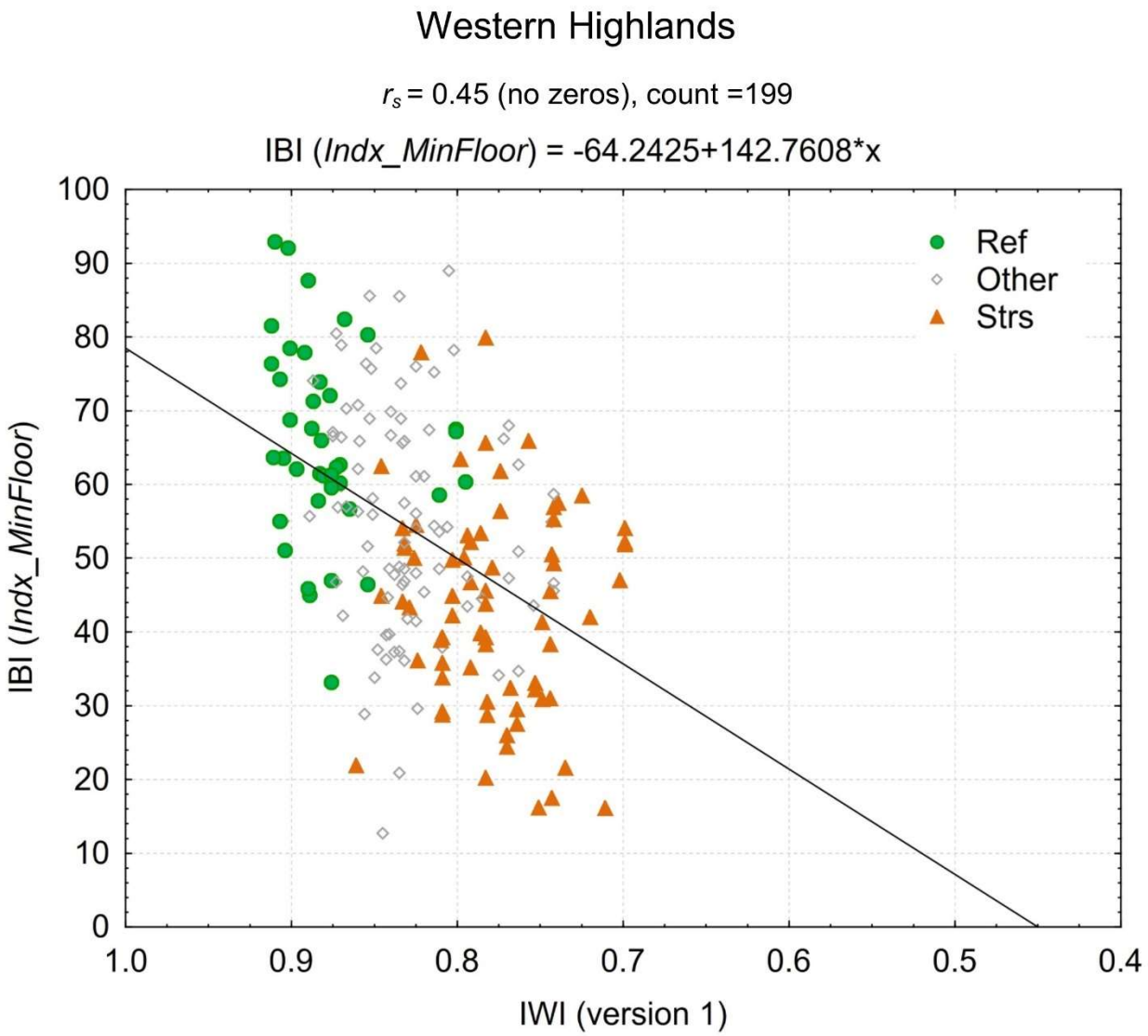


Figure C9. Scatterplot of IWI scores (Thornbrugh et al. 2018) vs. IBI scores (Indx_MinFloor) in the Western Highlands dataset. The ICI is scaled from 1 (best condition) to 0 (worst condition). All samples had IWI scores > 0 (no zeros). Also included (above the plot) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

