Pergamon

PII: S0025-326X(97)00049-0

Marine Pollution Bulletin, Vol. 34, No. 11, pp. 913–922, 1997

0025-326X/97 $17.00+0.00

Optimizing Temporal Sampling Strategies for Benthic Environmental Monitoring Programs

RAYMOND W. ALDEN III*, STEPHEN B. WEISBERG†, J. ANANDA RANASINGHE† and DANIEL M. DAUER‡
Applied Marine Research Laboratory, Old Dominion University, Norfolk, VA 23529, USA
†*Versar, Inc., 9200 Rumsey Road, Columbia, MD 21045, USA*
‡*Department of Biological Sciences, Old Dominion University, Norfolk, VA 23529, USA*

Trends in benthic macrofaunal monitoring data are often confounded by natural seasonal fluctuations in abundance. We used eight years of seasonally-collected benthic biological data to investigate the relative advantages of multi-season vs single season sampling strategies for assessing trends in the context of these seasonal fluctuations. The effect of seasonality on sampling strategy was examined in three ways. First, we tested long-term trends in benthic responses for homogeneity among seasons with the premise that a single season sampling strategy would be inappropriate if the direction of trends differed among seasons. Second, we compared the power for trend detection of several sampling regimes, simulating the distribution of a pre-determined number of samples across seasons in several ways. Third, we determined whether the magnitude of differences in benthic response between reference and degraded sites changed among seasons. Each test was applied to four benthic response measures: abundance; biomass; diversity; and proportional abundance of opportunistic taxa. The direction and magnitude of long-term trends were quite homogenous between seasons. No contradictory seasonal interactions were detected for any of the trends. Though advantageous, we found that four season sampling is not necessary for conducting trends analysis because the power of some two season alternatives (e.g. spring–summer) was only slightly less than the power for four seasons. For each of the four response variables we examined, power for trend detection was consistently higher when sampling in four seasons than when sampling in one or two seasons. Of the individual seasons, we found that summer is the best season in which to sample. Summer yielded the greatest power for trend detection, although power differences among seasons were mostly limited to the abundance measure. Summer was also the season when difference in benthic response between reference and degraded sites was greatest. © 1997 Elsevier Science Ltd

Benthic macrofaunal communities have been widely used as indicators of habitat condition in marine and estuarine environmental monitoring programmes (Bilyard, 1987; Wilson and Jeffrey, 1994). Macrobenthic communities are used as biological indicators because: *1.* they live in bottom sediments, where exposure to contaminants and oxygen stress is most frequent; *2.* most benthos are relatively sedentary and reflect the quality of their immediate environment; *3.* many benthic species have relatively long life spans and their responses integrate water and sediment quality changes over time; *4.* they include diverse species with a variety of life history characteristics and tolerances to stress and can usually be classified into different functional response groups; *5.* some species are commercially important or are food for commercially important species; and *6.* they affect fluxes of chemicals between sediments and water columns through bioturbation and suspension feeding activities (Boesch and Rosenberg, 1981; Aller, 1982; Dauer *et al.*, 1982; Hartley, 1982; Hargrave and Theil, 1983; Gray *et al.*, 1988; Warwick *et al.*, 1990; Weston, 1990; Weisberg *et al.*, 1997).

Although benthic macrofauna have many advantages as indicators, accounting for seasonal changes in abundance is a challenge when interpreting trends in benthic community condition. Most temperate species have large recruitment pulses during Spring, a smaller recruitment pulse during Autumn, and reduced numerical abundance during Summer and Winter (Buchanan and Warwick, 1974; McCall, 1977; Holland *et al.*, 1987). Several sampling and analytical strategies have been used to avoid confounding of seasonal abundance patterns with interannual trends in benthic macrofaunal monitoring data. Some programmes sample quarterly and stratify data seasonally for analysis (Dauer and Alden, 1995), while other programmes sample monthly and analyse data by modelling annual cycles (Holland *et al.*, 1987). An alternative strategy is to concentrate effort in a single season (Paul *et al.*, 1992). The single season strategy is least expensive and could be advantageous if within-season variability in the selected season is low.

