Of TITAN and straw men: an appeal for greater understanding of community data

Author: Baker, Matthew E.

Source: Freshwater Science, 32(2) : 489-506

Published By: Society for Freshwater Science

URL: https://doi.org/10.1899/12-142.1

Freshwater Science, 2013, 32(2):489–506

DOI: 10.1899/12-142.1
Published online: 26 March 2013

Of TITAN and straw men: an appeal for greater understanding of community data

Matthew E. Baker[1]
Department of Geography and Environmental Systems, University of Maryland, Baltimore County, Baltimore, Maryland 21250 USA

Ryan S. King[2]
Department of Biology, Baylor University, Waco, Texas 76706 USA

Abstract. Cuffney and Qian (2013) performed numerous simulations to demonstrate potential flaws in Threshold Indicator Taxa Analysis (TITAN), a method for interpreting taxon contributions to community change along novel environmental gradients. Based on their simulations, they concluded that: 1) TITAN is not an effective method for detecting different types of statistical thresholds in trend lines, 2) permutation results in highly significant *p*-values even for splits that are not thresholds, and 3) coincident change points may arise as an artifact of inaccuracies, imprecision, and systematic bias in both change-point estimation and TITAN's bootstrap. The critique raises some important concerns, but because of significant misunderstanding, it is based on analyses that violate basic assumptions of both TITAN and indicator species analysis (IndVal), and thus, constitutes a straw man that cannot be used to evaluate their performance. We demonstrate that the critique: 1) fundamentally misrepresents TITAN's primary goals; 2) simulates taxon abundances based on unrealistic statistical models that fail to represent important empirical patterns present in Cuffney and Qian's own published data sets (i.e., negative binomial distributions, frequent absences a function of the predictor); 3) tests TITAN's ability to identify breaks in trend lines distorted by log-transformation that do not match the greatest change in the simulated response, leading to misinterpretation of expected and previously documented behavior by TITAN as errors; 4) misinterprets TITAN's use of *p*-values while ignoring diagnostic indices of *purity* and *reliability* for identifying robust indicator taxa; and 5) asserts that bootstrapped change-point quantiles in TITAN are too narrow despite published results to the contrary. Last, in contrast to the claim that change-point synchrony may be an artifact of the technique, we show that: 6) analysis of published data using completely independent methods (i.e., scatterplots of abundance data or generalized additive models) also reveals synchrony in the nonlinear decline of numerous taxa in corroboration of TITAN and its underlying conceptual model. Thus, Cuffney and Qian have not identified any serious limitations of TITAN because their critique is based on misinterpretation of TITAN's assumptions and primary objectives. However, their critique does highlight the need for clarification of the appropriate uses, potential misuses, and limitations of TITAN and other methods for ecological analysis.

Key words: bioassessment, biodiversity, community analysis, indicator species, ecological thresholds, conservation, stream integrity, statistical analysis.

"*If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.*"

Leo Breiman (2001b, p. 199)

Cuffney and Qian (2013; hereafter C&Q) present a case for flaws in Threshold Indicator Taxa Analysis (TITAN; Baker and King 2010), a method for detecting and interpreting individual taxon contributions to patterns of community change along novel environmental gradients. TITAN uses binary partitioning by indicator value (IndVal; Dufrêne and Legendre 1997) scores to identify taxon-specific change points in community sampling units ranked along environmental gradients (fig. 1 in Baker and King 2010). Permutation methods are used to compare the magnitude of observed scores to random scores, and

[1] E-mail addresses: mbaker@umbc.edu
[2] ryan_s_king@baylor.edu

significance of change in robust indicator taxa is assessed via indices of *purity* (consistent direction), *reliability* (consistent magnitude), and narrow quantiles of change-point location across bootstrap replicates. Correspondence in the distribution of change-point locations for pure and reliable taxa and narrow bootstrap quantiles around a distinct peak in normalized changes summed across taxa (i.e., sum[z]) are used to infer evidence for thresholds in community composition and structure.

C&Q set up a series of simulations to assess sample partitioning by IndVal scores and the permutations used in TITAN to test their ability to detect statistical thresholds (i.e., slope changes or disjunctions) in trend lines representing different underlying patterns of abundance and differing levels of introduced variability. Their primary assertions are that: 1) TITAN's use of IndVal is not an effective method for detecting thresholds in species abundance trend lines, 2) the permutation methods result in highly significant p-values even for splits that are not thresholds, and 3) coincident change points are at least partially an artifact of inaccuracies, imprecision, and systematic bias resulting from both change-point estimation and subsequent bootstrapping.

TITAN is a relatively new method that combines well known techniques from quantitative ecology, and we expect detailed scrutiny. Our terminology is derived from these techniques, but the approach used in TITAN may require clarification to distinguish our use and interpretation from other perspectives. By design, TITAN combines a few of the most vetted and commonly used methods in ecological data analysis. IndVal (Dufrêne and Legendre 1997) is one of the most highly cited methods in the recent ecological literature (1849 Thomson Reuters® Web of Knowledge citations and 2761 Google Scholar® citations as of 10 January 2013; also see Podani and Csányi 2010). Binary partitioning forms the basis for frequently used methods such as Classification and Regression Trees (Breiman et al. 1984, De'Ath and Fabricius 2000), Random Forests (Breiman 2001a), and Boosted Trees (De'Ath 2007), whereas Random Forests, Boosted Trees, and Piecewise Regression (Toms and Lesperance 2003) use bootstrapping. However well these techniques are understood, their specific application within TITAN is novel and, thus, may generate confusion among potential users. Therefore, we welcome this opportunity to clarify the fundamental misunderstandings of TITAN upon which C&Q's critique rests.

Here we demonstrate that because of misunderstanding of TITAN, the critique's core premise that TITAN was designed to identify statistical thresholds in individual trend lines deemed by C&Q to represent ''species abundances'' is false. We then explore the nature of their simulation data to show that they are unrealistic and deviate strongly from patterns of real data assumed by both IndVal and TITAN. We show that entirely expected IndVal behavior is mistaken for errors because of misinterpretation, unrealistic simulations, data transformations that distort trend lines (C&Q analyzed transformed data but present TITAN's results with untransformed data), and modeled ''thresholds'' in trend lines that do not match the location of greatest change in simulated variables. As a result, the critique's claims of inaccuracy, imprecision, and bias are not valid. Further claims regarding statistical significance and confidence intervals based on the faulty premises are also misguided, and subsequent speculation about the causes of change-point synchrony in previously published work (e.g., King and Baker 2010, 2011; King et al. 2011) is not supported by any analysis and shown here to be unfounded. C&Q's critique raises some important questions surrounding threshold analysis, but because it does not present TITAN accurately and because its rationale for testing TITAN is at odds with all published descriptions or applications of the method, the results cannot reliably inform any judgment or evaluation of TITAN's ability to identify regions of maximum taxon-specific or community change. C&Q's claims should not discourage use of TITAN for analyzing taxa contributions to community change and detecting and interpreting ecological community thresholds, but they do highlight the need for clarification of uses and misuses of new analytical techniques.

A Flawed Premise

The premise at the core of C&Q's critique is that TITAN, through the use of the IndVal statistic, seeks to isolate statistical thresholds in a taxon's response to a disturbance gradient. We agree with C&Q that thresholds may be defined as a disproportionate ecosystem response to an incremental change in a driver (e.g., sensu Groffman et al. 2006). However, if C&Q's premise were true, then any curve that satisfies the statistical definition of a threshold would represent a valid test of IndVal partitioning and, by extension, of TITAN. Because the premise is false, threshold constructs used to identify ''errors'', the validity of subsequent test data, and the conclusions that follow all require much closer scrutiny.

TITAN does *not* seek to identify statistical thresholds in trend lines of a single response variable, and we have never described it this way. The goal

statement in our original methods paper and a summary statement from our subsequent description for managers were explicit:

> ''We introduce a new analytical approach…with the goals of (i) exploring and identifying abrupt changes in both the occurrence frequency and relative abundance of individual taxa along an environmental, spatial or temporal gradient; (ii) quantifying uncertainty around locations of abrupt change; and (iii) estimating the relative synchrony and uncertainty of those changes as a nonparametric indicator of a community threshold.'' (Baker and King 2010, p. 26)

> ''We developed TITAN specifically to address the problems described in our paper… [TITAN] can distinguish the response direction, magnitude, and location of change in individual taxa, and provide an assessment of uncertainty about the location and synchrony of taxon change points as evidence for community thresholds.'' (King and Baker 2010, p. 1006)

In TITAN, it is *immaterial* whether or not taxa exhibit a statistical threshold or other specific response form, because *change* at the taxon level is all that is required to develop and assess evidence for thresholds at the community level. All taxon thresholds represent change, but not all taxon changes constitute thresholds. IndVal partitioning is used to identify objectively the greatest change in abundance or occurrence frequency for individual taxa (i.e., change points or split points). Abrupt changes are indicated when resampling produces only limited modification of the location of the change point, whereas gradual changes produce broader distributions of resampled change-point locations. Said another way, IndVal partitioning *may* correspond to steepening trend lines and other so-called threshold patterns, but it is not required or even expected to do so in many cases (e.g., linear, wedge-shaped increases of tolerant taxa along urbanization or other novel gradients; King and Baker 2010, Bernhardt et al. 2012). Thus, what may seem to be a minor semantic distinction between statistical thresholds introduced by C&Q and change points is, in fact, a crucial mischaracterization of TITAN because it accounts for many of the disparities C&Q count as ''errors'' in their simulations. Furthermore, this misunderstanding leads to flawed interpretation of most of TITAN's core functions (see discussion below).

