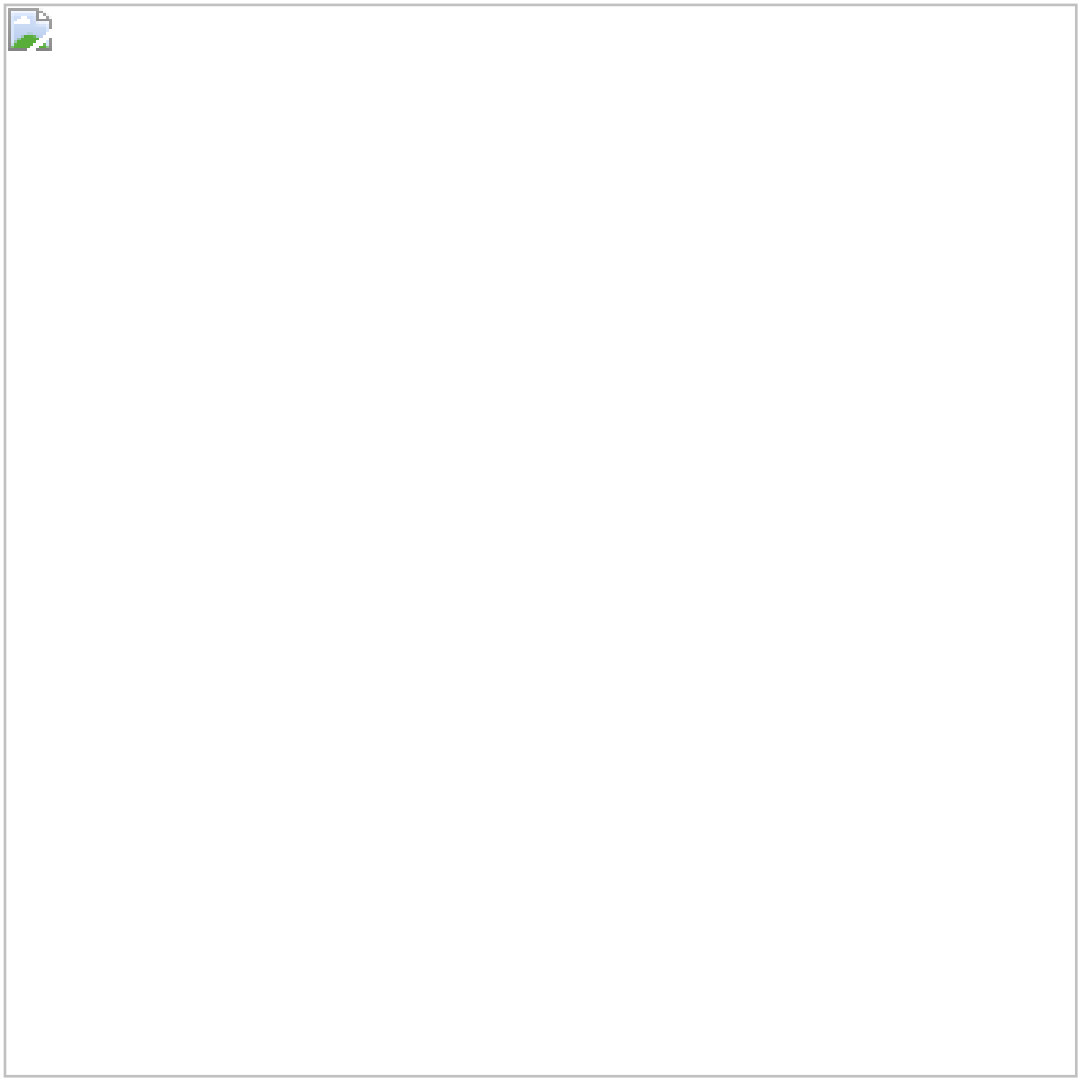

A critique of the use of indicator-species scores for identifying thresholds in species responses

Author: Cuffney, Thomas F.

Source: Freshwater Science, 32(2) : 471-488

Published By: Society for Freshwater Science

URL: https://doi.org/10.1899/12-056.1

Freshwater Science, 2013, 32(2):471–488

DOI: 10.1899/12-056.1
Published online: 26 March 2013

A critique of the use of indicator-species scores for identifying thresholds in species responses

Thomas F. Cuffney[1]
US Geological Survey, 3916 Sunset Ridge Road, Raleigh, North Carolina 27607 USA

Song S. Qian[2]
Nicholas School of the Environment, Duke University, Durham, North Carolina 27708 USA

Abstract. Identification of ecological thresholds is important both for theoretical and applied ecology. Recently, Baker and King (2010, King and Baker 2010) proposed a method, threshold indicator analysis (TITAN), to calculate species and community thresholds based on indicator species scores adapted from Dufrêne and Legendre (1997). We tested the ability of TITAN to detect thresholds using models with (broken-stick, disjointed broken-stick, dose-response, step-function, Gaussian) and without (linear) definitive thresholds. TITAN accurately and consistently detected thresholds in step-function models, but not in models characterized by abrupt changes in response slopes or response direction. Threshold detection in TITAN was very sensitive to the distribution of 0 values, which caused TITAN to identify thresholds associated with relatively small differences in the distribution of 0 values while ignoring thresholds associated with large changes in abundance. Threshold identification and tests of statistical significance were based on the same data permutations resulting in inflated estimates of statistical significance. Application of bootstrapping to the split-point problem that underlies TITAN led to underestimates of the confidence intervals of thresholds. Bias in the derivation of the *z*-scores used to identify TITAN thresholds and skewedness in the distribution of data along the gradient produced TITAN thresholds that were much more similar than the actual thresholds. This tendency may account for the synchronicity of thresholds reported in TITAN analyses. The thresholds identified by TITAN represented disparate characteristics of species responses that, when coupled with the inability of TITAN to identify thresholds accurately and consistently, does not support the aggregation of individual species thresholds into a community threshold.

Key words: benthic macroinvertebrates, change point, indicator species scores, management threshold, simulation, threshold.

Ecological thresholds are defined rather broadly as abrupt changes in the response of an ecological variable (e.g., species abundance) to a relatively small change in an environmental driver (e.g., urbanization) (Stringham et al. 2003, Groffman et al. 2006, Andersen et al. 2009). Thresholds can indicate the point at which an ecosystem transitions rapidly from one stable state to another with a concomitant change in ecosystem structure and function (Holling 1973, May 1977). Such changes usually represent a diminishment in ecological condition and the goods and services provided by the ecosystem. From a management perspective, threshold identification is important for establishing criteria that prevent degradation of the ecosystem and the subsequent high cost of restoration.

The ability of analytical methods to detect thresholds accurately and consistently is critical to their use as tools for establishing criteria that can protect aquatic resources. Methods that overestimate the threshold lead to degradation of the resource, whereas underestimates overprotect the resource, and inconsistent (i.e., highly variable) estimates provide unreliable protection. Many mathematical and statistical methods have been proposed for detecting thresholds (Brenden et al. 2008, Sonderegger et al. 2009, Dodds et al. 2010), but investigations of the accuracy and precision of these methods and the conditions under which they may be applied are relatively rare in the ecological literature (Daily et al. 2012).

[1] E-mail address: tcuffney@usgs.gov

[2] Present address: Department of Environmental Sciences, University of Toledo, 2801 West Bancroft Street, Toledo, Ohio 43606-3390 USA. E-mail: mdqian@gmail.com

TABLE 1. Indicator values ($IndVal_{ij}$) calculated for each 2-group cluster (1 and 2) associated with potential thresholds (i.e., midpoint between unique disturbance values) for a hypothetical data set. * indicates potential threshold that would not be considered in a threshold indicator analysis (TITAN) with a minimum cluster size criterion of 5.

Disturbance value	Abundance of		Potential thresholds									
	Species A	Species B	3.5*	13.0*	22.0*	32.5	41.0	51.0	62.5*	72.5*	83.5*	93.5*
0	200	152	1	1	1	1	1	1	1	1	1	1
7	180	160	2	1	1	1	1	1	1	1	1	1
19	190	155	2	2	1	1	1	1	1	1	1	1
25	185	158	2	2	2	1	1	1	1	1	1	1
25	180	155	2	2	2	1	1	1	1	1	1	1
40	60	162	2	2	2	2	1	1	1	1	1	1
42	70	158	2	2	2	2	2	1	1	1	1	1
60	55	154	2	2	2	2	2	2	1	1	1	1
65	45	149	2	2	2	2	2	2	2	1	1	1
80	80	55	2	2	2	2	2	2	2	2	1	1
87	65	53	2	2	2	2	2	2	2	2	2	1
100	68	57	2	2	2	2	2	2	2	2	2	2
$IndVal_{ij}$	Species A		53.9	54.0	54.5	55.8	54.8	54.4	53.7	52.4	52.6	52.3
	Species B		51.3	51.6	51.7	52.3	52.8	53.4	54.2	55.7	55.2	54.5

Baker and King (2010, King and Baker 2010) proposed a threshold-detection method based on the indicator-value analysis used by Dufrêne and Legendre (1997) to identify taxa that differentiate among clusters. They describe their threshold indicator analysis (TITAN) program as a nonparametric statistical method for identifying thresholds of individual species, testing the significance of the threshold, estimating uncertainty in threshold estimates, and combining thresholds into an aggregate measure that depicts the community threshold. The TITAN program identifies thresholds by ordering samples along a disturbance gradient (e.g., urban intensity) and then successively splitting the samples into 2 clusters each time the disturbance variable changes value (Table 1). The midpoints of the disturbance variables that differentiate the clusters define a series of potential thresholds in a species' response. Indicator values (*IndVal*s) are calculated for the clusters associated with each potential threshold using a modification of the method used by Dufrêne and Legendre (1997):

$$IndVal_{ij} = max\left(\frac{\bar{A}_{ij1}}{\bar{A}_{ij1}+\bar{A}_{ij2}} \times \frac{O_{ij1}}{n_{ij1}}, \frac{\bar{A}_{ij1}}{\bar{A}_{ij1}+\bar{A}_{ij2}} \times \frac{O_{ij2}}{n_{ij2}}\right) \times 100 \quad [1]$$

$$IndVal_i = \max\left(IndVal_{ij}\right) \quad [2]$$

where i is the species index, j is the potential threshold that defines the 2 clusters, $\bar{A}$ is the mean abundance in a cluster, O is the occurrence in a cluster, n is the number of samples in a cluster, and *max* is the maximum. The potential threshold associated with the cluster that has the maximum indicator value (32.5 and 72.5 for Species A and B in Table 1) is the threshold for the response of that species.