Little is known about which of these sampling/analytical strategies is most cost-effective for minimizing seasonal confounding in interannual trends assessments. In this paper, we use eight years of seasonally collected benthic biological data to assess the relative advantages of multi- vs single-season sampling strategies.

Methods

The effect of seasonality on sampling strategy was examined in three ways. First, we examined whether long-term trends in benthic responses were homogeneous among seasons. Our premise was that homogeneity among seasons is a prerequisite for sampling in a reduced number of seasons since observation of a trend in a single season is of limited value if dissimilar inter-annual patterns occur in other seasons. Second, we used simulations to assess how distribution of samples among seasons affected power for trend detection. Third, we assessed whether the magnitude of difference in benthic response at reference (nonstressed) and degraded (stressed) sites changed with season (stress detection analysis).

Analyses were conducted using benthic macrofaunal data collected quarterly at 16 Chesapeake Bay sites (see Dauer and Alden, 1995 for details of collection regime). The data used in this study were collected from March 1985 through December 1993. The sites represent a range of salinity, sediment silt-clay content, and dissolved oxygen exposure conditions from three different river systems (James, York and Rappahannock Rivers) and the Chesapeake Bay mainstem (Fig. 1). A summary of average salinity and sediment types is given in Table 1 of Dauer (1993). Three replicate benthic samples were collected from each site on each sampling date using a 184 cm^2 box corer with a minimum depth penetration of 25 cm. Samples were sieved through a 0.5 mm screen and preserved in buffered formalin. In the laboratory, samples were sorted, identified to the lowest possible taxonomic level, and enumerated. Biomass was measured as ash-free dry weight for each taxon.

Four benthic response variables were used for all analyses: Shannon-Weiner diversity; total abundance; total biomass; and opportunistic species abundance proportion. Weisberg *et al.* (1997) found that of all measures tested in a range of habitats, these variables were most often indicative of the condition of the benthos. These community measures have also been shown to be sensitive indicators of pollution stress (e.g. Pearson and Rosenberg, 1978; Dauer, 1993). Opportunistic species definitions followed Ranasinghe *et al.* (1993). Seasons were defined on the basis of three month intervals: Spring (April–June); Summer (July–September); Autumn (October–December); and Winter (January–March).

Homogeneity of trends

Homogeneity of trends among seasons was assessed by applying the Van Belle and Hughes (1984) test to each of our 16 sampling sites. The Van Belle and Hughes test is a non-parametric method that allows stratification by season and includes a term to assess whether there is homogeneity of response among seasons; the test has been widely employed for trend analyses of biological and water quality monitoring data in the Chesapeake Bay Program (Dauer and Alden, 1995). The patterns of significance of the seasonal homogeneity tests and the overall trend across all sites were examined. The directions and slopes of significant trends were determined using the Seasonal Kendall slope estimator (Gilbert, 1987).

Power for trend detection

The relative power for trend detection for several seasonal sampling regimes was assessed using Minimum Detectable Impact (MDI) analysis. MDI analysis is a simulation-based power analysis parameterized with empirical data. The MDI level is determined by iteratively introducing incrementally increasing trends into simulated time series until a statistically significant trend is detected at the desired α level. 350 simulations were run for each site-benthic response combination.

The model was parameterized with the correlation and variance structure of the real data, which were assessed from residuals after removing long-term trend and seasonality components. Trend removal from the real data was accomplished based on the Seasonal Kendall slope estimator. Seasonality was determined as a set of predicted values and a set of residuals which represent the unexplained variance associated with each of the seasons in the data set (termed 'within-season errors') using regression analysis, with dummy variables for each of the four seasons. The distributions of within-sampling event errors were evaluated by the maximum likelihood fitting process described by Box and Cox (1964) for the estimation of normalizing power transformations; the optimum power transformation coefficient (designated λ) required to normalize the error distribution for each sampling event was determined and retained for the simulation process. The temporal (between season) correlation matrix was also calculated for each residual data set.