If species exhibit modal distributions with respect to most natural environmental gradients (sensu Whittaker 1967, Austin and Smith 1989), novel gradients produced via anthropogenic activity can fall outside

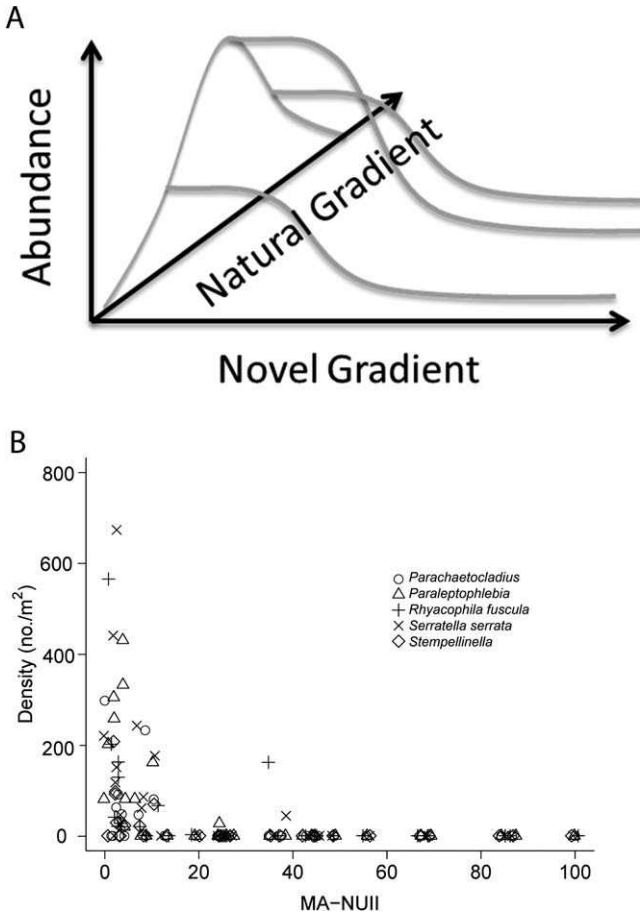

FIG. 1. Response of taxon abundance to novel environmental gradients showing idealized departures from a unimodal response often observed along natural gradients (A) and observed synchronous responses of 5 robust declining taxa identified by Threshold Indicator Taxa Analysis (TITAN) in US Geological Survey North American Water Quality Assessment (NAWQA) data (Boston; richest targeted habitat data) along a metro-area normalized urbanization intensity index (MA-NUII; Cuffney et al. 2010) (B).

the range of conditions experienced by a species over evolutionary time (Hobbs et al. 2006, Fox 2007, Williams and Jackson 2007). Depending on the intensity of the novel gradient and species-specific tolerances, we generally expect a decline in abundance and occurrence with increasing departures from natural conditions, similar to the truncated distributions often observed at extremes of natural gradients (Fig. 1A).

> ''In the context of ecological communities, we interpret a threshold to mean that the frequency and/or abundance of taxa will increase or decrease sharply at some level of an environmental gradient, such that an incremental change in a driver such as

urban intensity results in a disproportionately large change in community structure relative to elsewhere along the gradient.'' (King and Baker 2011, p. 2833)

TITAN is designed to compare change across taxa because we are interested in detecting correspondence at the community level. In principle, a series of wedge-shaped declines in species counts with synchronous IndVal change points could still produce a community response of interest in TITAN even though no single taxon exhibits threshold behavior. However, because of the gradual nature of the decline there would be broad uncertainty about the location of the change and, thus, only weak evidence for a community threshold. In practice, we have observed a variety of responses to novel gradients, and we refer interested readers and potential users to our description in fig. 7 of King and Baker (2010).

C&Q's misinterpretation of TITAN's purpose is understandable because *change point* and *threshold* are sometimes used as synonyms in the literature. Moreover, the term threshold appears to have different meanings for different people. We may have inadvertently contributed to the confusion because some of the simulations initially used to test TITAN involved underlying step-functions in the probability distribution of individual taxa (figs 3, 4 in Baker and King 2010), and we have documented that many taxa declining along anthropogenic disturbance gradients do show disproportionate change in frequency and abundance with an incremental change in the gradient (i.e., a threshold; King and Baker 2010, 2011, King et al. 2011). However, in TITAN, a *change point* really means what its name suggests (see next sections). We urge readers not to conflate change points (or ''split points'' following Breiman et al. 1984) exhibited by individual taxa (and detected by IndVal maxima) with ecological *thresholds* that, in TITAN, are inferred at the community level by considering aggregate changes across taxa.

Unrealistic and Inappropriate Simulated Data

In evaluating the C&Q critique, it is worth considering whether the simulated data they used represent a valid test of IndVal performance. As a measure of change, binary partitioning by IndVal may appear rather clunky, but the statistic itself was developed to deal elegantly with a well known, but sometimes overlooked property of biological survey data—sparse matrices (many 0s) with high levels of variability (i.e., negative binomial distributions with overdispersion; Pielou 1984, Legendre and Legendre 1998, McCune and Grace 2002, Podani and Csányi 2010). In most community matrices, absences can occur in nearly any sampling unit, occurrence frequency is not independent of mean abundance, and both may be a function of the gradient. We have repeatedly emphasized that TITAN was designed explicitly for this type of data, and we have generated our simulated abundances from distributions (i.e., negative binomial) that have some or all of these key characteristics (Baker and King 2010, p. 26; King and Baker 2010, p. 1000).

In their criticism of TITAN, Cuffney et al. (2011) and C&Q used 2 types of simulated data to test IndVal performance. Type 1 simulated data (figs 1–4 by C&Q) consists of abundances that follow specific trend lines and that have either *no absences* (i.e., Type 1 data $<25\%$ variance) or far fewer than would be expected for $>95\%$ of observed taxa (Fig. 2). Variance, when included, is *generated by a normal distribution* centered on the trend line (Type 1 data, variance 1–100%). Type 2 simulated data are similar to Type 1, except that C&Q force the response curve to intersect the *x*-axis away from where they define ''actual'' thresholds (e.g., fig. 5 by C&Q). In a few Type-2 simulations, C&Q place a group of 0 counts (absences) in locations inversely related to the trend line (e.g., fig. 9C, D by C&Q).

Collectively, Type 1 and 2 simulations are unambiguously unrealistic representations of taxon abundance data. In Fig. 2, we show a comparison of C&Q's simulation data to observed data from their own published work (Cuffney et al. 2010). The authors offer no substantiation for their claim that their simulated data are common. We dispute this claim and point to representative examples from our data (fig. 3 in King and Baker 2010, fig. 6 in King et al. 2011) and C&Q's empirical observations (Figs 1B, 2) that indicate otherwise. Had C&Q attempted to reproduce their own published empirical taxon distributions, we are confident they would have rejected their simulation data (as we do) on statistical, if not ecological, grounds.

The patterns C&Q simulated have more in common with their previous analysis of idealized *aggregate/composite community metrics* (see Cuffney et al. 2010, Qian and Cuffney 2012, Qian et al. 2012) than with realistic examples of *taxon* abundances. TITAN was developed for a fundamentally different kind of data, and we have presented it as an unambiguous *alternative* to aggregate metrics (King and Baker 2010). However, the reader need not rely on either of our viewpoints regarding properties of taxon counts. Following Zuur et al. (2010), we encourage investigators to plot representative examples of their data (taxon abundances along environmental gradients) and decide whether TITAN is appropriate for

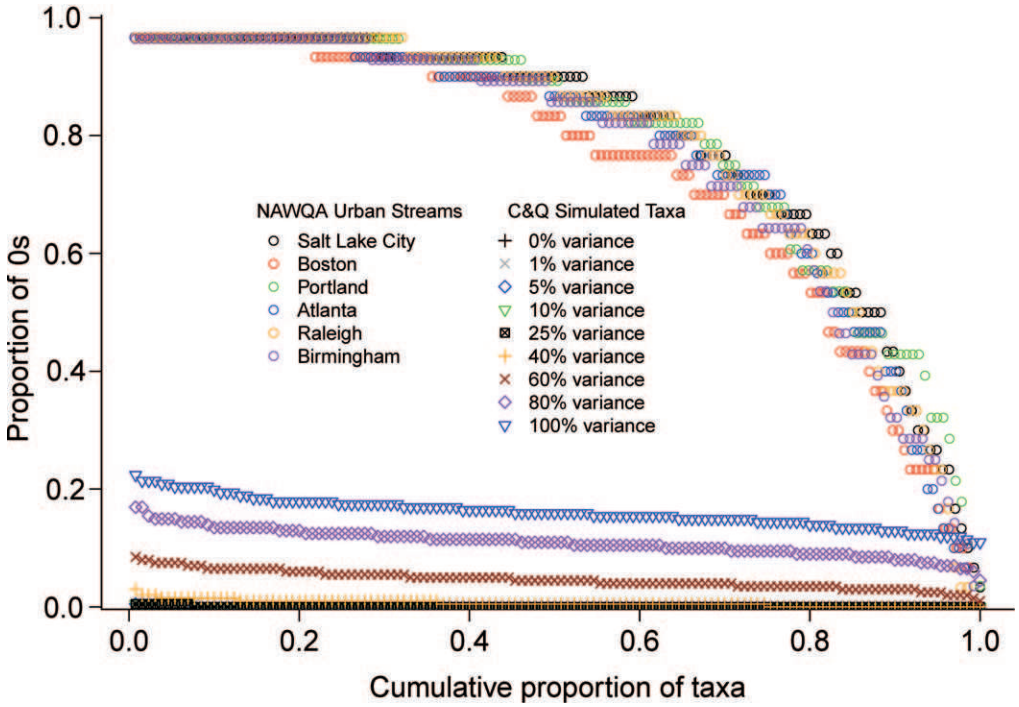

FIG. 2. Comparison of patterns of abundance with Type 1 (abundance trend lines with 0–100% normally distributed variability) simulation data used by Cuffney and Qian (C&Q) (2013) showing the proportion of 0s relative to the cumulative proportion of taxa vs empirical data from US Geological Survey North American Water Quality Assessment (NAWQA) of urban streams (Cuffney et al. 2010), which show strong evidence of frequent absences across most taxa.

their question (e.g., see Figs 1B, 3A–H). We also encourage users to plot these relationships post hoc for robust taxa identified by TITAN (e.g., Fig. 1B).