TITAN restricts the minimum sample size for a group to 3, though a minimum group size of 5 is recommended. Consequently, the search for the threshold often starts at some distance from the low end and ends at some distance from the high end of the gradient (Table 1) with the displacement dependent upon the distribution of the data along the gradient. Permutations of the data in the clusters are used to establish the statistical significance of the TITAN threshold under the null hypothesis of no change (or clustering) in the abundance or occurrence of the species along the gradient. The mean and variance of the *IndVal*s (μ_{ind}, σ_{ind}, respectively) estimated from the permutations are used to calculate the z-score ($z = \frac{IndVal - \mu_{ind}}{\sigma_{ind}}$) for each species that is used in the calculation of the community threshold. Confidence limits for each species' threshold are derived by bootstrapping.

We investigated the ability of TITAN to identify accurately and consistently thresholds in simple models with known response forms, thresholds, and variation. We calculated $IndVal_{ij}$s with a direct application of the methods of Dufrêne and Legendre (1997) to assess how $IndVal_{ij}$s change in these data sets and how closely the potential thresholds associated with the maximum $IndVal_{ij}$ correspond to the actual thresholds and the thresholds identified by TITAN. We also investigated the methods used to establish statistical significance and confidence intervals for thresholds in TITAN. The work presented here builds

on the analyses that were broadly outlined in Cuffney et al. (2011).

Methods

We assessed the ability of the TITAN method to detect thresholds accurately based on detection of known thresholds in data sets simulating the responses of a community of 225 species to a gradient of disturbance ranging from 0 (least disturbed) to 100 (most disturbed) (Supplemental data file; available online from: http://dx.doi.org/10.1899/12-056.1.s1). We modeled species responses as variations of 6 response models: broken-stick (BS), disjointed broken-stick (BSdj), step-function (SF), dose-response (DR), Gaussian (GA), and linear (LIN) (Fig. 1A–F). BS, BSdj, SF, and DR models (eqs 4–7 below) represented responses with statistical thresholds (i.e., model parameters changed at the threshold), GA models (eq. 8 below) represented responses with an ecological rather than a statistical threshold (i.e., the threshold is defined by the species optimum, μ, and model parameters do not change at the threshold), and LIN models (eq. 3) represented responses without a threshold. We based models on 201 observations evenly distributed across the gradient with no replication.

The 6 response models were defined as:

$$\text{LIN}: y_j = a + bx_j + \varepsilon \quad [3]$$

$$\text{BS}: y_j = \begin{cases} a + b(x_j - \varphi) + \varepsilon_1, \textit{if } x \le \varphi \\ a + (b+\delta)(x_j - \varphi) + \varepsilon_2, \textit{if } x > \varphi \end{cases} \quad [4]$$

$$\text{BSdj}: y_j = \begin{cases} a + b(x_j - \varphi) + \varepsilon_1, \textit{if } x \le \varphi \\ (a+\delta_1) + (b+\delta_2)(x_j - \varphi) + \varepsilon_2, \textit{if } x > \varphi \end{cases} \quad [5]$$

$$\text{SF}: y_j = \begin{cases} a + \varepsilon_1, \textit{if } x \le \varphi \\ (a+\delta) + \varepsilon_2, \textit{if } x > \varphi \end{cases} \quad [6]$$

$$\text{DR}: y_j = \begin{cases} a + \varepsilon_1, \textit{if } x \le \varphi_1 \\ a + b(x_j - \varphi_1) + \varepsilon_2, \textit{if } \varphi_2 \ge x > \varphi_1 \\ a + b(\varphi_2 + \varphi_1) + \varepsilon_3, \textit{if } x > \varphi_2 \end{cases} \quad [7]$$

$$\text{GA}: y_j = Ae^{-\frac{(x_j-\mu)^2}{2\sigma^2}} + \varepsilon \quad [8]$$

where y_j is the j^{th} observed species abundance, x_j is the j^{th} observed disturbance value, a is the intercept, b is the slope, φ is the threshold value, ε is the error term [$\sim N(0, \sigma^2)$], δ is the offset value, A is the maximum abundance, μ is the species optimum, and σ is the species tolerance.

LIN models lacked a statistical or ecological threshold and were defined by a single intercept and slope (eq. 3, Fig. 1F). BS models consisted of 2 joined line segments (A–B and B–C; Fig. 1A) that differed in slopes and intercepts (eq. 4, Fig. 1A). BSdj models were similar to the broken-stick models except that the end of the 1st line segment (B) and the beginning of the 2nd (C) were displaced along the y-axis with the displacement determined by δ_1 and δ_2 for intercepts and slopes, respectively (eq. 5, Fig. 1B). SF models were similar to the disjointed broken-stick models in that the ends of the 2 line segments (B and C) were displaced along the y-axis (a_1 and a_2; eq. 6, Fig. 1C), but differed in having 0 slopes for both line segments. Because slopes are 0, the intercepts in the SF model are the mean abundances in each line segment ($a_1 = \mu_1$, $a_2 = \mu_2$). The DR models consisted of 3 joined line segments (A–B, B–C, and C–D) in which the slope of segment B–C differed from A–B and C–D and the slopes of A–B and C–D may or may not have differed from one another or from 0 (eq. 7, Fig. 1D). The 3 segments of the DR model result in thresholds at points B and C (φ_1 and φ_2, respectively). GA models were modeled as a bell-shaped curve with the species' optimum (μ) identifying the ecological threshold (i.e., disturbance value at the transition between increasing and decreasing species abundance), σ representing the species' tolerance, and A the maximum abundance (eq. 8, Fig. 1E). The tails of the GA model were truncated by setting abundances <1 to 0.

We modeled 41 species as LIN, 42 as BS, 41 as BSdj, 36 as SF, 33 as DR, and 32 as GA. We varied model parameters (e.g., location of the thresholds, intercepts, slopes, offsets, maximum abundance, means, and standard deviations) to produce models with different abundance and occurrence characteristics. Disturbance values were uniformly distributed across the gradient and coupled with a single response value (i.e., no replication). We examined the effects of variability on the derivation of thresholds by introducing an additive error term with a normal distribution of mean 0 and a standard deviation expressed as a percentage (1, 5, 10, 25, 40, 60, 80, and 100%) of the response value resulting in 2025 simulated species responses.

These species response models are relatively simplistic and cannot provide a complete representation of all possible response models, but they do represent common patterns of response and provide important insights into the ability of TITAN to detect thresholds and assess uncertainty under a variety of conditions.

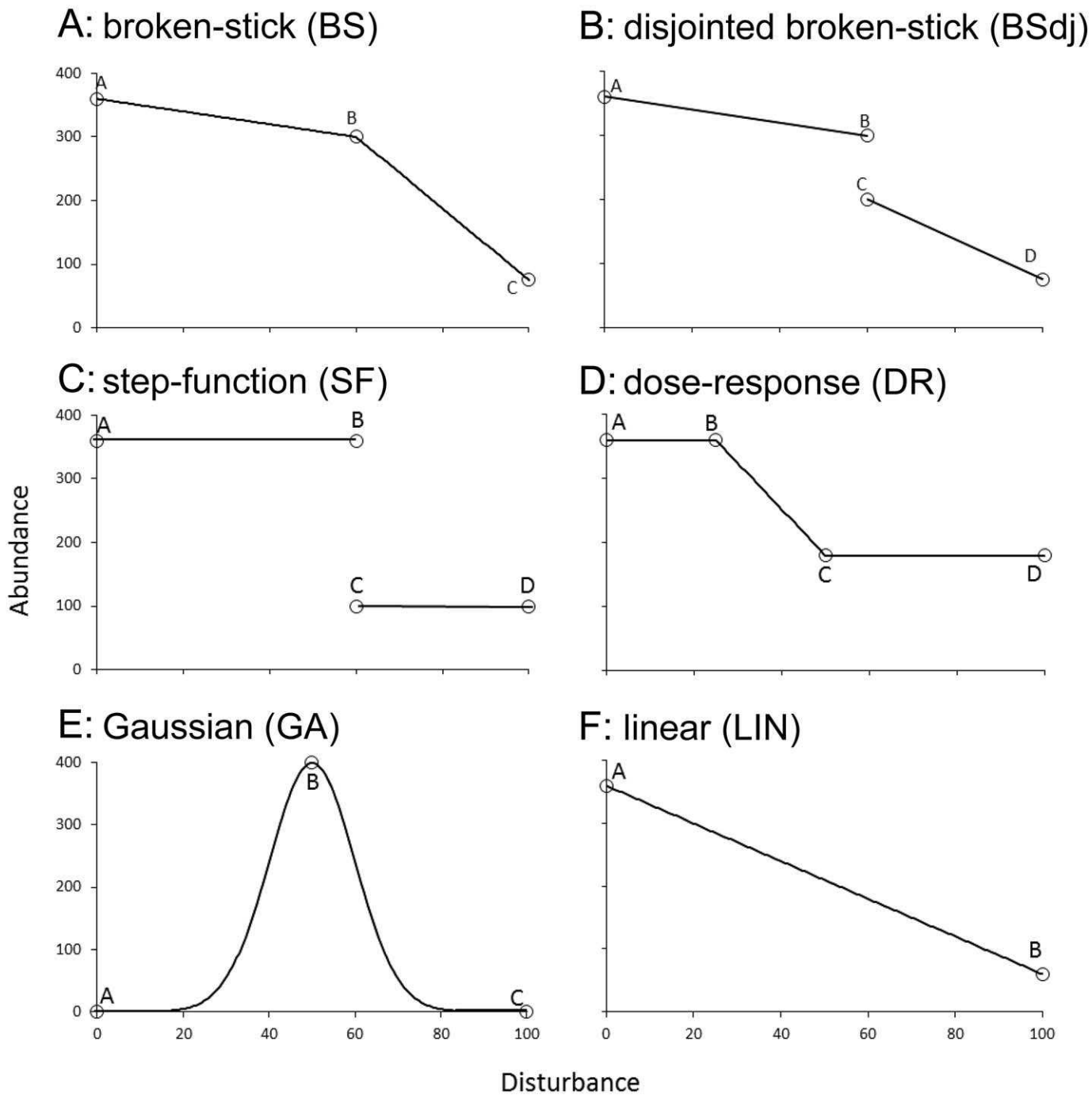

FIG. 1. Response models used to examine the performance of the Indicator Value (*IndVal*) and threshold indicator analysis (TITAN) methods: broken-stick (BS) (A), disjointed broken-stick (BSdj) (B), step-function (SF) (C), dose-response (DR) (D), Gaussian (GA) (E), and linear (LIN) (F) models.