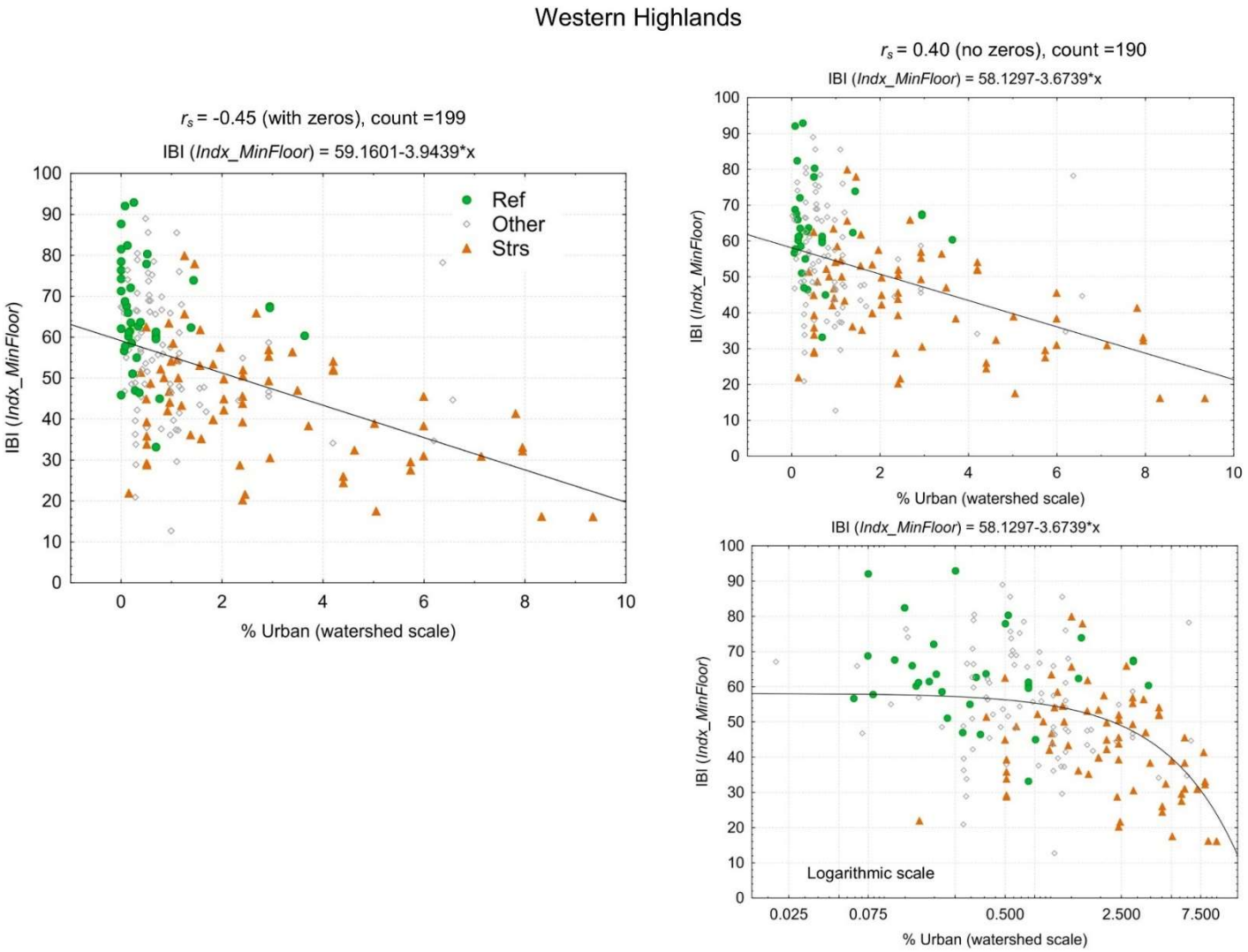


Figure C10. Scatterplot of % urban land cover (2011 NLCD, watershed-scale) vs. IBI scores (Indx_MinFloor) in the Western Highlands dataset. The plot on the left includes all samples (including the 9 samples with % urban values of 0). The plots on the right are limited to samples that have % urban values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