We developed more realistic simulations using negative binomial distributions with overdispersion in R (version 2.15.1; R Development Core Team, Vienna, Austria; Appendix S1; available online from: http://dx.doi.org/10.1899/12-142.1.s1). We set mean abundance across observations to range from 2 to 10 individuals, which approximates the range of empirical values we have observed in our analyses of various survey data sets (e.g., King and Baker 2010, 2011, Bernhardt et al. 2012). We set mean abundance to be a function of the response models suggested by fig. 1 of C&Q (Fig. 3A–D). However, we used the negative binomial distribution, which incorporated nonnormal error distributions and allowed specification of an appropriate level of overdispersion in the response, in our simulations. We set the dispersion parameter to 0.5 to simulate moderate overdispersion consistent with natural variability in taxon abundances in community data sets (e.g., the National Water Quality Assessment [NAWQA] data in Cuffney et al. 2010; Appendix S2; available online from: http://dx.doi.org/10.1899/12-142.1.s2). We did not log-transform the data prior to analysis.

Figure 3A–H illustrates the results of our negative binomial simulations. The most obvious characteristic is that observed values (Fig. 3E–H) only vaguely resemble the original model forms from which they derive (Fig. 3A–D). Noisy counts make discerning a trend in abundance difficult at best, and it is easy to imagine piecewise or quantile regression fits that look nothing like the underlying models. However, all of the data can be interpreted by binary partitioning, especially when one considers both changes in abundance and occurrence frequency across the gradient. Because each realization of the underlying distribution produces different patterns of both abundance and occurrence, the actual point of greatest change also varies. However, change points identified across 20 simulations for each response form illustrate that responses detected by TITAN are tracking the greatest changes in the underlying response model.

Wide bootstrap quantiles spanning the zone of greatest change for each response form in Fig. 3I–L are also highly informative because they demonstrate the high degree of uncertainty regarding the point of greatest change in the data. Linear, broken-stick, or dose-response models all have a zone, rather than a single point, of greatest change (steepest slope) that

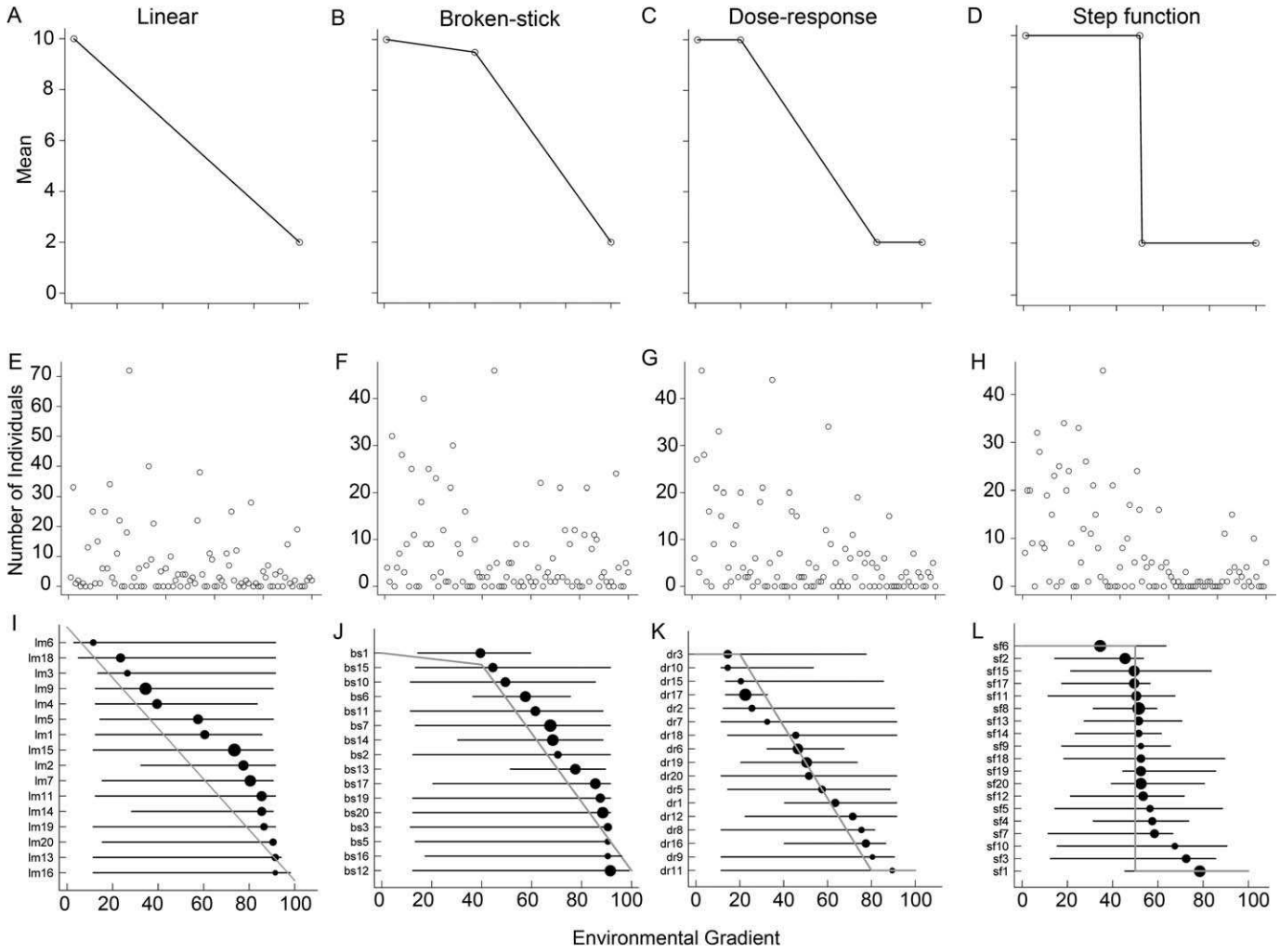

FIG. 3. Threshold Indicator Taxa Analysis (TITAN) change points obtained for 4 hypothetical taxon abundance distributions developed from negative binomial mean response curves from 2 to 10 individuals and a moderate overdispersion parameter of 0.5 for linear (A), broken-stick (B), dose-response (C), and step-function (D) declines. Panels E–H (row 2) show observed values generated by 1 simulation corresponding to each of the models in the same column. Panels I–L (row 3) show the distribution of indicator value (IndVal) maxima (i.e., change points, with symbol diameter proportional to *z*-score) and 90% bootstrap quantiles from 500 replicates across 20 simulations corresponding to each response curve. A total of 16 (linear), 16 (broken-stick), 17 (dose-response), and 19 (step-function) simulated taxa were deemed robust indicators (purity ≥ 0.95, reliability ≥ 0.95). Only pure and reliable indicator taxa are shown in panels I–L.

begins at the first break in the trend line and can continue for much of the remainder of the gradient (Fig. 3I–K). This continuum of greatest change is captured accurately by the distribution of IndVals and their respective bootstrap quantiles.

IndVal was developed specifically for the noise inherent in the type of data represented in Fig. 3E–H, and such variability underscores the fundamental problem in C&Q's idealized approach for critiquing IndVal. Similar constructs for analyzing community metric data may be familiar to many readers (e.g., Brenden et al. 2008, Dodds et al. 2010, Qian and Cuffney 2012), but C&Q's response models have little to do with the patterns encountered in site × species community matrices. Expecting a statistic developed specifically for noisy counts to detect with precision subtle breaks in trend lines or other smooth response forms ignores the reasons for developing IndVal in the first place.

C&Q's simulations contribute to the fallacy of their critique by disregarding known properties of biological survey data, an omission we are at a loss to explain given their recent exploration of negative binomial models (e.g., Qian et al. 2012). Taxa with a few absences occur in their simulation data (Fig. 2), but these are clearly the exception rather than the rule.

Without absences, the IndVal statistics assessed by C&Q frequently are reduced to comparisons of mean abundance, for which neither TITAN nor IndVal is ideal. Furthermore, without the nonnormal variability associated with overdispersion, C&Q's test data lack key signals that IndVal is specifically designed to exploit. In sum, C&Q's unrealistically biased simulations lead to inappropriate tests of TITAN's performance and a misapplication of the method as we designed it. We agree with C&Q that simple models can be heuristically useful, but only when they capture essential phenomena. C&Q's simulations are focused solely on detection of their specific response forms, but their simulations do not represent essential attributes of community data (McCune and Grace 2002).

Change-Point Identification

The core premise of C&Q's argument is false. Thus, the interpretation that follows is largely misguided and, when combined with unrealistic data, unfounded. However, we think it might be helpful to explain what C&Q found and why they found it. Many practitioners accustomed to thinking about biotic change in terms of nonlinear trend lines produced by aggregate community metrics (e.g., Walsh et al. 2005, Brenden et al. 2008, Cuffney et al. 2010, Qian and Cuffney 2012) may wonder at discrepancies presented by the critique, despite the flaws in its analytical construct.

An important modification made by C&Q may not be obvious to the casual reader because it is mentioned only briefly in their methods. C&Q specifically developed their idealized curves to test TITAN's ability to detect threshold response forms, and then $\log(y + 1)$-transformed their simulation data prior to analysis. We have transformed our empirical data in the past when using TITAN (e.g., Baker and King 2010, King and Baker 2010), but only after careful consideration of the effect of transformation on properties of sparse community matrices with predominant 0s and very high overdispersion for most taxa. More recently, we have found that TITAN is relatively robust to, and transformation is unnecessary for, all but the most extreme cases of overdispersion (King and Baker 2013). Transformations are common in ecological data analysis, but we consider transformation of C&Q's simulation data unnecessary because their simulation data are neither sparse nor overdispersed (Fig. 2). Moreover, C&Q presented their untransformed idealized curves with TITAN's change points, even though they analyzed transformed data. Figure 4A–F illustrates the effect of this transformation on the data actually analyzed with TITAN for 3 selected response forms (linear [LIN], broken-stick [BS], and Gaussian [GAU]). In the transformed data, smoothly varying functions are distorted (evident in curves that should have uniform slopes), the magnitudes of breaks in slope are altered, and specific locations of greatest changes in abundance may be shifted. Thus, presentation of untransformed response curves and TITAN's change points is potentially misleading to readers who might fail to notice that TITAN was performed on, and change points identified from, transformed values rather than those presented in the figures.