The simulated responses represent conditions in which the response form and thresholds are known and unambiguous (variance = 0), and conditions in which the threshold and response form, while known, are obscured by varying amounts of random variability. Our approach complements the hypothetical response curves analyzed in Baker and King (2010) and King and Baker (2010) that, while addressing realistic responses, were not able to address issues related to the conditions (e.g., response form, level of variability) under which TITAN can and cannot accurately identify thresholds. We used additional models to

TABLE 2. Percentage of species-response models (Online Appendices) that were identified as having statistically significant thresholds ($p \leq 0.05$) by the threshold indicator analysis (TITAN) program at levels of introduced variability ranging from 0 to 100% of the response. BS = broken stick, BSdj = disjointed broken stick, DR = dose response, SF = step function, GA = Gaussian, LIN = linear.

Variability (%)	Response model					
	BS	BSdj	DR	SF	GA	LIN
0	100.0	100.0	100.0	100.0	100.0	100.0
1	100.0	100.0	100.0	100.0	100.0	100.0
5	100.0	100.0	100.0	100.0	100.0	100.0
10	100.0	100.0	100.0	100.0	100.0	100.0
20	100.0	100.0	100.0	100.0	100.0	100.0
40	100.0	100.0	97.0	100.0	100.0	100.0
60	97.6	95.1	81.8	100.0	96.9	97.6
80	85.7	95.1	78.8	88.9	96.9	97.6
100	78.6	90.2	69.7	94.4	93.8	92.7

investigate the effects of replicated data, data distributions (uniform, right-, and left-skewed), and distributions of 0 values on threshold detection in TITAN.

We calculated thresholds using the TITAN program (Baker and King 2010) with the following default values: 250 permutations, taxa present at a minimum of 5 sites, minimum cluster size of 5, and $\log_{10}(y + 1)$-transformation of abundance data. We repeated TITAN analyses 25 times to assess variability in the estimates of thresholds. We used a smaller set of representative data to examine issues related to bootstrapping because of the extremely long calculation times required for bootstrapping.

We calculated $IndVal_{ij}$s independently of TITAN with the method of Dufrêne and Legendre (1997). We placed samples in ascending order along the disturbance gradient and defined potential thresholds as the mid-points between successive unique disturbance values. We $\log_{10}(y + 1)$-transformed abundances prior to calculating *IndVal*s. Table 1 illustrates the process of identifying potential thresholds, calculating *IndVal*s values (eqs 1, 2), and the effects of TITAN's minimum cluster size (5) on threshold detection in a small data set representing 2 species responses that follow a SF model. Two samples had replicate disturbance values (25, 25) resulting in potential thresholds at 22 ([19+25]/2) and 32.5 ([25+40]/2), but not at 25 ([25+25]/2). The clusters defined by each potential threshold are represented as 1s and 2s in Table 1. *IndVal*s can be calculated for all potential thresholds, but TITAN requires a minimum number of samples in each cluster. For example, only 3 (32.5, 41.0, and 51.0) of the 10 potential thresholds in Table 1 would be considered in a TITAN analysis if a minimum cluster size of 5 were used. Independently calculated $IndVal_{ij}$s were used to determine the response of $IndVal_{ij}$s across the gradient and the relationship between the maximum $IndVal_{ij}$ (e.g., 32.5 for Species A and 72.5 for Species B) and the TITAN threshold.

Results and Discussion

The TITAN program reported statistically significant ($p < 0.05$) thresholds for all species response models with <25% variability (Table 2). However, even under ideal conditions (0% introduced variability) the thresholds detected by the TITAN program were consistently accurate estimates of the actual thresholds only for SF models (Fig. 2A–F). Under less-ideal conditions (1–100% introduced variability), the deviation of TITAN thresholds from actual thresholds tended to increase when random variability was ≥40%, particularly for SF models (Fig. 3). Deviations for GA models were the largest of all the models studied and exhibited little change as variability increased. The level of variability (40%) at which statistical significance dropped off and the deviation of SF thresholds increased would not be unexpected for replicate macroinvertebrate samples depending upon the sampling methods used, the abundance of the taxon, and the spatial distribution of the taxon (Elliott 1977, Resh 1979).

Plots of $IndVal_{ij}$s showed that the maximum value corresponded to the actual threshold only for SF models (Fig. 4A–H, Online Appendices; available online from: http://dx.doi.org/10.1899/12-056.1.s2). Maximum $IndVal_{ij}$ for BS, BSdj, and LIN models tended to occur at either end of the species distribution along the gradient (Fig. 4A–C, H, Online Appendices). If the species' abundance increased across the gradient, $IndVal_{ij}$ decreased across the

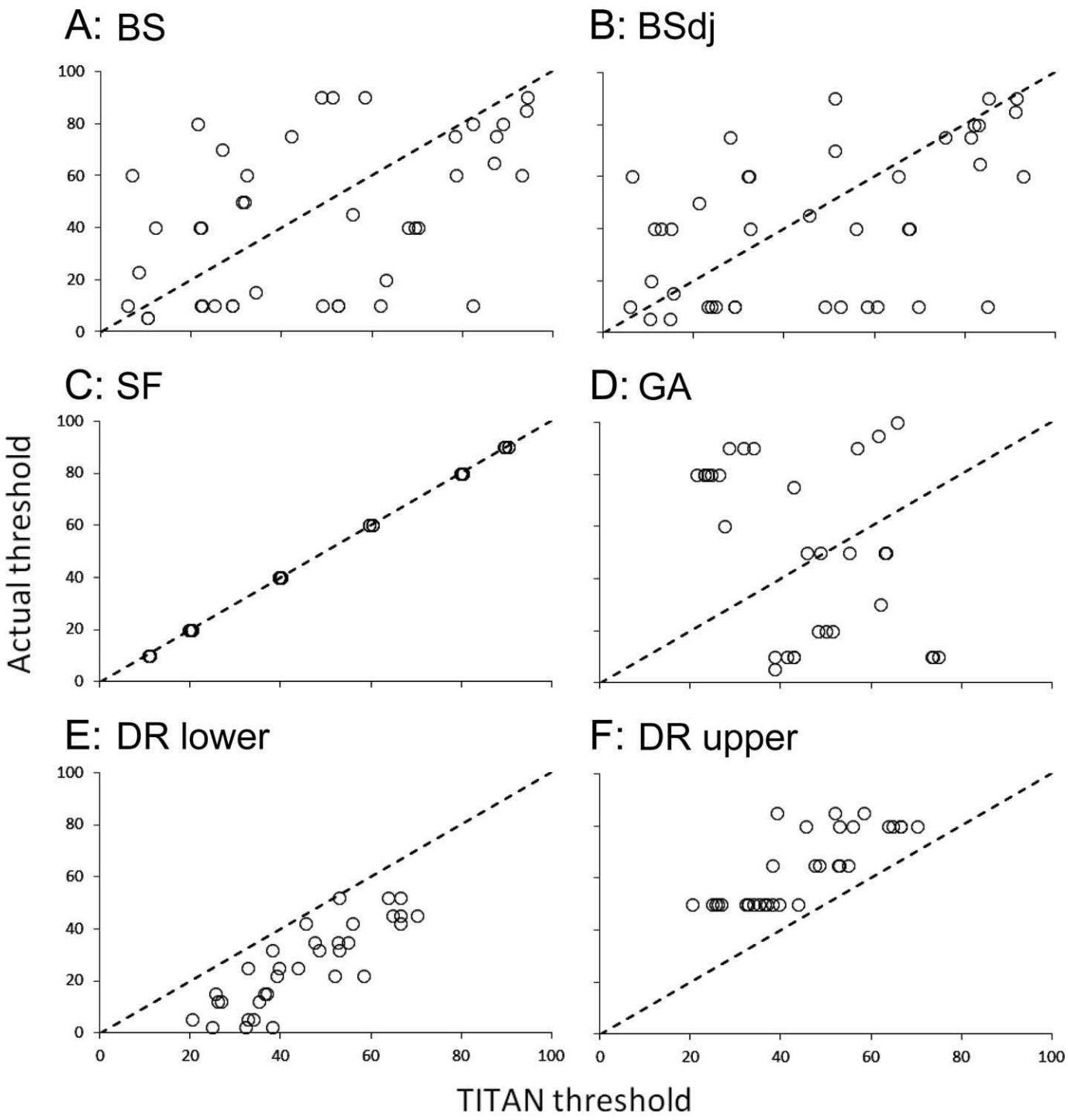

FIG. 2. Correspondence between thresholds estimated by threshold indicator analysis (TITAN) and the actual thresholds for the broken-stick (BS) (A), disjointed broken-stick (BSdj) (B), step-function (SF) (C), Gaussian (GA) (D), and lower (E) and upper dose-response (DR) (F) models (variance = 0%). The dashed line indicates the 1:1 correspondence between the actual thresholds and the thresholds estimated by TITAN.

gradient. Conversely, if abundance decreased across the gradient, $IndVal_{ij}$ increased across the gradient. If the species occurred over only a portion of the gradient (truncated species distributions; Fig. 5A–F), $IndVal_{ij}$s reached a peak at the point where the response curve intersected the *x*-axis rather than at the actual threshold. This corresponds to the point on the gradient that divided the data into a cluster