For the simulation process, the characteristics of the data sets were reproduced in a five-step process by re-tracing the steps of the decomposition in reverse order. First a data set of random normal deviates was produced with a random number generator. Second, the temporal correlation structure was introduced using Capra and Elster (1971) techniques. Third, the simulated data were 'back-transformed' to the distribution of the original error matrix using the λ values determined for the Box and Cox power transformation. Fourth, each error term (simulated season residual) in the matrix was multiplied by the standard deviation of the corresponding term in the original residual data set in order to scale its distribution. Finally, the stochastic components of the simulation (the simulated residuals)

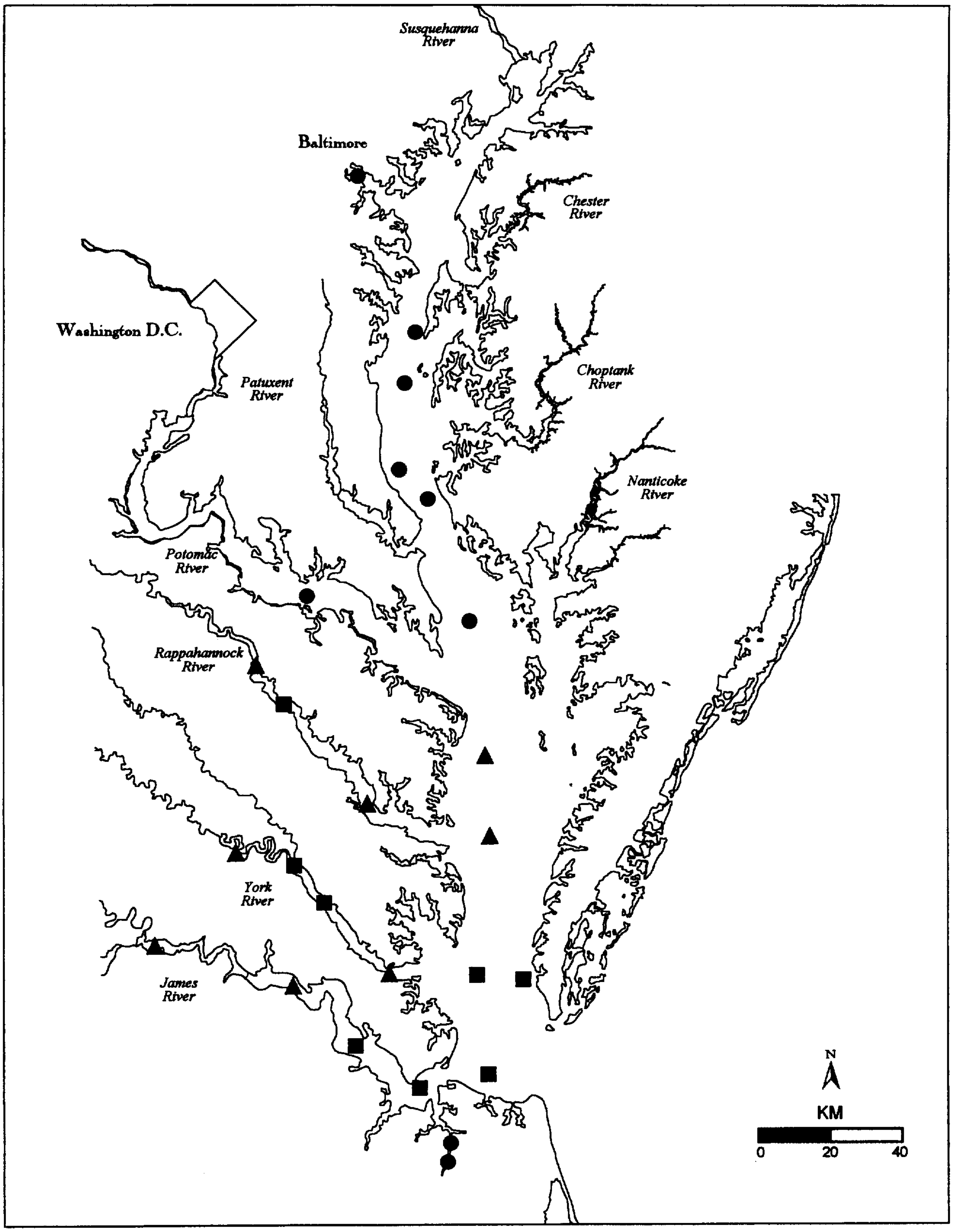

Fig. 1 Sample sites used in analysis. Triangles represent sites used for trend analysis; circles represent sites used for stress detection analysis; squares represent sites used for both.

were iteratively added to the predicted values from the nonlinear regression model to produce the final simulated data set used for power analysis. Details of the estimation equations for power curves are provided in the Appendix.

We conducted MDI power analyses to determine the trend detection sensitivity of several seasonal sampling regimes: allocating twelve samples equally to all four seasons (3 samples each); to every combination of two seasons (6 samples each); and to each individual season (12 samples each). The relative sensitivity of each sampling regime was determined by comparing relative detectable trend slopes (expressed as a percentage of the grand median for the benthic response variable at each site) for $\alpha=0.01$ and $\beta=0.2$ using Friedman's test, with the 16 sampling sites serving as replicates. Comparisons were first conducted among each of the single season regimes and among each of the two season combinations. The most sensitive single season and two-season combinations were then compared to the four season regime as differences in slope expressed as a percentage of the four-season sampling regime slope (i.e. % loss in sensitivity).

Stress detection analysis