C&Q indicate that TITAN failed to identify ''actual'' threshold locations in trend lines representing BS, dose-response (DR), and GAU response forms unless they were step changes in mean abundance. The ''actual'' thresholds in some of these simulations may be legitimate split points of analytical interest, but they depart from the literature definition that C&Q cite (Groffman et al. 2006) when they do not correspond to locations of greatest change in abundance (Table 1), particularly after log-transformation (Fig. 4A–F). For example, in their BS and DR simulations, C&Q define thresholds at breaks in slope even though all locations anywhere along the steepest part of the trend line are equivalent locations of disproportionate change relative to an incremental change in the environment (i.e., an ecological threshold; Groffman et al. 2006). For GAU response curves, C&Q define their ''actual'' threshold at the peak of the unimodal response, where change is at a *minimum* (e.g., figs 4F, G, 5D, E by C&Q). Considering these departures from preceding literature, TITAN's stated goals (see A Flawed Premise above), our previous description of IndVal behavior along step function (SF) (taxon A), Gaussian (taxon C), and dose-response (taxon D) models (fig 2 in Baker and King 2010), distortion of curves via log-transformation, and graphical representation of the point or zone of greatest change in taxon responses to environmental gradients (fig. 7 in King and Baker 2010), it should not be surprising that IndVal maxima did not correspond well with C&Q's ''actual'' thresholds.

Given known behavior of IndVal partitioning when confronted with various response forms, it also should come as no surprise that IndVal performs well when there is a step change in mean abundance or that such partitioning is sensitive to the pattern of absences because the statistic was designed specifically to detect such changes. As the simulations in Fig. 3A–L demonstrate, binary partitioning is one of the few ways to interpret the noise in count data, and IndVal integrates 2 of its strongest signals. The IndVal

TABLE 1. Differences with Cuffney and Qian (C&Q) (2013) regarding interpretation of indicator value (IndVal) partitioning of simulated response forms by Threshold Indicator Taxa Analysis (TITAN).

Response form	Actual location	C&Q interpretation	Baker and King response
Step function	Step break	TITAN and IndVal identify this location accurately	Expected, especially for data without absences
Gaussian	Mode of distribution	TITAN and IndVal inaccurate; they always identify points to either side of the correct location	Greatest change is to either side of the mode, especially when transformed; TITAN identifies these locations accurately; C&Q's definition is inconsistent with literature
Dose-response	Upper and lower slope breaks	TITAN and IndVal inaccurate; they always identify points between the correct locations	For partitioning algorithm, greatest change is always along declining limb; TITAN working as expected
Broken stick	Slope break	TITAN and IndVal inaccurate; they identify points other than slope break	Partitioning works on simulated data, not idealized curve; IndVal maximum depends on magnitude of slope break and range of introduced variability; TITAN working as expected, especially for transformed data when there is no change in occurrence frequency

notation used by C&Q (eq. 1 in C&Q) and similar notation elsewhere (Dufrêne and Legendre 1997, Legendre and Legendre 1998, McCune and Grace 2002) shows that the term maximized across groups by IndVal scores is the product of 2 components: 1) the proportion of occurrences (O_{i1}/n_{i1}) and 2) the mean abundance in 1 sample group ($\bar{A}_{i1}$) relative to total mean abundance across both groups ($\bar{A}_{i1} + \bar{A}_{i2}$). Notice that the difference between a species' presence (occurrence = 1, abundance > 0) and its absence (occurrence = 0, abundance = 0) affects both terms, and therefore, has a compounded effect on IndVal.

Without absences, the IndVal statistic in TITAN finds a split that produces the greatest difference in mean abundance. Without changes in abundance, the statistic reverts to a comparison of within-group occurrence frequency. As used in TITAN, IndVal can be viewed as the difference in abundance-weighted frequency between 2 groups. IndVal scores use *both* terms to assess whether each taxon shows a distinct association with a group of samples split according to an environmental variable.

As C&Q discovered, and as should be apparent from published descriptions of both IndVal and TITAN, absences matter. When variability is introduced into each of their idealized response curves (fig. 3 by C&Q), we would expect similar patterns of such ''errors'' to occur until increased variability starts to produce counts of 0 (absences). When distributed randomly and uniformly across the gradient (i.e., Fig. 5A, Type 1 data, >60% variation) rather than in correlation with abundance (e.g., Fig. 3E–H), absences add random noise to IndVal partitioning rather than interpretable signal. When absences are introduced nonrandomly, unrealistically, and away from decreases in abundance (e.g., fig. 9C, D by C&Q), they necessarily affect the ability of IndVal partitioning to detect change because they reduce apparent differences in mean abundance. However, the change points shown in fig. 9C, D by C&Q are not just unrealistic, they are also an artifact of misapplied data transformations that are not apparent because the figure shows untransformed data and ignores the results of TITAN's bootstrap. In Fig. 6A, transformation enhances the effect of unrealistic absences by minimizing the break in abundance and producing an increasing (not decreasing) change point with broad uncertainty. In Fig. 6B, the uncertainty is so great that the resulting change point is both impure and unreliable (<0.95) over 500 bootstrap replicates. When we analyzed the untransformed data provided in C&Q's supplemental data (available online at: http://dx.doi.org/10.1899/12-056.1.s2), TITAN identified pure and reliable change points at or near 52 (90% quantiles: 48.5–58.5) in the case of fig. 9C by C&Q (Fig. 6C) and at or near 85 (90% quantiles: 74.0–87.0) in the case of fig. 9D by C&Q (Fig. 6D). We do not know how many ''errors'' identified by C&Q are attributable to transformations because we have not reanalyzed all of their data, but this example does raise concerns regarding their findings and interpretation.

When absences are introduced via truncated distributions (i.e., Type 2 data, fig. 5 by C&Q), the IndVal statistic is no longer driven by a comparison of mean abundance. Because C&Q have so few absences in their simulations, the IndVal statistic is maximized at or near the point of intersection of each response curve with the *x*-axis when all abundance and occurrence are entirely in one partition (e.g., Fig. 5B)

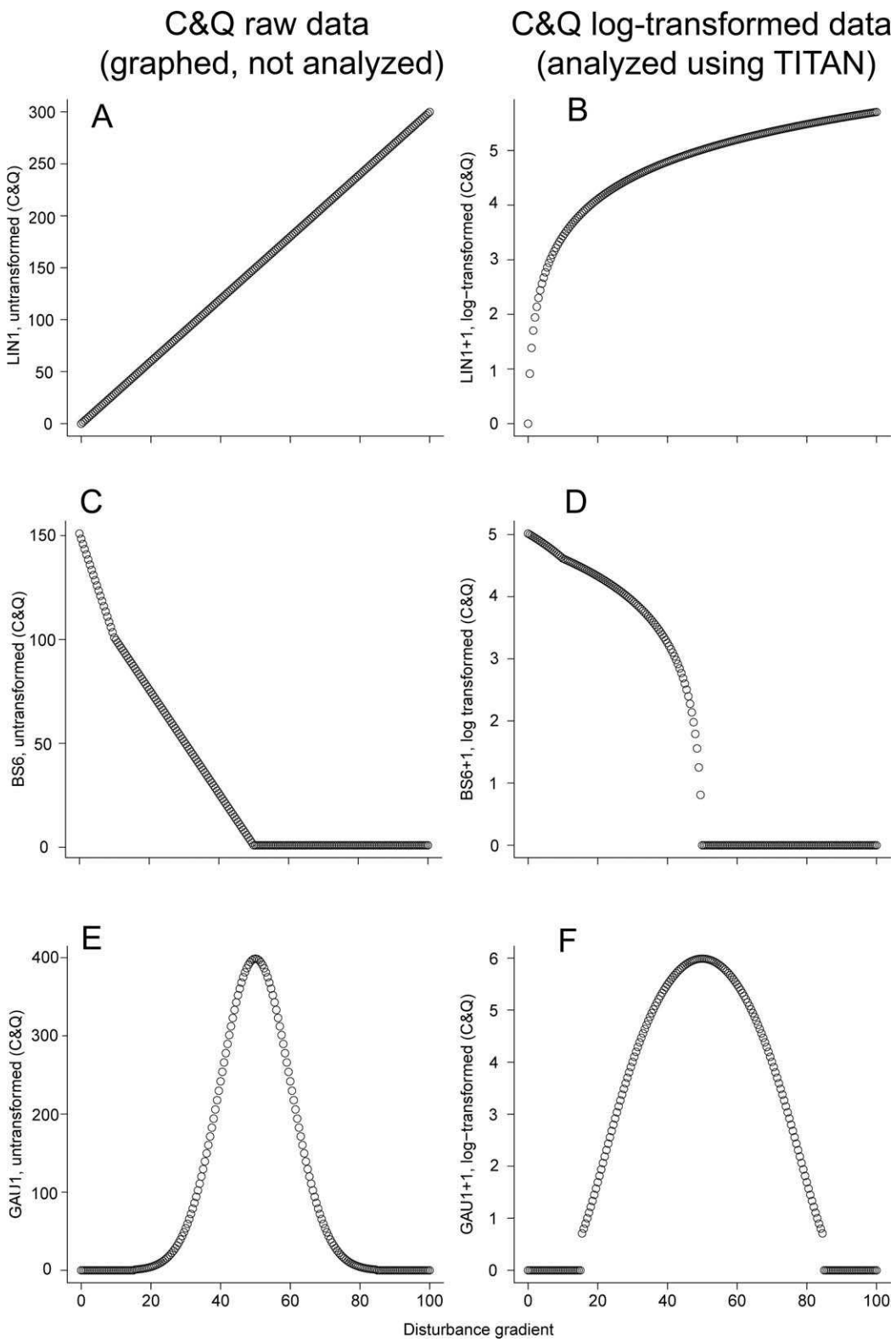

FIG. 4. Comparison of selected response curves used in tests of Threshold Indicator Taxa Analysis (TITAN) presented by Cuffney and Qian (C&Q) (2013) (A, C, E) with the log(y + 1)-transformed data actually analyzed by Cuffney and Qian (2013) (B, D, F). y-axis titles refer to specific models presented in their online appendices (available at: http://dx.doi.org/10.1899/12-052.1.s1). LIN = linear, BS = broken stick, GAU = Gaussian.