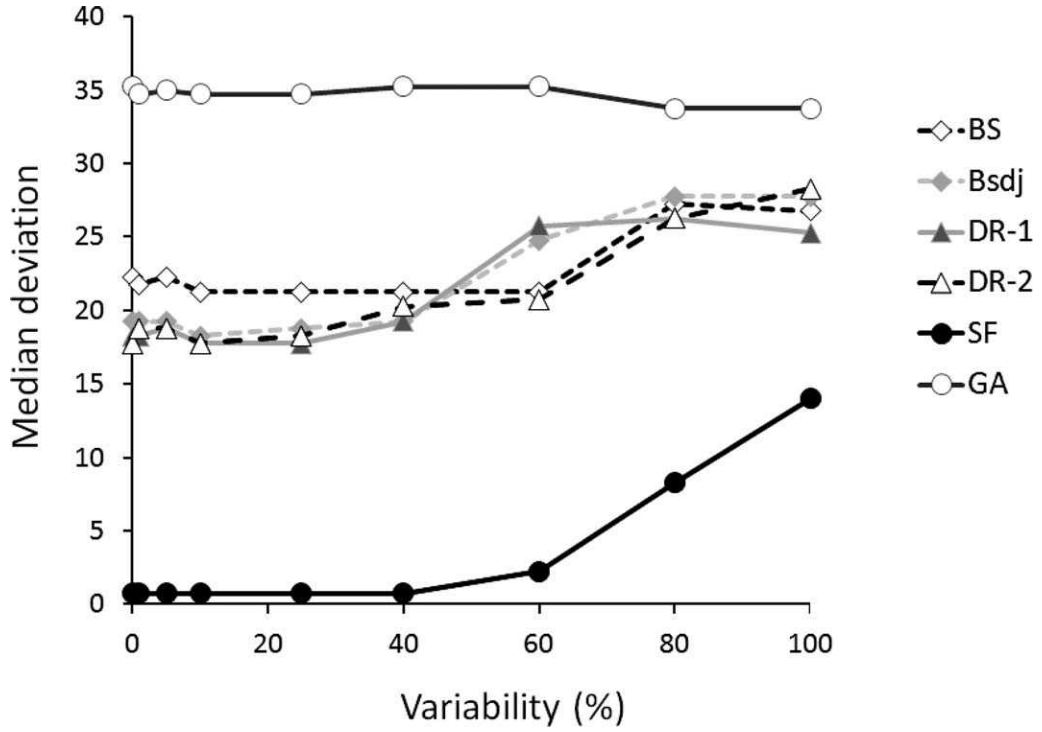

FIG. 3. Median deviation between the threshold indicator analysis (TITAN)-estimated threshold and the actual threshold at various levels of introduced random variability.

consisting of all 0 values and a cluster consisting of non-0 values.

$IndVal_{ij}$s in GA models (Figs 4F, G, 5D, E, Online Appendices) showed responses that were similar to BS, BSdj, and LIN models when the model was asymmetrically located along the gradient resulting in an overall increasing or decreasing trend in the abundance (GA13 and GA10; Figs 4G, 5D). When GA models were symmetrically located across the gradient (GA4 and GA1; Figs 4F, 5E), peak $IndVal_{ij}$s corresponded to the ends of the distribution and reached a minimum, rather than a maximum, value at the threshold (μ). In these cases, multiple runs of the TITAN program detected one or the other end of the GA distribution resulting in a mean threshold estimate that was much closer to the actual threshold than any of the individual TITAN threshold estimates. As with BS, BSdj, and LIN models, the $IndVal_{ij}$s in GA models peaked at the point or points where the abundance curve intercepted the *x*-axis (GA1 and GA10, Fig. 5D, E) rather than at the threshold when the species occurred over only a portion of the disturbance gradient.

$IndVal_{ij}$s in DR models typically showed a peak that lay between the 2 thresholds (DR4; Fig. 4E), though some DR models showed a monotonic response across the gradient with a maximum value at the ends of the disturbance gradient if line segment A–B or C–D was very short (e.g., DR17; Online Appendices). Step-function models (SF3; Fig. 4D, Online Appendices) were the only response models that showed a sharp peak in the $IndVal_{ij}$s that corresponded to the actual threshold and to the threshold identified by TITAN.

Threshold identification in TITAN

Thresholds estimated by TITAN varied each time the data were run through the program (Fig. 6) even though the underlying data, potential thresholds, and $IndVal_{ij}$s did not change (Figs 4A–H, 5A–F, Online Appendices). The variability in threshold estimates occurs because TITAN uses the maximum *z*-score to identify the threshold rather than the maximum $IndVal_{ij}$ value (Fig. 7A–C). *z-scores* are derived from a relatively small subset (e.g., 250–500) of all possible permutations of the data in the clusters. Consequently, the statistics used to calculate *z*-scores (μ_{ind} and σ_{ind}) and to identify the threshold vary each time the program is run. Increasing the number of permutations would appear to address this issue, but other problems with the derivation of *z*-scores cannot be addressed simply by increasing the number of permutations.

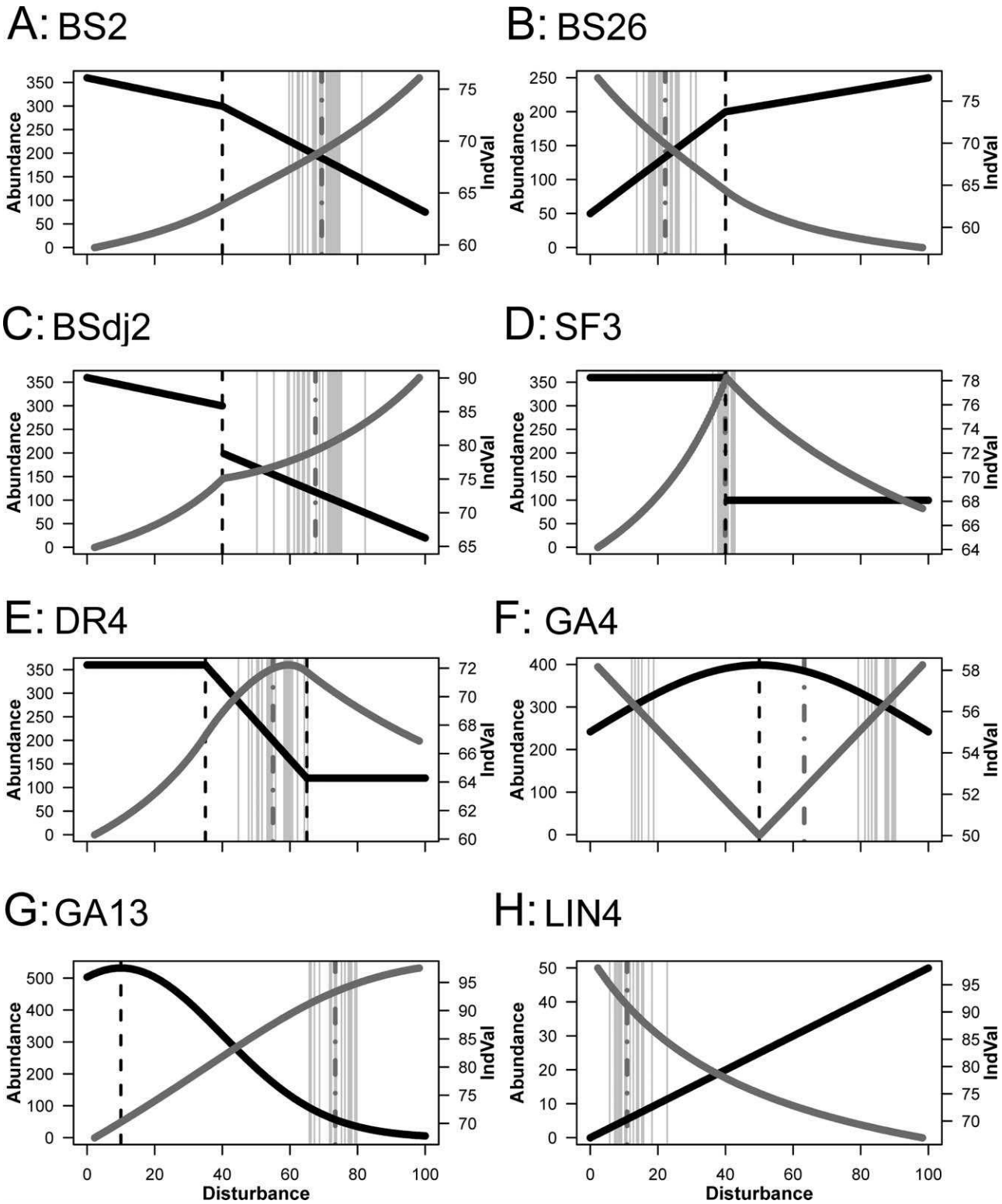

FIG. 4. Broken-stick (BS) models 2 (A) and 26 (B), disjointed broken-stick (BSdj) model 2 (C), step-function (SF) model 3 (D), dose-response (DR) model 4 (E), Gaussian (GA) models 4 (F) and 13 (G), and linear (LIN) model 4 (H) showing the response model (solid black line), the actual threshold (vertical dashed black line), the 25 thresholds estimates obtained from threshold indicator analysis (TITAN) (vertical light gray lines), the mean TITAN threshold (vertical gray dash-dot line), and the independently calculated indicator value ($IndVal_{ij}$) (dark gray response line). Model abbreviations and numbers refer to the models in Online Appendices.