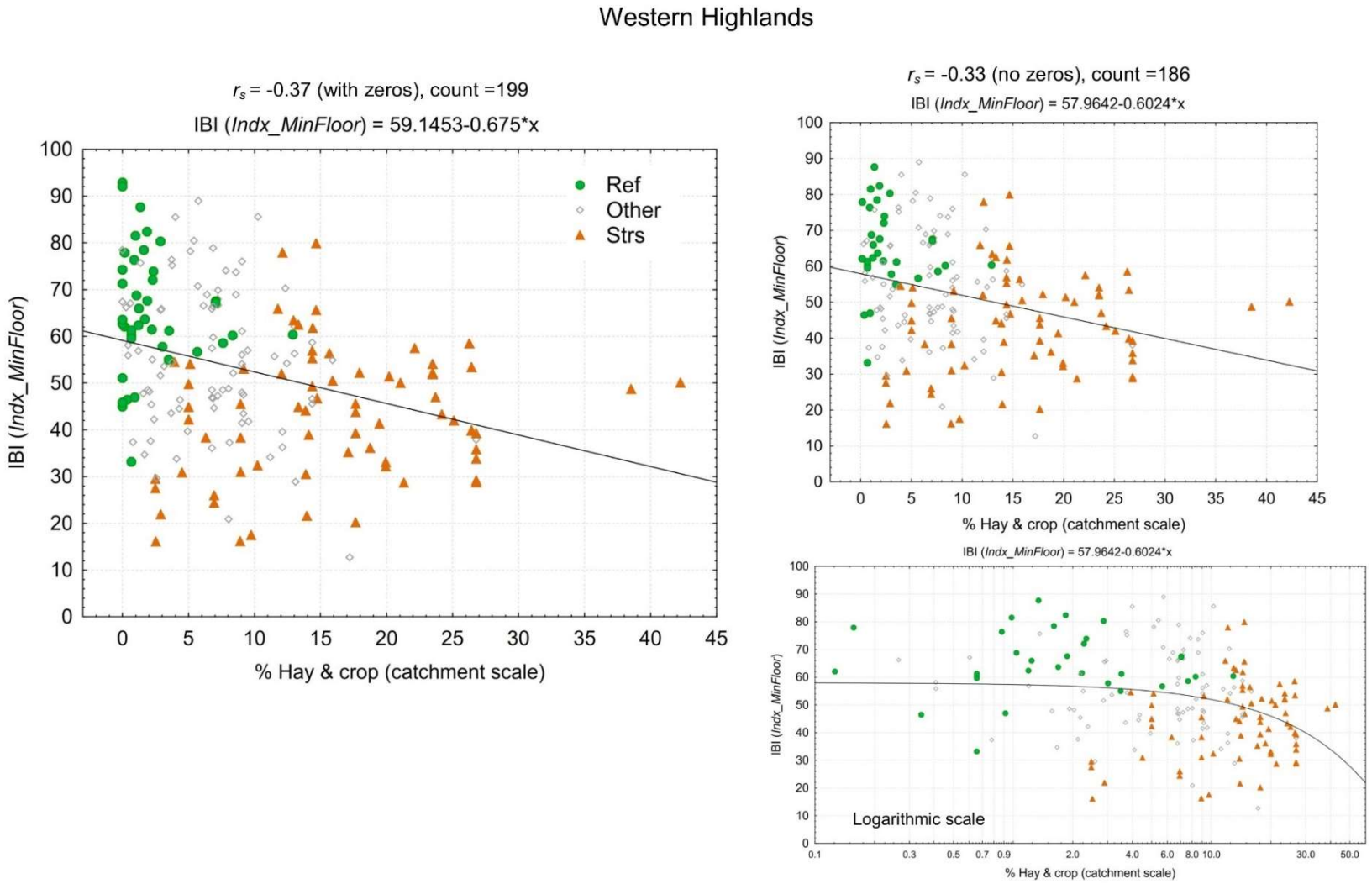


Figure C11. Scatterplot of % hay + row crop land cover (2011 NLCD, local catchment-scale) vs. IBI scores (Indx_MinFloor) in the Western Highlands dataset. The plot on the left includes all samples (including the 13 samples with % hay + crop values of 0). The plots on the right are limited to samples that have % hay + crop values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