Stress detection analysis was conducted by classifying a subset of our sampling sites as reference (minimally affected) sites or as degraded (anthropogenically stressed) sites and comparing the benthic response between them in each season. Classification followed the approach of Weisberg *et al.* (1997) in which sites were treated as 'minimally affected' if no contaminant exceeded the ER-M concentration benchmark established by Long *et al.* (1995) and bottom dissolved oxygen concentration measurements (DO) were consistently high. Consistently high was defined as DO never recorded at less than 2 ppm, at least 90% of the observations were greater than 3 ppm, and at least 80% of the measurements greater than 5 ppm. Sites were classified as degraded if one or more chemicals exceeded ER-M concentrations or DO was low, where low was defined as below 0.5 ppm on any occasion, or below 2 ppm on at least three different days during the year. Sites that met either set of these criteria are shown in Fig. 1.

For the stress detection analysis, data from the 16 Dauer and Alden (1995) sites were supplemented with data collected from nine Chesapeake Bay monitoring program sites which had similar sampling protocols but had been sampled for a fewer number of years (Fig. 1). Two of the supplemental sites were sampled using the same gear but were only sampled between September 1989 and December 1993 (Dauer, 1993). Seven of the supplemental sites were sampled between July 1984 and June 1989 using a 250 cm^2 petite ponar grab (Holland *et al.*, 1987).

Eleven reference and nine degraded sites were identified based on our classification criteria. These sites were stratified for analysis into three habitats based on salinity and substrate: low mesohaline mud; high mesohaline mud; and polyhaline mud, following Chesapeake Bay Program definitions for these habitats (Ranasinghe *et al.*, 1993). Benthic response was compared between the reference and degraded sites within each habitat using a Wilcoxon rank sum test. No sites with sandy sediments (<40% silt-clay content) were classified as stressed. This observation is consistent with those from previous studies that indicate that most contaminants entering coastal and estuarine bodies of water become particle-bound and are eventually concentrated in fine-grained sediments (Swartz and Lee, 1980) and that most low-dissolved oxygen events occur over fine-grained bottoms (Pihl *et al.*, 1991; Dauer *et al.*, 1992; Diaz and Rosenberg, 1995).