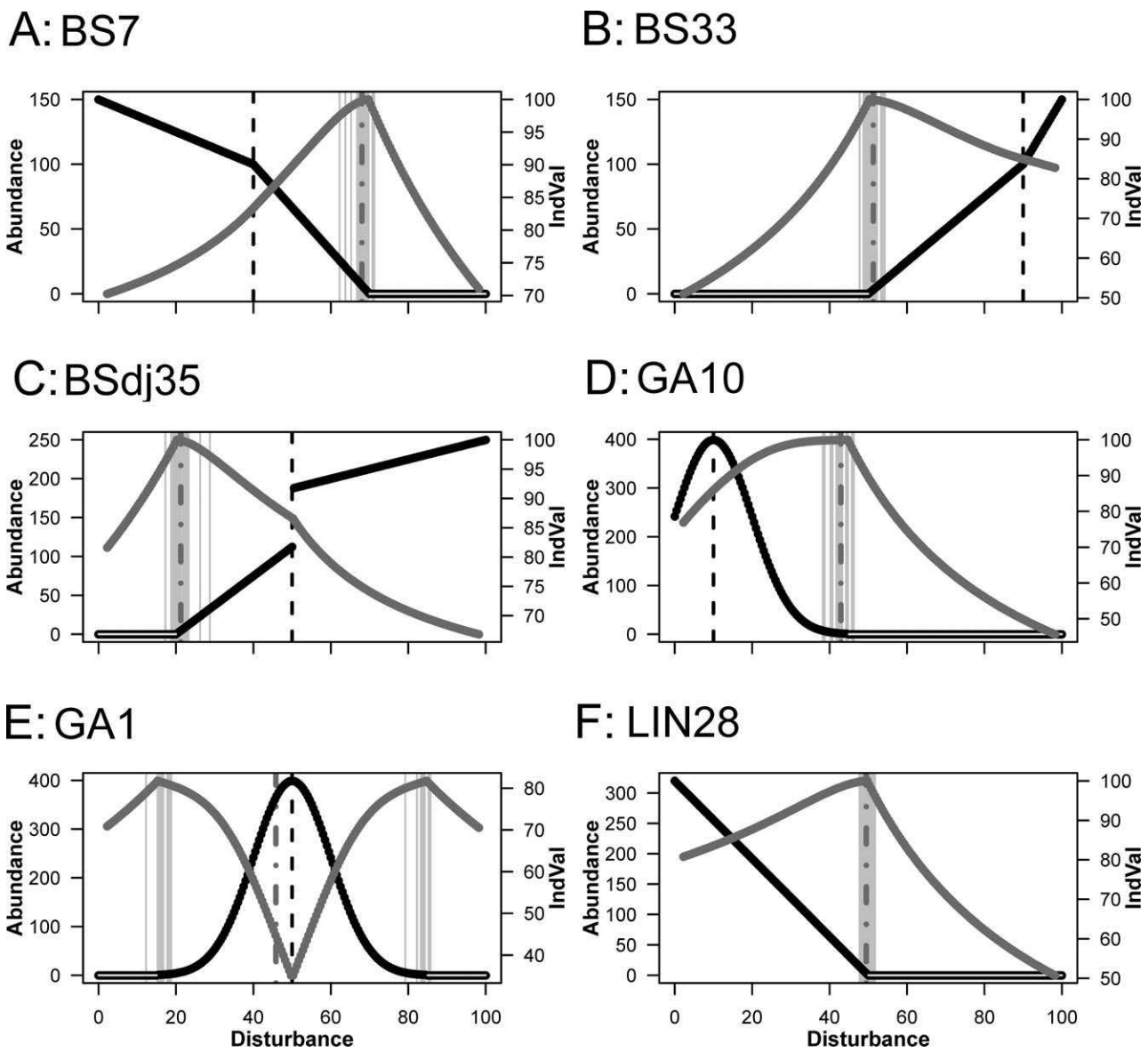

FIG. 5. Effect of truncated species distributions (structural 0s) on the ability of TITAN to detect thresholds accurately in broken stick (BS) models 7 (A) and 33 (B), disjointed broken stick (BSdj) model 35 (C), Gaussian (GA) models 10 (D) and 1 (E), and linear (LIN) model 28 (F) showing the response model (solid black line denotes the response domain, black line with white center denotes 0 abundances), the actual threshold (vertical dashed black line), the 25 thresholds estimates obtained from threshold indicator analysis (TITAN) (vertical light gray lines), the mean TITAN threshold (vertical gray dash-dot line), and the independently calculated indicator value ($IndVal_{ij}$) (dark gray response line). Model abbreviations and numbers refer to the models in Online Appendices.

Disparities in cluster sizes across the disturbance gradient introduce a systematic bias in the μ_{ind} and σ_{ind} statistics used to derive the *z*-scores and identify thresholds. This bias results in a U-shaped distribution across the gradient (Fig. 7A–C) with larger values of μ_{ind} and σ_{ind} at the ends of the gradient and smaller values in the middle. This pattern was evident for all model forms (BS, BSdj, SF, DR, GA, LIN) examined and results from disparities in cluster sizes associated with potential thresholds across the gradient. For example, if data are evenly distributed across the gradient (uniform; Fig. 7B), the larger disparity in cluster sizes at the ends (low and high) of the gradient will bias μ_{ind} and σ_{ind} toward high values and *z*-scores toward low values. This bias results in *z*-scores that can peak near, but not at, the ends of the gradient and

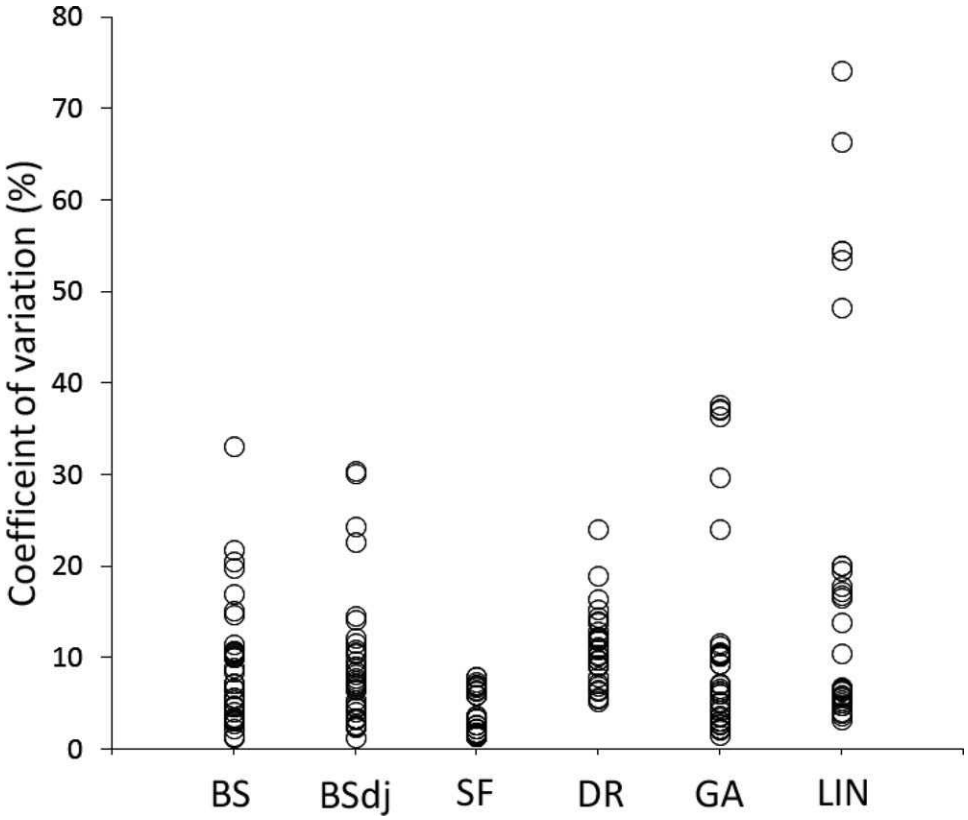

FIG. 6. Variability (coefficient of variation) in the threshold estimates produced by the threshold indicator analysis (TITAN) program for the broken-stick (BS), disjointed broken-stick (BSdj), step-function (SF), dose-response (DR), Gaussian (GA), and linear (LIN) response models. Variability is based on 25 iterations of the TITAN program using the default value of 250 permutations.

account for why TITAN does not identify thresholds at the ends of the gradient when the $IndVal_{ij}$s show a monotonic response across the gradient. The bias in μ_{ind}, σ_{ind}, and *z*-scores persist even if the distribution of data across the gradient is highly skewed (Fig. 7A, C).

TITAN accurately and consistently detected thresholds in responses only for SF models. It could not accurately detect thresholds that were associated with abrupt changes in rates (slopes) or direction (GA) of response nor could it differentiate responses that lacked a threshold (LIN). A more comprehensive approach to threshold detection would be to compare multiple alternative models (Qian et al. 2003, Qian and Cuffney 2012) to determine whether the data fit a response model that is indicative of a threshold or whether the response cannot be differentiated from a model without a threshold (e.g., linear model).

Statistical significance and confidence limits

TITAN uses permutation tests to determine the statistical significance of threshold estimates. On the basis of these tests, the thresholds derived for all simulated species responses with <25% introduced variability were statistically significant ($p \leq 0.004$), even those derived from LIN responses that have no thresholds (Table 2). The percentage of statistically significant thresholds tended to decrease as variability increased >25%, but most response models still had a high percentage of statistically significant thresholds at relatively high levels of variability.

TITAN assesses the statistical significance of thresholds using the same permutations that were used to calculate *z*-scores and identify thresholds (Table 3). The maximum *z*-score across the gradient (15.66) identifies the threshold (49.5), and the proportion of permuted *IndVal*s that are larger than the $IndVal_{ij}$ associated with the threshold (52.99) is used to determine the statistical significance of the threshold ($p = 0.002$). This approach is problematic because the permutations used to identify the threshold (i.e., find the best 2-cluster classification) also are used to test the statistical significance of the threshold (classification). Evaluation of the statistical significance of classifications using permutations requires that the classification must be derived a priori for the tests to