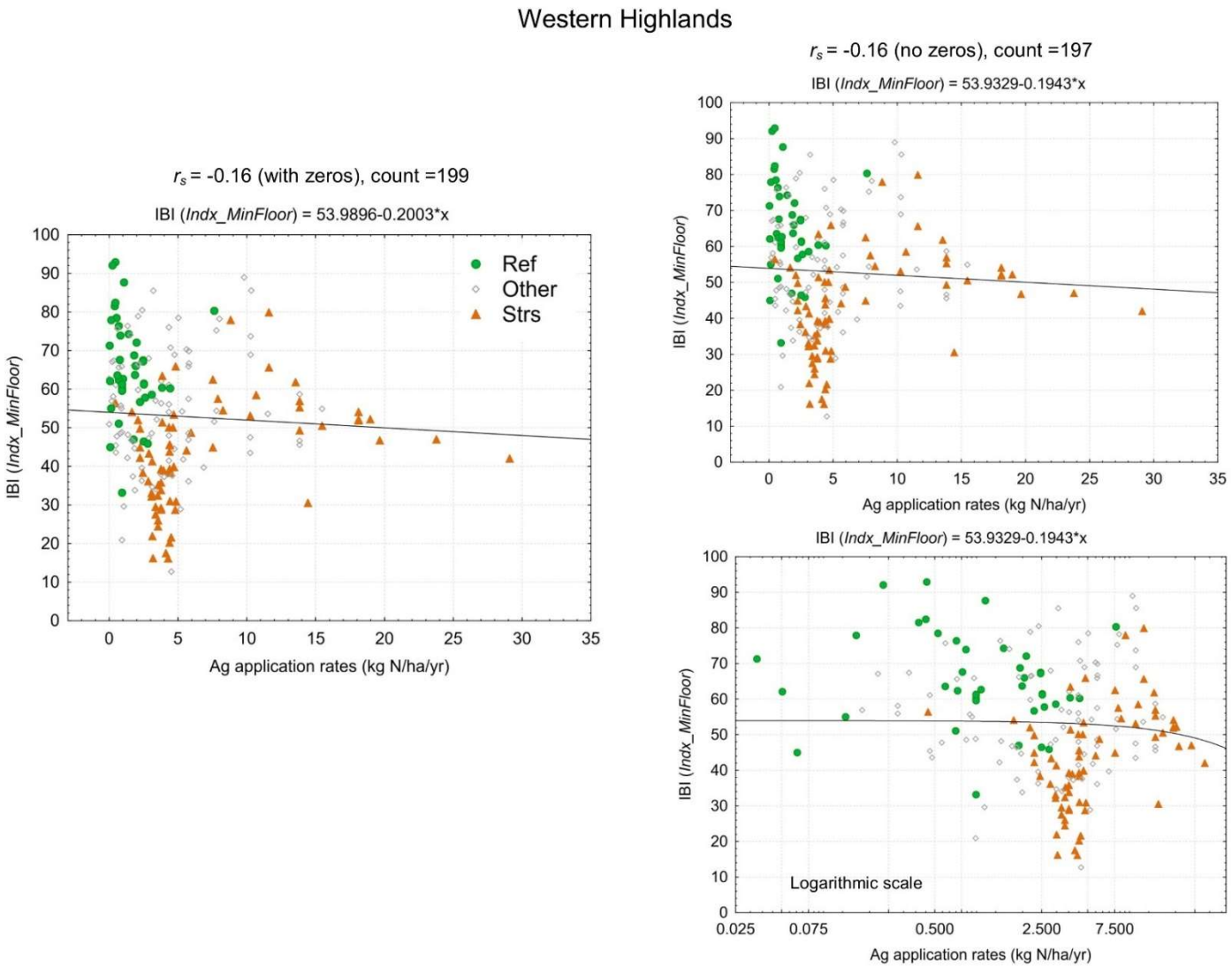


Figure C12. Scatterplot of agricultural (ag) application rates vs. IBI scores (Indx_MinFloor) in the Western Highlands dataset. The plot on the left includes all samples (including the 2 samples with ag application rate values of 0). The plots on the right are limited to samples that have % ag application rate values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