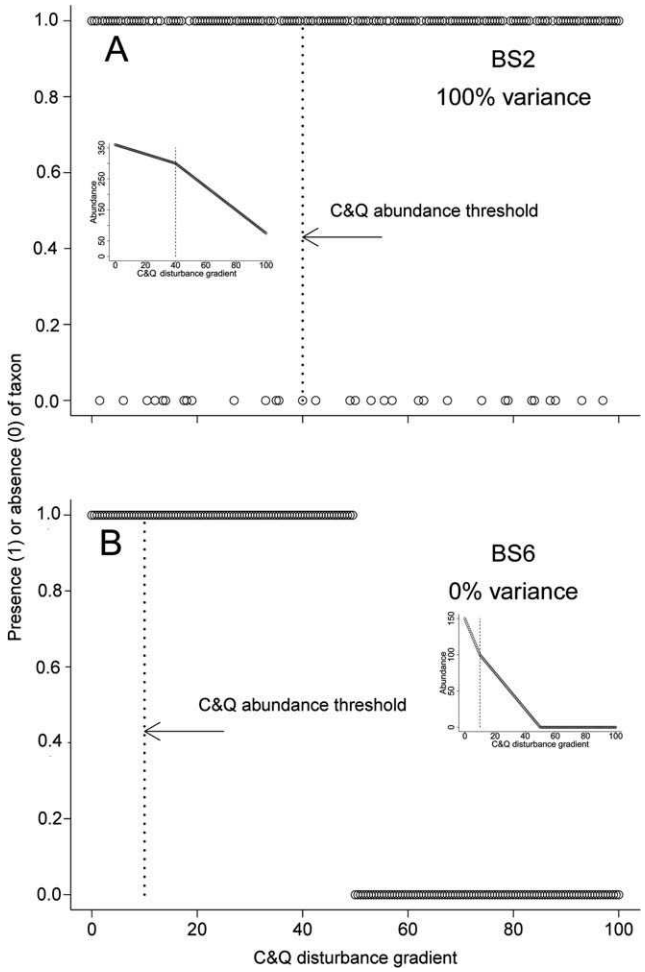

FIG. 5. Occurrence frequency response in broken-stick (BS) models used in Cuffney and Qian (C&Q) (2013) representative of problems with Type 1 (BS2 with 100% introduced variance) (A) or Type 2 (BS6 with 0% introduced variance) data (B). Insets show the modeled taxon abundance response curve. See text for details of Type 1 and Type 2 data. C&Q state that a threshold occurs at a break in the slope of abundance decline (vertical dashed line). Threshold Indicator Taxa Analysis (TITAN) integrates changes in mean relative abundance and occurrence frequency. In A, the pattern of absences is unrelated to the pattern of abundance or to the gradient. In B, the break in the slope is a trivial signal (especially after log[y + 1] transformation) when compared to the truncated distribution. C&Q count these differences as errors in TITAN, but they represent expected behavior when TITAN is presented with unrealistic data or unreasonable definitions.

and especially when the data are log-transformed (Fig. 4D). In addition to problems with C&Q's ''actual'' threshold locations relative to the zone of greatest change in the response, C&Q's definition of thresholds in Type 2 data is problematic. Without absences that increase in frequency with decreasing mean abundance, we would expect TITAN to find the coincident loss of abundance and occurrence instead of a break in slope or the peak of a unimodal distribution (e.g., taxa E and F in fig. 2 by Baker and King 2010). In empirical data, truncated distributions usually are preceded by low abundances and a marked increase in absences (fig. 7 by King and Baker 2010), so TITAN typically identifies this zone rather than the point of intersection. Here, we differ philosophically with C&Q because we think that a point of sudden extinction along an environmental gradient is both statistically and ecologically a location of far greater change in a taxon's response than either a break in slope or a peak in a unimodal distribution, especially when compared to the subtle change in slopes used to define Type 2 thresholds.

The results presented by C&Q as problems with IndVal partitioning are, in fact, entirely predictable given their simulated trend lines, log-transformation of data, confusion regarding TITAN's purpose, and threshold criteria that depart from literature definitions. Thus, their critique does not pose a concern for users analyzing empirical community data with typical negative binomial distributions and frequent 0s. C&Q suggest that comparing multiple alternative models offers a more comprehensive approach to threshold detection across taxa. We agree, but we note that the alternative models they suggest (i.e., Qian and Cuffney 2012) may be appropriate for some community metrics, but would violate their own assumptions (e.g., homogeneity of errors and normal distribution of observations) if applied to a matrix of taxon counts (e.g., Figs 1B, 3E–H).

IndVal, *z*-Score Maxima, and Skew

In their critique, C&Q pointed out differences between change points identified by independently calculated IndVal maxima and those identified by TITAN, which are based on *z*-score maxima. The difference occurs because, as C&Q note, *z*-scores generated via permutation are different each time TITAN runs. This difference was designed to draw attention to uncertainty arising from near-maximal change-point realizations (i.e., IndVals with near-maximum magnitudes can be identified at different locations along the gradient by different permutations; e.g., see pattern of IndVal *z*-scores relative to IndVals in Fig. 6A, C). C&Q's more serious claim is that the permutations introduce bias into the identification of change points. We are aware of this possibility, and C&Q are correct that it can happen, especially with certain kinds of data. In general, the effect is to move the change point closer to the center

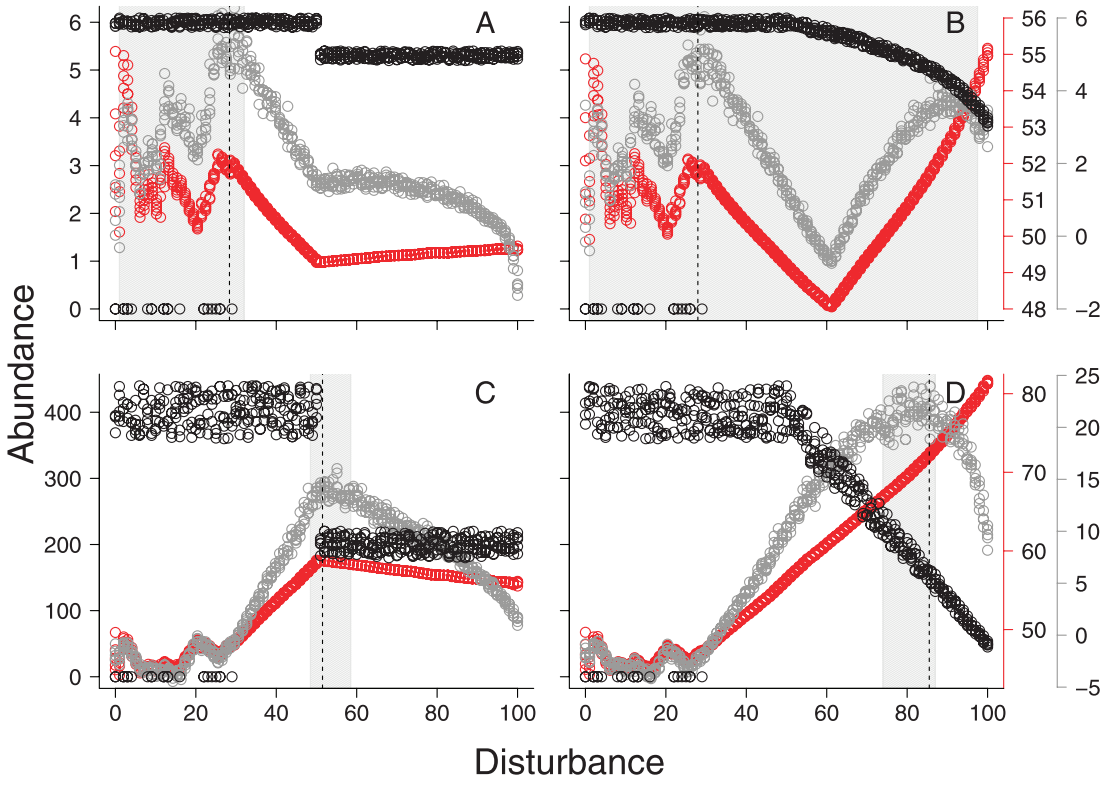

FIG. 6. Scatterplots of log(y + 1)-transformed simulated taxon abundances (black) corresponding to a step-function (A) and broken-stick (B) abundance response patterns with introduced absences analyzed by Cuffney and Qian (C&Q) (2013) and the untransformed data presented by C&Q in their fig. 9C (C) and fig. 9D (D). C&Q stated that by introducing absences to idealized abundance patterns, they were able to generate large differences in observed change points identified by Threshold Indicator Taxa Analysis (TITAN), but these differences, change-point directionality, and associated uncertainty were an artifact of the transformation. When we provided the untransformed data (obtained from C&Q supplemental data files; available at: http://dx.doi.org/10.1899/12-052.1.s2) to TITAN, we obtained results similar to those obtained by C&Q when they analyzed the same response patterns without absences in their fig. 9A and B (C, D). Observed change points are shown as a dashed vertical line, with 90% bootstrap quantiles as a shaded area. Change points in panel A denote an increase in occurrence frequency, change points in panel B are impure and unreliable, and change points in panels C and D are pure and reliable decreasers. IndVal scores for all candidate change points are shown in red (1st right-hand axis), whereas IndVal z-scores are shown in grey (2nd right-hand axis).

of observations along the x-axis. All TITAN objects output in R contain IndVals and z-scores for each taxon, but updates of TITAN will, by default, use IndVal maxima to simplify interpretation and to reduce concerns about bias.