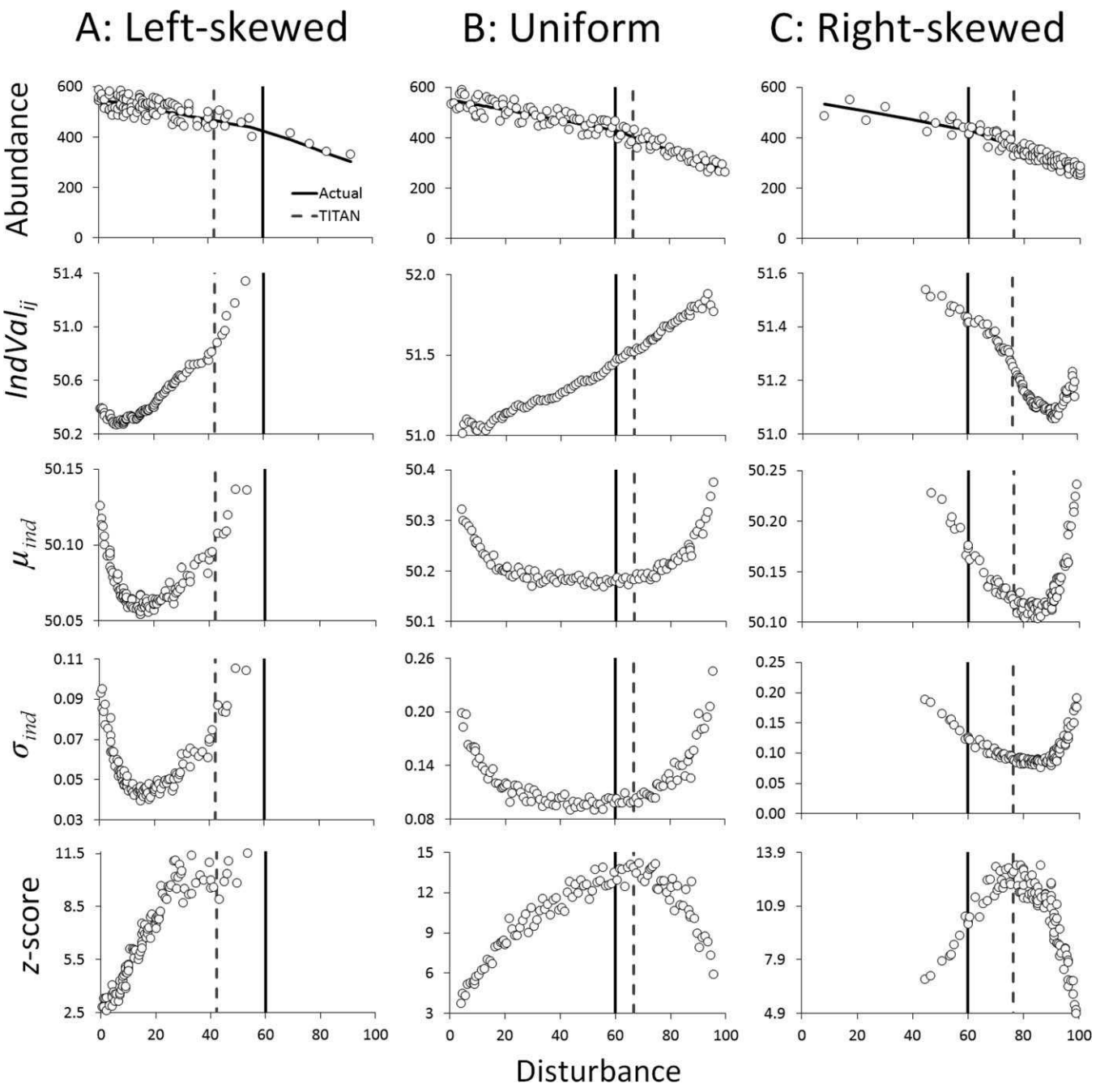

FIG. 7. Variables (abundance and indicator values [$IndVal_{ij}$]) and statistics (μ, σ, and z-scores of $IndVal_{ij}$s) used to identify thresholds in threshold indicator analysis (TITAN) based on a common broken-stick model with left-skewed (A), uniform (B), and right-skewed (C) data distributions. The solid vertical line identifies the actual threshold and the dashed line indicates the TITAN threshold.

be statistically valid, as emphasized in analysis of similarity (ANOSIM; Clarke and Gorley 2006) and multiresponse permutation procedure (MRPP; McCune and Mefford 1999) analyses. Because the permutation tests in TITAN are used both to derive and to test the classifications, the p-values derived from TITAN overestimate the statistical significance of the thresholds. This accounts for the very low p-values (1/[number of permutations + 1]) commonly observed even for LIN models.

TITAN uses a bootstrap resampling method to calculate confidence intervals for the selected threshold. Bootstrapping uses Monte Carlo simulation to obtain an approximation of the sampling distribution

TABLE 3. Indicator values ($IndVal_{ij}$s), means (μ_{ind}), variances (σ_{ind}), and z-scores obtained from 500 permutations of the data in the clusters defined by the potential thresholds.

	Potential threshold								
Permutation	5.5	6.5	7.5	…	49.5	…	94.5	95.5	96.5
1	50.5	50.2	50.4	…	50.2	…	50.2	50.5	50.3
2	50.0	50.2	50.8	…	50.1	…	50.2	50.5	50.9
3	50.5	50.2	50.4	…	50.1	…	50.2	50.0	50.3
:	:	:	:	:	:	:	:	:	:
498	50.5	50.2	50.0	…	50.1	…	50.2	50.5	50.9
499	50.5	50.2	50.4	…	50.1	…	50.2	51.1	50.3
500	50.5	50.7	50.4	…	50.0	…	50.7	50.5	51.7
Observed $IndVal_{ij}$	51.58	51.60	51.61	…	52.99	…	51.69	51.67	51.66
μ_{ind}	50.50	50.51	50.46	…	50.23	…	50.50	50.49	50.60
σ_{ind}	0.41	0.33	0.37	…	0.18	…	0.35	0.43	0.40
z-score	2.67	3.24	3.13	…	15.66	…	3.46	2.74	2.63

of the variable of interest by substituting random samples from the existing data for random samples from the target population (Efron and Tibshirani 1993). However, bootstrapping is not appropriate for a split-point problem (Bühlmann and Yu 2002, Banerjee and McKeague 2007), such as TITAN, because the estimated standard deviation of the threshold will always be smaller than the true standard deviation, thereby leading to a narrower confidence interval. In a TITAN analysis with k unique gradient values, the number of potential thresholds ($k - 1$) and their values are fixed. The number of potential thresholds in a bootstrap sample is always $< k - 1$ and represents a subsample of the same $k - 1$ potential thresholds in the original data. In other words, the bootstrapping process repeatedly selects the threshold from the same pool of $k - 1$ potential thresholds. Consequently, the bootstrapped estimated standard deviation is much smaller than it should be and the estimated confidence interval is much narrower than it should be. This problem can be illustrated by comparing the distribution of thresholds derived using bootstrapping in TITAN to $IndVal_{ij}$s (Fig. 8A–C). The displacement of the distribution toward the center of the gradient is apparent and is similar to the bias introduced in z-score calculations using statistics derived from permutations.

The uncertainty analyses incorporated into TITAN (permutation tests and bootstrapping) have statistical problems that lead to overestimates of the significance of the thresholds and underestimates of confidence intervals. A better approach to estimating uncertainty would be to use Bayesian methods to estimate model parameters (e.g., φ in eqs 4–7, μ in eq. 8) from which statistics (mean, standard deviation, and confidence limits) could be derived and tested (Qian et al. 2003, Qian and Cuffney 2012). Bayesian analysis would provide better estimates of uncertainty that could be applied to a wide range of model types and compared to identify the most appropriate representation of the response.

Abundance, occurrence, and thresholds

TITAN accurately detects thresholds in SF models because the abrupt change in abundance that defines these models (Fig. 1C, eq. 6) divides the data into 2 clusters that maximize the difference in average abundance relative to other cluster pairs. Consequently, the potential threshold associated with the maximum $IndVal_{ij}$ and maximum z-value corresponds to the actual threshold. In contrast, models in which thresholds are defined by an abrupt change in the rate (slopes in BS, BSdj, or DR; Fig. 1A, B, D) or direction of response (GA; Fig. 1E) produce gradual changes in average abundance across successive clusters. For these models, neither maximum $IndVal_{ij}$ nor maximum z-value corresponds to the actual threshold unless the model approximates a SF model (e.g., BSdj model with slopes close to 0).

Threshold detection in TITAN also is strongly influenced by the distribution of occurrences across the gradient. As with abundance, TITAN readily detects thresholds that correspond to an abrupt change in occurrence. Such changes are often associated with the limits of a species distribution across the gradient (i.e., species' response domain; Fig. 9A, B) or with the distribution of occurrences within a species' response domain (Fig. 9C, D). TITAN's sensitivity to the distribution of occurrences is evident when species responses include abrupt changes in both occurrence and abundance. In these situations, TITAN identifies thresholds associated with relatively small changes in occurrence while ignoring thresholds

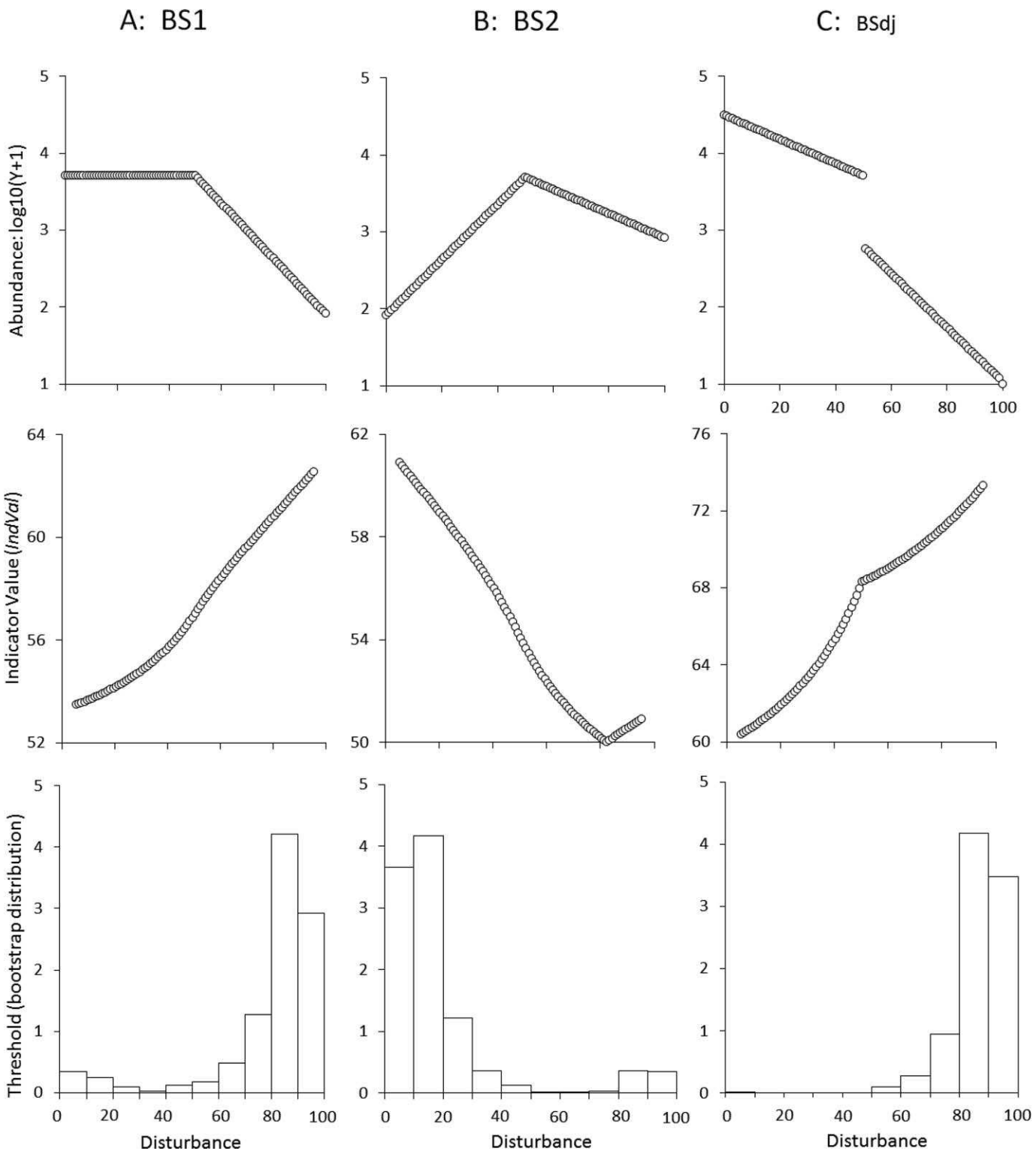

FIG. 8. Three response models that illustrate the changes in abundance, indicator values ($IndVal_{ij}$), and distribution of TITAN threshold estimates obtained from bootstrapping along the disturbance gradients. A.—Broken stick (BS)1. B.—BS2. C.—Disjointed broken stick (BSdj).