Results

Homogeneity of trends

Significant temporal trends were observed for more than half of the 64 site-benthic response combinations tested, with the greatest number of significant trends observed for the diversity and abundance measures (Table 1). Most of these trends were homogeneous across all seasons, with only 2 sites displaying significant heterogeneity in trend slopes between seasons for diversity and for abundance. For these two sites, the heterogeneity resulted from differences in magnitude of the slopes among seasons; there were no cases in which trends were in contradictory directions among seasons.

TABLE 1

Summary of long-term trend analyses of benthic response parameters at our sixteen sampling sites. The values are the number of sites that displayed significant trends, season-specific interactions for trends and contradictory seasonal interactions.

Variable	No. of sites with significant trends	No. of sites with significant seasonal interactions	No. of interactions with contradictory slopes
Diversity	12	2	0
Abundance	12	2	0
Biomass	7	0	0
% Opportunists	2	0	0

TABLE 2

Average ranks of scores for relative trend slopes ($\alpha = 0.01$ and $\beta = 0.2$) among each of the single season sampling regimes. The p values are based on Friedman's test.

Variable	Winter	Spring	Summer	Autumn	p
Abundance	3.3	2.7	1.3	2.8	<0.001
Biomass	2.4	2.8	2.5	2.3	>0.8
Diversity	2.5	2.6	2.3	2.7	>0.8
% Opportunists	2.9	1.8	2.6	2.8	<0.05

Power for trend detection

Among the single season sampling regimes, seasonal differences in power to detect trends were found for only two of the four response variables: abundance and proportion of opportunistic species abundance (Table 2). For abundance, summer yielded about 25% greater power for trend detection than any of the other seasons (Fig. 2). For opportunist abundance, spring yielded the greatest power, but the difference from other seasons was less than 10%.

Significant differences in power among two-season sampling regimes were detected only for the same two variables (Table 3). Again, abundance was the only variable for which the median difference was more than 10%, and all two season combinations that included summer had more power than those that did not (Fig. 3). For opportunistic abundance, the spring–summer and summer–autumn combinations were most powerful, but only by small amounts.

For all four response variables, the four season sampling regime yielded significantly greater power than the best single season (summer) regime and the best two-season (spring–summer) sampling regime (Fig. 4). The four season regime had at least 25% more power for trend detection than the single-season regime (Table 4). Difference in power for trend detection between four-season and two-season regimes was