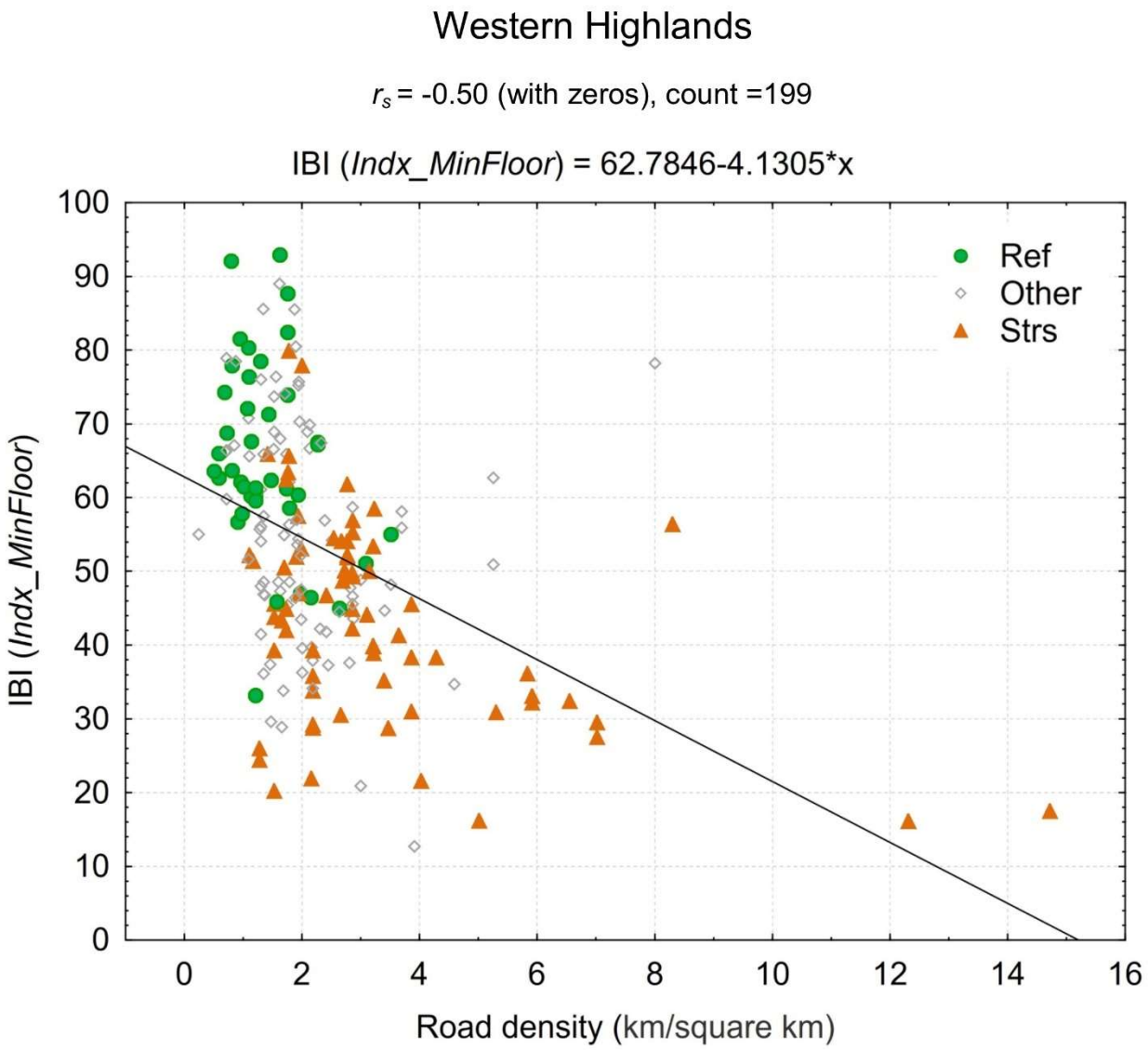


Figure C13. Scatterplot of road density vs. IBI scores (Indx_MinFloor) in the Western Highlands dataset. All samples had road density values > 0 (no zeros). Also included (above the plot) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.