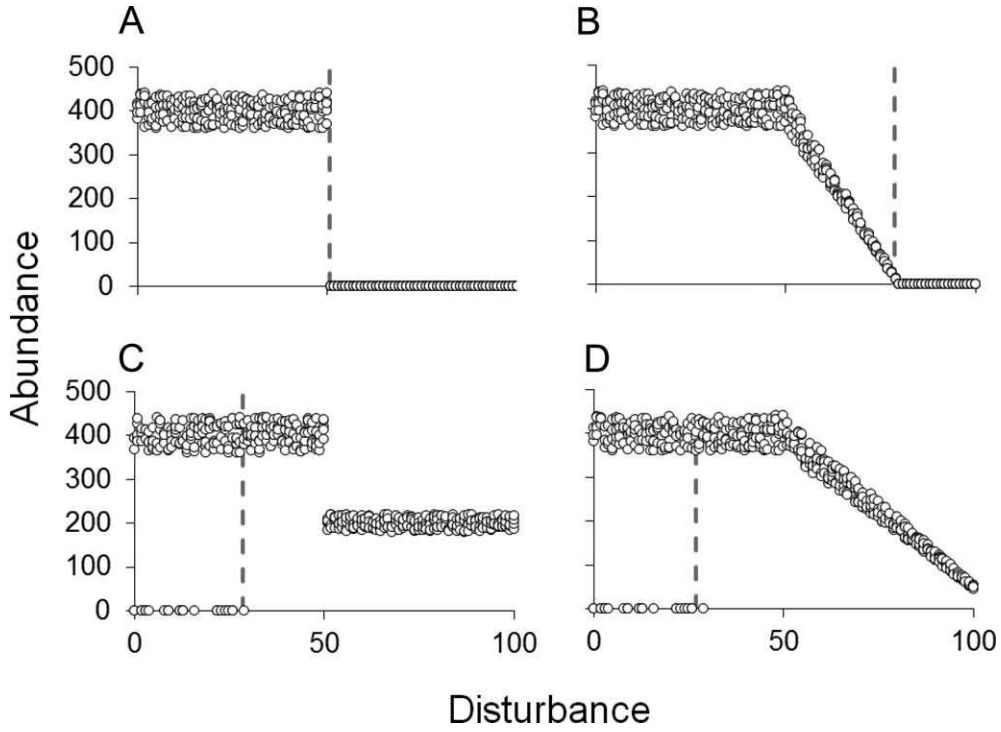

FIG. 9. The effect of an abrupt change to 0 abundance (A), a gradual change to 0 abundance (B), and changes in the distribution of 0 values within the response domain of a step-function (C) and broken-stick (D) model on thresholds identified by TITAN (vertical gray dashed line).

associated with large changes in abundance (Fig. 9B–D). Abrupt changes in occurrence represent interesting characteristics of a species response, but whether they should be interpreted as thresholds (i.e., an abrupt change in an ecological attribute) (Groffman et al. 2006) and combined into a community threshold is a matter of debate. For example, if the change in occurrence arises because the species is present over only a portion of the disturbance gradient, the resulting threshold will be consistent with the threshold definition if the transition to 0 constitutes an abrupt change (Fig. 9A), but not if it constitutes a gradual change (Fig. 9B). When the change in occurrence is associated with the distribution of 0 values within the response domain of the species (Fig. 9C, D), these 0 values constitute errors in sampling (failure to detect the species when it is present) or experimental design (response to environmental factors not captured in the disturbance measure). A relatively small difference in the distribution of these errors (0 values) can cause TITAN to identify a threshold associated with the errors even if other thresholds are present (Fig. 9C, D). Identifying thresholds in the distribution of errors within a species' response domain is ecologically and statistically meaningful, but this threshold should not be combined with thresholds representing other types of response characteristics (e.g., abrupt change in slopes, change in direction of response, or limits of the response domain) to form a community threshold.

Confusing thresholds that arise from sampling 0s (errors), structural 0s (response domains), or changes in abundance (slopes and direction) can lead to errors regarding the presence and ecological significance of thresholds. Unfortunately, distinguishing between the various causes of threshold detection to which TITAN is sensitive can be extremely difficult, particularly in real world situations where gradients may be short, experimental control may be tenuous, and levels of organism identification less than optimal (e.g., mix of taxonomic levels), all of which make differentiating responses along the disturbance gradient difficult. Only when analyses are conducted on simple, idealized responses with known forms, thresholds, and levels of random variability can the limitations of threshold methods be determined (Daily et al. 2012).

Synchronicity and data distributions

Baker and King (2010, King and Baker 2010, King et al. 2011) identify synchronicity of thresholds among species as a major finding of their work with TITAN. However, at least some of this apparent synchronicity probably results from a combination of inaccurate threshold estimates, biases in *z*-scores, and effects of

TABLE 4. Ranges in threshold indicator analysis (TITAN) thresholds expressed as a percentage of the range in actual thresholds (20, 40, 60, 80) for broken-stick (BS), disjointed broken-stick (BSdj), and step-function (SF) models with different data distributions (uniform, left- or right-skewed data) and directions of change in abundances (i.e., increasing or decreasing across the gradient).

Model	Direction of change in abundance across gradient	Distribution of data		
		Left	Uniform	Right
BS	Decreasing	9.5	12.7	14.7
	Increasing	9.3	22.2	33.0
BSdj	Decreasing	36.3	32.2	18.3
	Increasing	31.3	68.2	42.5
SF	Decreasing	57.5	98.3	55.5
	Increasing	57.7	98.7	54.2

skewed data distributions. Because of *z*-score bias, models (BS, BSdj, LIN, GA) in which $IndVal_{ij}$s exhibit a monotonic response over the gradient will yield TITAN thresholds near the low (increasing abundance) or high (decreasing abundance) end of the gradient regardless of the actual threshold. This result gives the impression that species have similar (synchronous) thresholds when actual thresholds may be very different or nonexistent. Strongly skewed data distributions also cause TITAN to identify species thresholds that are more similar to one another than are the actual thresholds, particularly when the minimum cluster size constraint (5) results in inclusion of the actual threshold in the first or last cluster (e.g., highest threshold in left-skewed and lowest threshold in right-skewed data). This situation causes TITAN to underestimate the actual threshold and increases the similarity among thresholds. These effects can be illustrated by comparing TITAN thresholds derived for BS, BSdj, and SF models that have thresholds at 20, 40, 60, and 80 along a gradient that ranges from 0 to 100. The ranges in TITAN thresholds for each model were strongly affected by the data distribution (uniform, left-, or right-skewed; Fig. 7A–C) and the direction of change (increasing or decreasing) in abundances across the gradient (Table 4). The thresholds derived from SF models with uniform data distributions captured essentially all of the range in the actual thresholds (>98%). However, thresholds derived from SF models with left- or right-skewed data distributions captured only 54 to 58% of the range because the low (20) and high (80) thresholds could not be represented by potential thresholds. Thresholds derived from BS and BSdj models represented 9.3 to 68.2% of the range in actual thresholds. These models were affected by both data distributions (lowest thresholds associated with left- and highest with right-skewed data) and direction of change in abundance (lower thresholds associated with increasing abundance).

These results reinforce the close association between the *IndVal* method and the SF model and underscore the need to use threshold detection methods that are appropriate for the underlying model and data distribution. Applying TITAN to responses that follow BS, BSdj, DR, GA, and LIN models is likely to produce thresholds that are inaccurate, strongly affected by data distributions, and that indicate a greater level of synchronicity than is supported by the actual models. Selecting an appropriate threshold analysis requires a careful understanding of the capabilities and limitations of the detection method and its applicability to the underlying data model (Daily et al. 2012). Characterization of individual taxon responses and detection of thresholds can be improved by comparing multiple alternative models (Qian and Cuffney 2012) to determine which model best represents the data and whether that model contains a threshold.

Community threshold

The community threshold is intended to provide an assessment of community responses by aggregating the standardized thresholds for the species in the community. The ecological significance of the community threshold depends on TITAN's ability to extract accurate thresholds that represent ecologically equivalent information. Unfortunately, the thresholds detected by TITAN are often imprecise and inaccurate, particularly if the underlying response model is linear or contains a threshold defined by a change in response rate (slope) or direction (BS, BSdj, DR, GA, LIN). TITAN thresholds can also represent very different characteristics of a species response, such as an abrupt change in average abundance (Fig. 9A), the limits of a species' distribution (Fig. 9A, B), or variation in the distribution of sampling errors (Fig. 9C, D) within a species response domain. The diagnostics provided by TITAN (*p*-values, confidence intervals) are not sufficient to assess which thresholds

are statistically meaningful and which thresholds represent ecologically equivalent information about the species responses. Without this information, combining TITAN thresholds into a community threshold carries a significant risk of combining ecologically disparate information into a single index. Before a community threshold can be created, each threshold identified by TITAN must be evaluated to establish that it is a valid threshold and to determine whether it represents equivalent characteristics of the species responses. This evaluation has to occur outside of the TITAN program and should involve examining multiple alternative models to identify the proper model form against which the TITAN-derived threshold can be compared (Qian et al. 2003, Qian and Cuffney 2012). Failure to conduct such in-depth analyses can lead to erroneous conclusions regarding the presence of thresholds and the effect of disturbance on the community. This situation could be problematic if the community threshold were used to establish criteria for protecting the resource, and the resulting threshold was not protective.