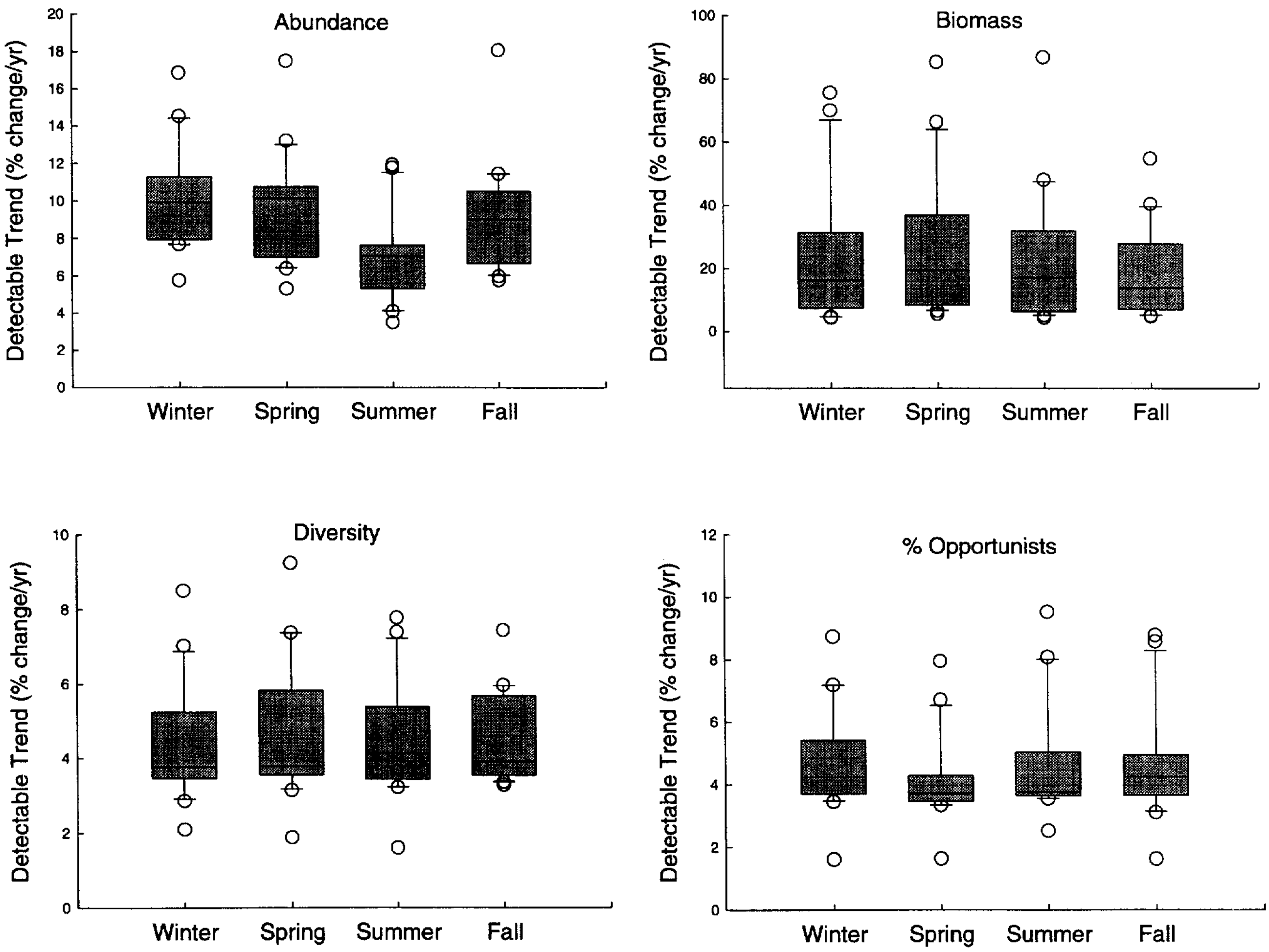

Fig. 2 Box and whisker plots of detectable relative trend slopes (% of grand median for $\alpha = 0.01$ and $\beta = 0.2$) for benthic response variables with single season collection regimes. The horizontal lines are the medians for the sixteen sites, the shaded box represents the 25% and 75% centiles, the vertical lines indicate the 5% and 95% centiles and the circles are the outliers.

TABLE 3

Average rank scores for relative trend slopes ($\alpha=0.01$ and $\beta=0.2$) among each of the two-season sampling regimes. The p values are based on Friedman's test.

Variable	Winter and Spring	Winter and Summer	Winter and Autumn	Spring and Summer	Spring and Autumn	Summer and Autumn	p
Abundance	5.1	2.6	4.9	1.9	4.4	2.0	<0.001
Biomass	4.5	3.4	3.4	3.8	3.4	2.6	>0.1
Diversity	3.5	3.2	3.9	3.5	3.8	3.2	>0.8
% Opportunists	2.9	4.3	4.8	2.5	3.1	2.4	<0.01