Our own comparisons indicate that z-score bias has not affected interpretation of our published analyses of community thresholds along urbanization gradients. If bias were substantial, its effect would be to make apparent responses to urbanization less extreme (i.e., toward the center of observations) and *not* more extreme at the low end of the gradient. When a taxon's change point is the result of a strong break in its distribution (i.e., SF3 in figs 4D, 5A–F by C&Q; Figs 5B, 6C), the variation introduced by permutation is likely to be trivial compared to the uncertainty generated by the bootstrap, which C&Q ignore. When used to describe an indistinct or smoothly varying and unrealistic IndVal maximum, such as those provided by the linear (LIN) or BS curves (figs 4A–C, H, 7A–C in C&Q; Fig. 6D), the permutation can introduce enough variability to change the rank order of z-scores. However, the same variability is introduced repeatedly during bootstrapping, broadening change-point quantiles (what C&Q call taxon-specific confidence intervals; e.g., Fig. 3I–L), and making the observed location suspect according to prescribed use of TITAN (Baker and King 2010, King and Baker 2010).

C&Q raise the issue of skewed sample distributions as a further claim of bias in change-point locations and present examples in their fig. 7A–C. C&Q's skew simulations suffer from a lack of absences and a smoothly declining pattern of abundance, which

makes the ''threshold'' they define subtle, especially after transformation (data not shown in fig. 7A–C by C&Q). Under these specific conditions of reduced signal, bias introduced by the *z*-score maximum is enhanced, pushing the identified change point closer to the center of observations along the *x*-axis.

Distributional skew within a sample set is a serious concern with important implications for nearly any form of regression (most assume a normal or uniform distribution in the predictor) and for binary partitioning because the location of candidate change points (considered as the midpoint between successive, ranked observations) is not evenly distributed. Thus, in regions of reduced sample density, change-point precision will necessarily be reduced. Analyses of regression trees suggest that skewed samples can affect the pattern of variance when applied to idealized and smoothly varying functions (i.e., linear; Daily et al. 2012). However, IndVal partitioning (as opposed to *z*-score partitioning) in TITAN appears relatively robust to strongly skewed samples when parameters from observed data are used with more realistic negative binomial simulations (see appendix 1 by Bernhardt et al. 2012 for a detailed example). To aid users in better understanding the implications of skewed samples in specific data sets, forthcoming versions of TITAN will offer guidance for generating data-set-specific simulations and IndVal, rather than *z*-score, partitioning.

Statistical Significance

Another premise of C&Q's argument is that TITAN uses permutations to evaluate the significance of the IndVal or *z*-score maximum for each taxon (i.e., its observed change point). In their simulations, C&Q find many examples where apparently erroneous threshold locations nonetheless produce very small ($p < 0.01$) *p*-values. C&Q argue that because the permutation procedure is used to identify thresholds (by *z*-scores) in addition to testing their statistical significance (by *p*-values), the significance tests, and by implication TITAN itself, are invalid.

We agree that IndVal *p*-values should not be used to assess the significance of taxon-specific thresholds. However, C&Q's claim is incorrect because TITAN does not use *p*-values to compare the quality of one candidate split-point location relative to others. Rather, it uses *p*-values to compare the magnitude of the change relative to random noise and to provide an initial filter for candidate partitions assessed during bootstrapping. In fact, although permutation results are used to assist in evaluating many candidate partitions, *no statistical evaluation* is made of the comparison among the resulting *z*-scores that produces the maximum (i.e., a test for a threshold change). We bear some blame here because our descriptions have not been explicit enough, and we did not anticipate C&Q's misinterpretation. In our published descriptions of TITAN, we noted that Dufrêne and Legendre (1997) used permutation for evaluating the statistical significance of IndVal scores. The *p*-values in TITAN were calculated in precisely the same way as in IndVal analysis to provide continuity with the original method. However, Dufrêne and Legendre (1997) permuted IndVal scores for a small number of clustered sets defined a priori by independent (usually environmental) criteria in a hierarchical cluster analysis and used IndVal scores to select optimal levels of grouping in a hierarchical cluster. Dufrêne and Legendre (1997) did not permute IndVal scores repeatedly for each value of an environmental variable—as is done in TITAN—and we were careful to note this distinction (Baker and King 2010, p. 27). Therefore, as we discussed in our original description (underline added):

''…many (>40%) <u>randomly</u> generated distributions were nonetheless deemed to contain significant change points following permutation. This pattern illustrated how frequent or abundant taxa with only modest differences in IndVals between groups are often statistically significant ($p \leq 0.05$) in large data sets despite dubious ecological significance. However, such patterns are readily distinguished from more meaningful responses through the diagnostic use of reliability and purity. We note that those taxa deemed significant by permutation do not always achieve reliability or purity ≥ 0.95, but taxa with reliability or purity ≥ 0.95 are by definition significant at $p \leq 0.05$ or much lower.'' (Baker and King 2010, p. 35)

Thus, we reported our finding that the *p*-values from the permutation (or, for that matter, the magnitude of *z*-scores) were not, by themselves, a useful criterion for significance of change points, and their interpretation is not Bonferroni-adjusted for repeated computing at each possible change point. Instead, *p*-values indicate whether observed changes are large enough to be distinct from random noise. TITAN uses *z*-scores to normalize the relative magnitude of change across taxa with inherently different abundance patterns (Baker and King 2010, pp. 28–29; after Dufrêne and Legendre 1997), as in any generic standardization approach (e.g., Euclidean distances). C&Q misunderstood TITAN's use of the *p*-values suggested by Dufrêne and Legendre (1997) because they misinterpreted the objective of IndVal partitioning in TITAN. However, by disregarding

purity and reliability, they have misrepresented the method.

C&Q's criticism of change-point significance highlights another—largely implicit—premise that underpins their case against TITAN: disregarding the integral role of bootstrap resampling. Their critique implies that significance of any IndVal-based change point in TITAN is assessed *solely* through permutation, but they completely ignore the role that bootstrap results play in interpreting and filtering IndVal results. Our original description of TITAN was explicit about the integral role of the bootstrap:

> "The bootstrap procedure is necessary because unlike *a priori* group classification required by indicator species analysis, optimal group partitioning along x is initially unknown in TITAN, and is in fact the objective of the analysis. Whereas the permutation procedure is used to estimate the probability that an equal or larger IndVal could be obtained from random data, the bootstrap procedure estimates uncertainty around change point locations as well as consistency in the response direction of each taxon. Variability in change-point location, directionality, and magnitude constitute the information content of the *indicator response* for each taxon." (Baker and King 2010, p. 27)

In TITAN, the 3-fold information content of each taxon's indicator response is used to assess its significance, and this 3-fold approach is an addition to the original conceptualization of the IndVal statistic. TITAN does not provide an estimate of overall likelihood for a change point. Instead, it uses the smallest p-values across all candidate change points derived from permutation of *each* bootstrap replicate to assess whether response magnitudes are consistently large (i.e., *reliability*; Baker and King 2010). Taxa that fail to maintain small p-values across bootstrap replicates are considered unreliable indicators because the strength of their apparent response depends strongly on the specific sample analyzed (e.g., taxa with large abundances in few samples). Likewise, taxa that fail to maintain fidelity to the observed response direction (e.g., unimodal distributions or weak signals variably interpreted as increasing or decreasing) are considered impure indicators of change (*purity*; Baker and King 2010). We have found that reliability is somewhat redundant with purity (e.g., taxa with purity ≥ 0.95 are usually reliable) except in cases of a modal response (i.e., reliable taxa that are not pure). TITAN uses purity and reliability—both outcomes of bootstrap resampling—to remove those taxa with weak signal relative to background noise, reduce error, and minimize the effects of bias.

TITAN's bootstrap diagnostics were not used by Cuffney et al. (2011) and C&Q as significance criteria in their simulations because of the time needed to complete the procedure. Instead, they evaluated the bootstrap diagnostics based on analysis of small subsets of their simulation data. Thus, their assertion that bootstrapped confidence limits are too narrow is not based on empirical evidence. The theoretical arguments C&Q present are all based on split-point problems in which the objective is to estimate the true population standard deviation by resampling observed values. However, TITAN does not use the bootstrap to estimate a true standard deviation (i.e., as if it were attempting to detect taxon-specific thresholds), nor does TITAN estimate an observed standard deviation. Thus, how C&Q can demonstrate that its quantiles are too narrow is not clear. The IndVal statistic is sensitive to the distribution of relative mean abundance and occurrence frequency within each partition. Resampling is used to examine a range of alternative observation sets to assess relative sensitivity in the location of change points across taxa. Moreover, we explicitly discouraged strict interpretation of taxon-specific change-point quantiles as confidence intervals (Baker and King 2010, p. 28).

C&Q refer to simulations in their fig. 8 to support their claim, but these results do not provide evidence that taxon-specific bootstrap quantiles are too narrow. Instead, the figure shows unrealistic abundance trend lines without absences, where IndVals would normally be maximized at one end of the gradient. Under such conditions resampling naturally tends to find a range of alternate change points. The bootstrapped change-point distributions span the entire gradient. Thus, they (appropriately) convey broad uncertainty associated with finding a difference in mean abundance along each trend line. We do not see how these distributions could get any wider. In our simulations, true change points sampled using negative binomial generators were frequently captured more often than expected for 90% quantiles, results indicating that intervals might be too *wide* rather than too narrow. We prefer to be conservative in conveying such uncertainty.

Given the empirical patterns described above, C&Q's subsequent endorsement of Bayesian approaches over those used in TITAN is contradictory. We are unaware of any rigorous test of the method they promote, and the sole publication where both approaches were compared (Qian et al. 2003) revealed that Bayesian intervals were *narrower* than intervals similar to those calculated by TITAN (2-sample t-test assuming unequal variances, $n = 8$ observations, $p < 0.01$). Thus, C&Q's claims regarding the bootstrap in

TITAN are not valid. Without mentioning TITAN's core diagnostic indices and without using the bootstrap results in their interpretations, the critique does not actually test TITAN, but selectively misapplies portions of the analysis for purposes that have never been suggested or used in a published or unpublished work.

Synchronicity and Community Change

C&Q imply that coincident changes described in our published work are at least partially an artifact of inaccuracies, imprecision, and bias in TITAN's identification of change points. Of all of C&Q's claims, this suggestion of artifactual synchrony is perhaps the least well supported and most speculative. None of the concerns raised about *z*-scores, skewed samples, or minimum split size indicate that change points across taxa should be *more* coincident as a result.