Potential thresholds and replication

$IndVal_{ij}$s derived independently of TITAN (Figs 4A–F, 5A–H, Online Appendices) were identical to those produced by the TITAN program (variable obsiv in TITAN) as long as the data set did not contain replicate samples. When replicate samples were present, TITAN reported $IndVal_{ij}$s for potential thresholds that corresponded to the midpoints between all disturbance values rather than between unique values. For example, TITAN identified 8 potential thresholds (35, 45, 50, 50, 50, 50, 55, and 65) in a data set containing abundances measured at disturbance values of 30, 40, 50, 50, 50, 50, 50, 60, and 70 rather than the 4 potential thresholds (35, 45, 55, 65) that corresponded to the midpoints of the 5 unique disturbance values (30, 40, 50, 60, 70). This approach is problematic because: 1) successive clusters associated with the replicates do not represent a change in disturbance, 2) no basis exists for determining the order in which replicate responses should be incorporated into successive clusters, and 3) successive clusters contain varying numbers of responses for the same disturbance level. As an example, the 4 potential thresholds at disturbance = 50 define 4 clusters that contain 1, 2, 3, and 4 of the replicate responses in the left-hand clusters, and 4, 3, 2, and 1 in the right-hand clusters, respectively. In contrast, the potential threshold at 45 would combine all the observations at disturbance $\leq$40 into one cluster and all values $\geq$50 into the other. Treating replicate samples as independent potential thresholds apportions the variability in replicate responses among multiple clusters representing the same disturbance levels. This variability should be associated with a single potential threshold and 2-group cluster that includes all replicates in the same cluster. The TITAN program should be modified to drop $IndVal_{ij}$s that are derived from replicate samples to better quantify response variability associated with disturbance levels.

Simplistic models

Our evaluation of the *IndVal* method and its implementation in TITAN is based on the analysis of relatively simple models. One could argue that these models are unrealistic given the large amounts of variability normally observed in species responses. However, using simple models has enormous advantages when it comes to evaluating the performance of a threshold-detection method (Daily et al. 2012). The form of the response model can be defined in precise mathematical terms (eqs 3–8), which allows method performance to be assessed across different model types (e.g., BS, BSdj, DR, SF, GA, and LIN). The characteristics that define each model (e.g., slopes, offsets, thresholds, replication, variability, occurrences, and distribution of data across the gradient) can be manipulated individually and collectively to identify effects on threshold detection. The use of simple models allowed us to assess the accuracy and precision of thresholds detected in TITAN for various response models and to identify the effects of the response domain (sampling and structural zeros) and data distributions on threshold detection, neither of which would have been feasible with more realistic and less well understood data sets. The philosophy that underlies using simple models for method evaluation is very simple: 1) a method that cannot accurately detect a threshold in a simple model that represents the best statistical conditions for threshold detection is unlikely to detect a threshold accurately in a more complicated situation, and 2) the behavior of an analytical method can be best understood when the characteristics of the underlying models are clearly understood prior to analysis.

Conclusions

The TITAN program has been advanced as a nonparametric method of threshold detection that can be applied to a wide variety of response models (Baker and King 2010) and that can detect thresholds at very low levels of disturbance (King et al. 2011). It is an easy-to-use method that is beginning to appear in the literature (Kail et al. 2012). However, our analysis of the performance of TITAN indicates that it has important limitations that need to be considered.

The problems with TITAN primarily arise from 3 sources: 1) the use of the *IndVal* method (Dufrêne and Legendre 1997) without consideration of its relation to the underlying response model (i.e., a discrete method applied to continuous data), 2) biases in statistics (μ_{ind}, σ_{ind}) derived from permutations of the data, and 3) the use of the same permutations to identify and test the statistical significance of thresholds. Collectively, these problems affect the accuracy, precision, and consistency of threshold identification. The concept of creating a community threshold by aggregating thresholds in TITAN is ecologically attractive, but great care must be exercised to establish that the thresholds extracted by TITAN are accurate and that they represent equivalent ecological characteristics of the species' responses. Combining disparate characteristics of species responses can easily lead to an aggregate threshold that has little ecological meaning and that is unsuitable for supporting management decisions or criteria development. Because the threshold detection method used by TITAN (*IndVal*) gives reliable results only for certain model forms (e.g., SF), it is imperative that the response of each species be evaluated carefully to know whether the response is appropriate for analysis with TITAN. The diagnostic tools provided with the TITAN program are not sufficient to do this evaluation. Instead, multiple alternative models should be evaluated and compared (Qian et al. 2003, Qian and Cuffney 2012) to determine which model best describes each species' response, whether that model contains a threshold, and whether it is amenable to TITAN analysis.

Acknowledgements

We thank Roxolana Kashuba, Ibrahim Alameddine, and Gerald McMahon for their insightful discussions and comments. Ian Waite (US Geological Survey [USGS], Portland, Oregon), John Van Sickle (US Environmental Protection Agency, Corvallis, Oregon), and Michael Lavine (University of Massachusetts, Amherst, Massachusetts) reviewed an early version of the manuscript and provided many valuable comments and suggestions. Comments and suggestions provided by 2 anonymous referees and Associate Editor Heikki Mykrä are also greatly appreciated. The research was completed while Song Qian was supported by the USGS through a USGS–Duke University cooperative agreement (08HQAG0121).

Literature Cited

ANDERSEN, T., J. CARSTENSEN, E. HERNANDEZ-GARCIA, AND C. M. DUARTE. 2009. Ecological thresholds and regime shifts: approaches to identification. Trends in Ecology and Evolution 24:49–57.

BAKER, M. E., AND R. S. KING. 2010. A new method for detecting and interpreting biodiversity and ecological community thresholds. Methods in Ecology and Evolution 1:25–37.

BANERJEE, M., AND I. W. MCKEAGUE. 2007. Confidence sets for split points in decision trees. Annals of Statistics 35: 543–574.

BRENDEN, T. O., L. Z. WANG, AND Z. M. SU. 2008. Quantitative identification of disturbance thresholds in support of aquatic resource management. Environmental Management 42:821–832.

BÜHLMANN, P., AND B. YU. 2002. Analyzing bagging. Annals of Statistics 30:927–961.

CLARKE, K. R., AND R. N. GORLEY. 2006. PRIMER v6: user manual/tutorial. PRIMER-E Ltd, Plymouth, UK.

CUFFNEY, T. F., S. S. QIAN, R. A. BRIGHTBILL, J. T. MAY, AND I. R. WAITE. 2011. Response to King and Baker: limitations on threshold detection and characterization of community thresholds. Ecological Applications 21:2840–2845.

DAILY, J. P., N. P. HITT, D. R. SMITH, AND C. D. SNYDER. 2012. Experimental and environmental factors affect spurious detection of ecological thresholds. Ecology 93:17–23.

DODDS, W. K., W. H. CLEMENTS, K. GIDO, R. H. HILDERBRAND, AND R. S. KING. 2010. Thresholds, breakpoints, and nonlinearity in freshwaters as related to management. Journal of the North American Benthological Society 29: 988–997.

DUFRÊNE, M., AND P. LEGENDRE. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67:345–366.

EFRON, B., AND R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman and Hall, New York.

ELLIOTT, J. M. 1977. Some methods for the statistical analysis of samples of benthic invertebrates. 2nd edition. Publication 25. Freshwater Biological Association, Ambleside, UK.

GROFFMAN, P., J. BARON, T. BLETT, A. GOLD, I. GOODMAN, L. GUNDERSON, B. LEVINSON, M. PALMER, H. PAERL, G. PETERSON, N. POFF, D. REJESKI, J. REYNOLDS, M. TURNER, K. WEATHERS, AND J. WIENS. 2006. Ecological thresholds: the key to successful environmental management or an important concept with no practical application? Ecosystems 9:1–13.

HOLLING, C. S. 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics 4: 1–23.

KAIL, J., J. ARLE, AND S. C. JÄHNIG. 2012. Limiting factors and thresholds for macroinvertebrate assemblages in European rivers: empirical evidence from three datasets on water quality, catchment urbanization, and river restoration. Ecological Indicators 18:63–72.

KING, R. S., AND M. E. BAKER. 2010. Considerations for analyzing ecological community thresholds in response to anthropogenic environmental gradients. Journal of the North American Benthological Society 29:998–1008.

KING, R. S., M. E. BAKER, P. F. KAZYAK, AND D. E. WELLER. 2011. How novel is too novel? Stream community thresholds

at exceptionally low levels of catchment urbanization. Ecological Applications 21:1659–1678.

MAY, R. M. 1977. Thresholds and breakpoints in ecosystems with a multiplicity of stable states. Nature 269:471–477.

MCCUNE, B., AND M. J. MEFFORD. 1999. PC-ORD. Multivariate analysis of ecological data. Version 4. MjM Software Design, Gleneden Beach, Oregon.

QIAN, S. S., AND T. F. CUFFNEY. 2012. To threshold or not to threshold, that is the question. Ecological Applications 15:1–9.

QIAN, S. S., R. S. KING, AND C. J. RICHARDSON. 2003. Two statistical methods for the detection of environmental thresholds. Ecological Modelling 166:87–97.

RESH, V. H. 1979. Sampling variability and life history features: basic considerations in the design of aquatic insect studies. Canadian Journal of Fisheries and Aquatic Sciences 36:290–311.

SONDEREGGER, D. L., H. N. WANG, W. H. CLEMENTS, AND B. R. NOON. 2009. Using SiZer to detect thresholds in ecological data. Frontiers in Ecology and the Environment 7:190–195.

STRINGHAM, T. K., W. C. KRUEGER, AND P. L. SHAVER. 2003. State and transition modeling: an ecological process approach. Journal of Range Management 56:106–113.

Received: 5 April 2012
Accepted: 16 January 2013