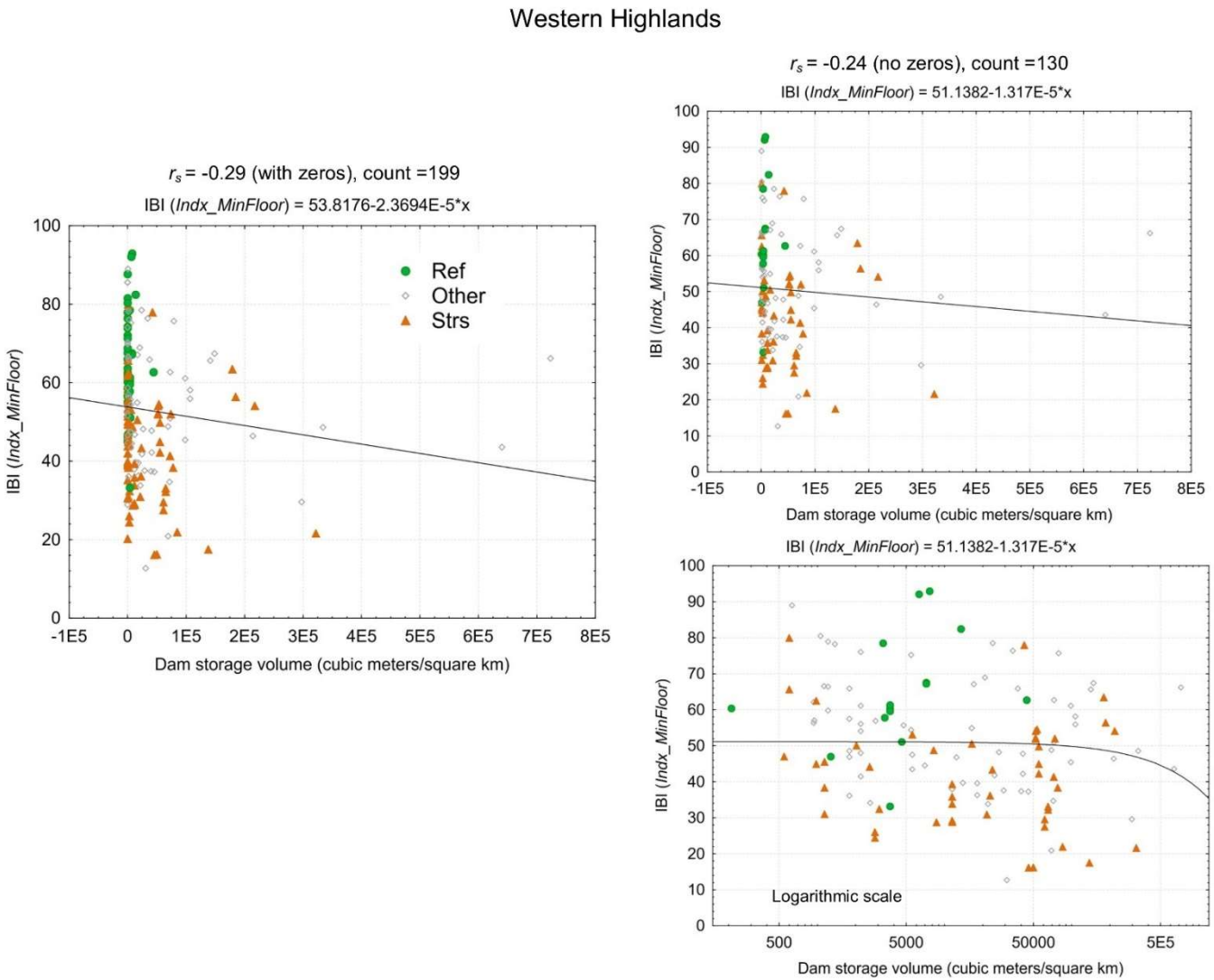


Figure C14. Scatterplot of dam storage volume vs. IBI scores (Indx_MinFloor) in the Western Highlands dataset. The plot on the left includes all samples (including the 69 samples with dam storage volume values of 0). The plots on the right are limited to samples that have dam storage volume values > 0. The x-axis in the bottom plot is transformed to logarithmic scale. Also included (above the plots) is the linear regression equation, Spearman correlation coefficient (r_s) and sample size.