For example, C&Q noted that *z*-scores introduce *stochastic variability* into change-point maxima. Stochastic variability should decrease (not increase) the likelihood that they will generate synchrony. Moreover, the *z*-score bias that C&Q demonstrated over smooth, monotonic functions is clearly larger and more directional than what they found over more abrupt changes (e.g., SF3 in fig. 4D, 5A–F by C&Q). Nevertheless, the effect of skew on *z*-score change points identified by C&Q is *toward the center* of the sample distribution and not toward the margin of the gradient. Bias toward the center of the distribution should decrease the likelihood of artifactual synchrony in change points at gradient extremes, where our empirical analyses have detected synchrony (King and Baker 2010, 2011, King et al. 2011). As C&Q suggest, using a minimum cluster size does make identifying change points at extremes impossible in observed data, but identifying change points at extremes certainly is possible during bootstrap resampling, a point that C&Q did not mention. In any case, the support for change points at gradient extremes is necessarily weak (i.e., low sample densities and number of observations), a problem that should be highlighted by the bootstrap. On the other hand, TITAN identified coincident change points in our published simulations (Baker and King 2010, King and Baker 2010) with high accuracy. In all of our simulations, TITAN's diagnostic filters were far more likely to interpret coincidently declining taxa as either not robust or asynchronous than to indicate synchrony where none existed.

Change-point synchrony is well known along thermal (e.g., Wehrly et al. 2003) and nutrient gradients (e.g., King and Richardson 2003) among others. In TITAN, synchronous change at the community level is identified by sharp peaks in sum(*z*) maxima with narrow confidence intervals, and corresponding alignment of change points for robust indicator taxa. TITAN is focused on detecting change, but *synchronous* change can be identified by other means. Verification of change-point synchrony is as easy as plotting the abundances of robust (i.e., pure and reliable) indicator taxa vs the gradient (e.g., Fig. 1B, Appendix S2). C&Q did not use this approach in their simulations. Yet another approach involves fitting models to the relative abundance of each taxon (e.g., Qian et al. 2012). Below we describe our use of all 3 approaches, and we encourage those skeptical of TITAN's component analyses to do the same.

We used previously published US Geological Survey NAWQA data from Boston, Massachusetts (richest targeted habitat data [RTH]; Cuffney et al. 2010), which we previously analyzed at the community level and presented graphically using combined Quantitative–Qualitative (QQ) presence–absence data along a metro-area normalized urbanization intensity index (MA-NUII) in King and Baker (2011). Cuffney et al. (2010) concluded that this assemblage exhibited linear change in response to urbanization. TITAN showed that 13 taxa present at low levels of urbanization declined sharply below 12 MA-NUII, leading to a community threshold estimate at 11.6 MA-NUII (with a 90% confidence interval of 3.9–25.0) (Fig. 7A, B). Eleven additional taxa declined between 20 and 53 MA-NUII, indicative of additional change that was less synchronous but certainly not a linear function of the remaining gradient. TITAN also described 4 increasing taxa whose change points ranged from 3 to 61 MA-NUII with 90% confidence intervals at the community level of 10.4 to 77.0 MA-NUII.

We analyzed the same data set by fitting negative binomial generalized additive models (GAMs) using the *gam* function in the *mgcv* package in R following Zuur et al. (2009). We standardized densities of taxa with significant ($p < 0.05$) model fits to the maximum predicted density for each taxon for clarity (i.e., response curves could not exceed a value of 1). The data points ($n = 30$) were too few for us to interpret most of the resulting curves with great confidence, but nonetheless, the results are relevant. Of the 37 taxa with ≥ 3 occurrences selected by TITAN or GAMs, 23 were identified by both GAMs and TITAN, 9 were identified by GAMs but not TITAN, and 5 were identified by TITAN but not GAMs (Appendices S2, S3; available online from: http://dx.doi.org/10.1899/12-142.1.s3). The primary pattern was strong agreement between TITAN and an independent method, a

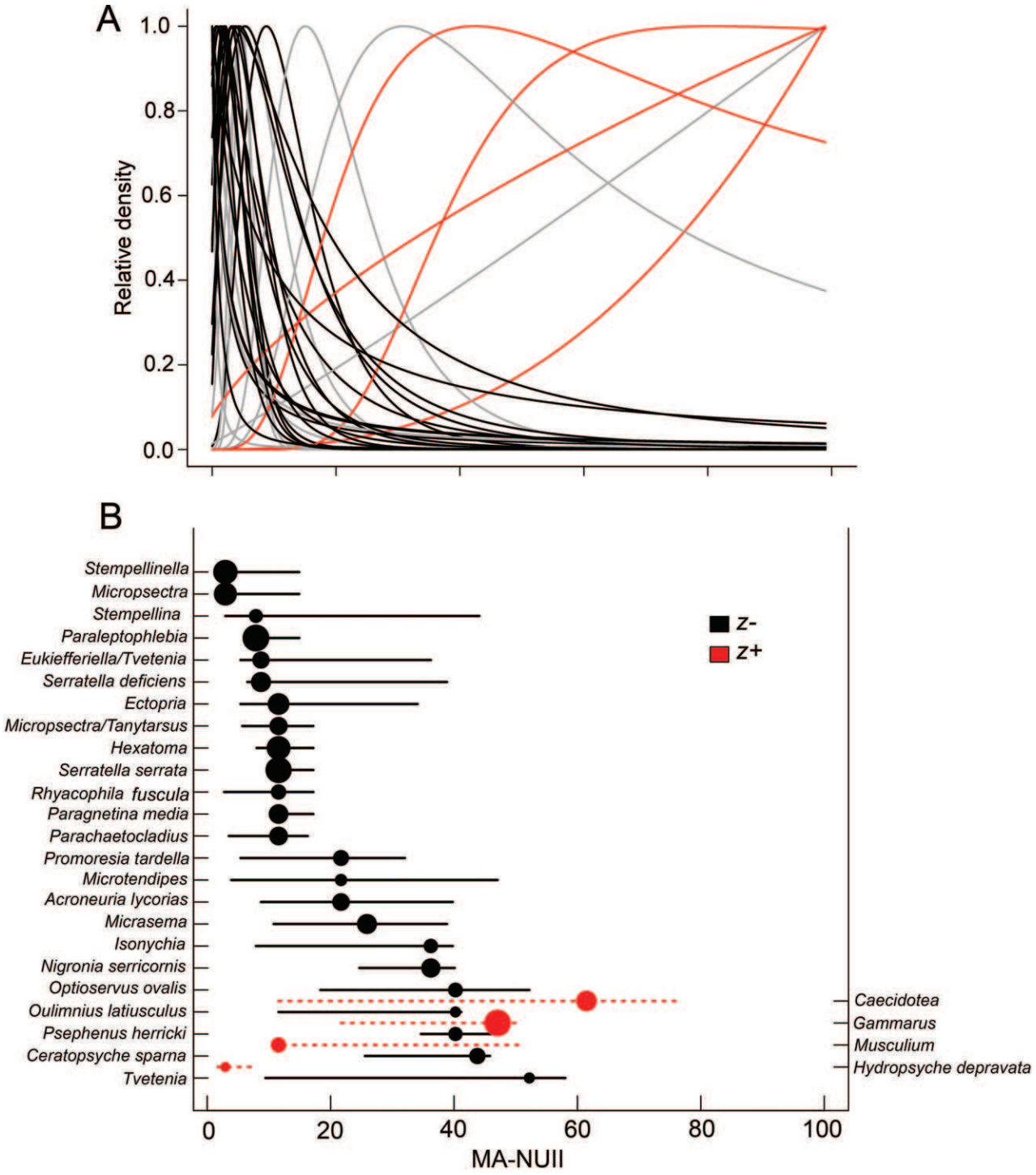

FIG. 7. Plots of significant negative binomial Generalized Additive Model (GAM) fits to taxa abundance with ≥3 occurrences for a metro-area normalized urbanization intensity index (MA-NUII) (A) and Threshold Indicator Taxa Analysis (TITAN) change points with 90% bootstrap quantiles for robust indicator taxa showing increasers (red) or decreasers (black), with symbol size proportional to IndVal *z*-scores (B). In panel A, black lines indicate robust declining indicator taxa from panel B ($n = 19$), red lines are robust increasing taxa ($n = 4$), and gray lines are taxa that did not meet purity and reliability criteria in TITAN ($n = 9$). Many taxa corroborate TITAN results of declines at <12 MA-NUII, with other declines occurring with greater urbanization. Two increasing taxa change most steeply <15 MA-NUII, whereas one changes at ~50, and a 4th exhibits greater change when MA-NUII > 60. See Appendices S2, S3 for detailed results and scatterplots of each taxon in response to MA-NUII, respectively.

plurality of taxa declining within a narrow band of low levels of MA-NUII, and independent corroboration of change-point synchrony.

Conclusion

Informed and thoughtful criticism is part of the scientific process. As the authors of TITAN, we welcome constructive critiques of the method or suggestions for improvement, and we have received many from academic colleagues and insightful users. The approach used in TITAN is novel, and its superficial similarity to other, well-known techniques may actually hinder initial understanding among analysts well versed in quantitative methods. Therefore, we bear responsibility to distinguish TITAN's goals and to explain clearly and carefully how it works in our publications. We note that several independent groups have interpreted our published descriptions successfully and applied TITAN in a manner consistent with our original intent (Kail et al. 2012, Payne et al. 2013, Smucker et al. 2013). On the other hand, our experience with others (e.g., Cuffney et al. 2011) has convinced us that greater explanation is necessary (King and Baker 2013). In return, we expect a fair and complete representation of TITAN from its critics.

The critique offered by C&Q shows evidence of misunderstanding of TITAN's primary functions. Crucial components of its analysis were omitted, simulated data for which we would recommend other approaches were tested, and the data were log-transformed, a step we would not recommend. Many of the thresholds C&Q used as a baseline for comparison with TITAN violated literature definitions they themselves cited, and at least some of the results they attributed to TITAN were potentially misleading to uninformed readers. It is easy to imagine that readers accustomed to looking at trends in multimetric indicator responses might ask "how TITAN does that" even though it does not, and thereby conflate a distortion of TITAN with more familiar approaches. We hope that our explanations have clarified serious misunderstandings and will help users assess whether TITAN is appropriate for their data.

C&Q did identify 2 valid issues (i.e., *z*-score bias, extreme sample skew) that could be problematic for users working to identify change points with small data sets, disturbance extremes, or the occasional ubiquitous taxon. These concerns are easily addressed via modification of TITAN's output (i.e., IndVal maxima are output as part of every TITAN object) or post hoc simulation (see appendix 1 in Bernhardt et al. 2012) with minor effect on TITAN's core analysis. TITAN is relatively unique in its goal of deconstructing community responses into taxon-specific change as a complement to ecotoxicology, species distribution modeling, and community analysis. We will continue working to refine and improve TITAN's performance (and the clarity of our explanation) as we and other investigators explore new applications and discover new challenges.

We developed TITAN to address what we view as a growing concern in community ecology that can be broadly defined as over-reliance on summary metrics (e.g., taxon richness, ordination scores) and biotic indices (e.g., Ephemeroptera, Plecoptera, Trichoptera richness, indices of biotic integrity) in system-level analysis, assessment, and management (Suter 2009, Baker and King 2010, King and Baker 2010, 2011). Like many investigators, we struggled with various existing analytical techniques in both basic and applied contexts until we realized that our problem was not statistical but conceptual. Understanding any level of a hierarchical system requires appreciating the behavior of components and broader contextual constraints (Allen and Starr 1982, O'Neill et al. 1986). It should not be controversial to state that to comprehend community behavior fully, we must at least consider what individual taxa are doing, just as surely as we need to examine spatial or environmental contexts.

Previously, we demonstrated that a simulated community built entirely of step-function decliners, wedge-shaped increasers, and random noise produced single-variable community metrics (e.g., number of taxa) where nothing but smooth and monotonic declines were evident (King and Baker 2010). We hoped such findings would call attention to the need for species-level analysis during community-level investigations, regardless of whether TITAN or other appropriate analytical techniques are used. The surge of recent papers on species distribution modeling for enhancing bioassessment supports this view (e.g., Olden et al. 2006, Elith and Leathwick 2009, Esselman and Allan 2011). If a better approach exists for linking community and species-level analyses, we suspect it does not involve fitting a population of candidate threshold regression models to each species distribution within a sampled assemblage.

We have difficulty imagining an underlying rationale for analyzing community change that requires community thresholds to be composed solely of responses that match threshold criteria or ecologically equivalent change among component taxa. Different species exhibit different tolerances and can occupy different trophic positions, leading to distinct responses along environmental gradients. We see no a priori reason to

expect that all species will respond in the same manner or at the same level along a novel environmental gradient or to constrain community-level analysis by such an assumption. Thus, the very existence of coincident change points is exciting from a theoretical and practical standpoint. The information aggregated by TITAN is *change*: change that is robust to permutation, change that is robust to resampling, change that maintains its direction and magnitude (i.e., remains pure and reliable). We encourage users to explore that change as they interpret community data.

Acknowledgements

We thank the many users of TITAN who have shared their problems, confusion, successes, and insights with us. They have enriched our understanding. Emily Bernhardt, Jason Taylor, Chris Swan, and Stu Schwartz provided helpful comments on an early draft of this manuscript. We are grateful to John Van Sickle and 1 anonymous referee for their close reading and constructive criticism of our submission.

Literature Cited

ALLEN, T. F. H., AND T. B. STARR. 1982. Hierarchy: perspectives for ecological complexity. University of Chicago Press, Chicago, Illinois.

AUSTIN, M. P., AND T. M. SMITH. 1989. A new model for the continuum concept. Vegetatio 83:35–47.

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.

BERNHARDT, E. S., B. D. LUTZ, R. S. KING, J. P. FAY, C. E. CARTER, A. M. HELTON, AND D. CAMPAGNA. 2012. How many mountains can we mine? Assessing regional degradation of central Appalachian rivers by surface coal mining. Environmental Science and Technology 46:8115–8122.

BREIMAN, L. 2001a. Random forests. Machine Learning 45: 5–32.

BREIMAN, L. 2001b. Statistical modeling: the two cultures. Statistical Science 16:199–231.

BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE. 1984. Classification and regression trees. Wadsworth Statistics Series, Chapman and Hall/CRC Press, Washington, DC.

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

CUFFNEY, T. F., R. B. BRIGHTBILL, J. T. MAY, AND I. R. WAITE. 2010. Responses of benthic macroinvertebrates to environmental changes associated with urbanization in nine metropolitan areas. Ecological Applications 20:1384–1401.

CUFFNEY, T. F., AND S. S. QIAN. 2013. A critique of the use of indicator species scores for identifying thresholds in species responses. Freshwater Science 32:471–488.

CUFFNEY, T. F., S. S. QIAN, R. B. 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.

DE'ATH, G. 2007. Boosted trees for ecological modelling and prediction. Ecology 88:243–251.

DE'ATH, G., AND K. E. FABRICIUS. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

DODDS, W. K., W. H. CLEMENTS, K. GIDO, R. H. HILDERBRAND, AND R. S. KING. 2010. Thresholds, breakpoints, and nonlinearity in aquatic ecosystems 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.

ELITH, J., AND J. R. LEATHWICK. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

ESSELMAN, P. C., AND J. D. ALLAN. 2011. Application of species distribution models and conservation planning software to the design of a reserve network for the riverine fishes of northeastern Mesoamerica. Freshwater Biology 56: 71–88.

FOX, D. 2007. Back to the no-analog future? Science 316: 823–825.

GROFFMAN, P. M., J. S. BARON, AND T. BLETT. 2006. Ecological thresholds: the key to successful environmental management or an important concept with no practical application? Ecosystems 9:1–13.

HOBBS, R. J., S. ARICO, J. ARONSON, J. S. BARON, P. BRIDGEWATER, V. A. CRAMER, P. R. EPSTEIN, J. J. EWEL, C. A. KLINK, A. E. LUGO, D. NORTON, D. OJIMA, D. M. RICHARDSON, E. W. SANDERSON, F. VALLADARES, M. VILÀ, R. ZAMORRA, AND M. ZOBEL. 2006. Novel ecosystems: theoretical and management aspects of the new ecological world order. Global Ecology and Biogeography 15:1–7.

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., AND M. E. BAKER. 2011. An alternative view of ecological community thresholds and appropriate analyses for their detection. Ecological Applications 21: 2833–2839.

KING, R. S., AND M. E. BAKER. 2013. Use, misuse, and limitations of Threshold Indicator Taxa Analysis (TITAN) for natural resource management. *In* G. Gunten-

spergen (editor). Ecological thresholds for management applications. Springer, New York (in press).

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.

KING, R. S., AND C. J. RICHARDSON. 2003. Integrating bioassessment and ecological risk assessment: an approach to developing numerical water-quality criteria. Environmental Management 31:795–809.

LEGENDRE, P., AND L. LEGENDRE. 1998. Numerical ecology. Developments in ecological modelling. Elsevier, Amsterdam, The Netherlands.

MCCUNE, B., AND J. B. GRACE. 2002. Analysis of ecological communities. MjM Software Design, Gleneden Beach, Oregon.

OLDEN, J. D., M. K. JOY, AND R. G. DEATH. 2006. Rediscovering the species in community-wide predictive modelling. Ecological Applications 16:1449–1460.

O'NEILL, R. V., R. L. DEANGELIS, J. B. WADE, AND T. F. H. ALLEN. 1986. A hierarchical concept of ecosystems. Princeton University Press, Princeton, New Jersey.

PAYNE, R. J., N. B. DISE, C. J. STEVENS, D. J. Gowing, BEGIN PARTNERS. 2013. Impact of nitrogen deposition at the species level. Proceedings of the National Academy of Sciences 110:984–987.

PIELOU, E. C. 1984. The interpretation of ecological data: a primer on classification and ordination. John Wiley and Sons, New York.

PODANI, J., AND B. CSÁNYI. 2010. Detecting indicator species: some extensions of the IndVal measure. Ecological Indicators 10:1119–1124.

QIAN, S. S., AND T. F. CUFFNEY. 2012. To threshold or not to threshold? That's the question. Ecological Indicators 15: 1–9.

QIAN, S. S., T. F. CUFFNEY, AND G. MCMAHON. 2012. Multinomial regression for analyzing macroinvertebrate assemblage composition data. Freshwater Science 31: 681–694.

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

SMUCKER, N. J., N. E. DETENBECK, AND A. C. MORRISON. 2013. Diatom responses to watershed development and potential moderating effects of near-stream forest and wetland cover. Freshwater Science 32:230–249.

SUTER, G. W. 2009. A critique of ecosystem health concepts and indices. Environmental Toxicology and Chemistry 12:1533–1539.

TOMS, J., AND M. L. LESPERANCE. 2003. Piecewise regression: a tool for identifying ecological thresholds. Ecology 84: 2034–2041.

WALSH, C. J., A. H. ROY, J. W. FEMINELLA, P. D. COTTINGHAM, P. M. GROFFMAN, AND R. P. MORGAN. 2005. The urban stream syndrome: current knowledge and the search for a cure. Journal of the North American Benthological Society 24:706–723.

WEHRLY, K. E., M. J. WILEY, AND P. W. SEELBACH. 2003. Classifying regional variation in thermal regime based on stream fish community patterns. Transactions of the American Fisheries Society 132:18–38.

WHITTAKER, R. H. 1967. Gradient analysis of vegetation. Biological Reviews 42:207–264.

WILLIAMS, J. W., AND S. T. JACKSON. 2007. Novel climates, no-analog communities, and ecological surprises. Frontiers in Ecology and the Environment 5:475–482.

ZUUR, A. F., E. N. IENO, AND C. S. ELPHICK. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1:3–14.

ZUUR, A. F., E. N. IENO, M. J. WALKER, A. A. SAVELIEV, AND G. M. SMITH. 2009. Mixed effects models and extensions in ecology with R. Springer, New York.

Received: 24 September 2012
Accepted: 23 January 2013