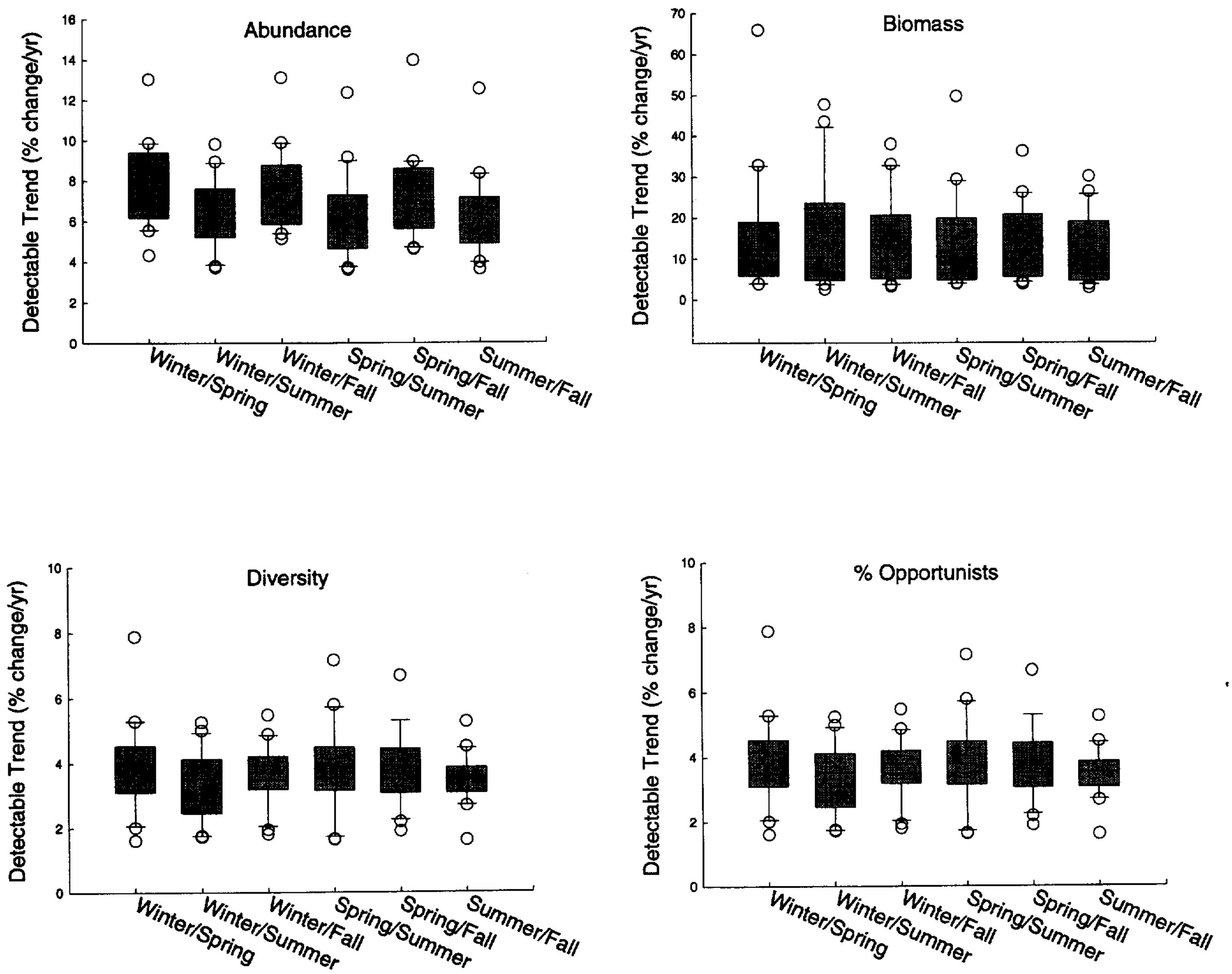

Fig. 3 Box and whisker plots of detectable relative trend slopes (% of grand median for $\alpha=0.01$ and $\beta=0.2$) for benthic response variables with two season collection regimes. The horizontal lines are the medians for the sixteen sites, the shaded box represents the 25% and 75% centiles, the vertical lines indicate the 5% and 95% centiles and the circles are the outliers.

smaller, with a difference of 10% or less for all variables except biomass (Table 4).

Stress detection analysis

In all three habitats, the greatest number of significant differences between reference and degraded sites was observed for Summer (Table 5). Of 12 possible habitat-response combinations, 11 were significant in Summer, 9 in Autumn, 7 in Spring, and only 6 in Winter (Table 5).

The magnitude of difference between stressed and reference sites was generally greatest in Summer and

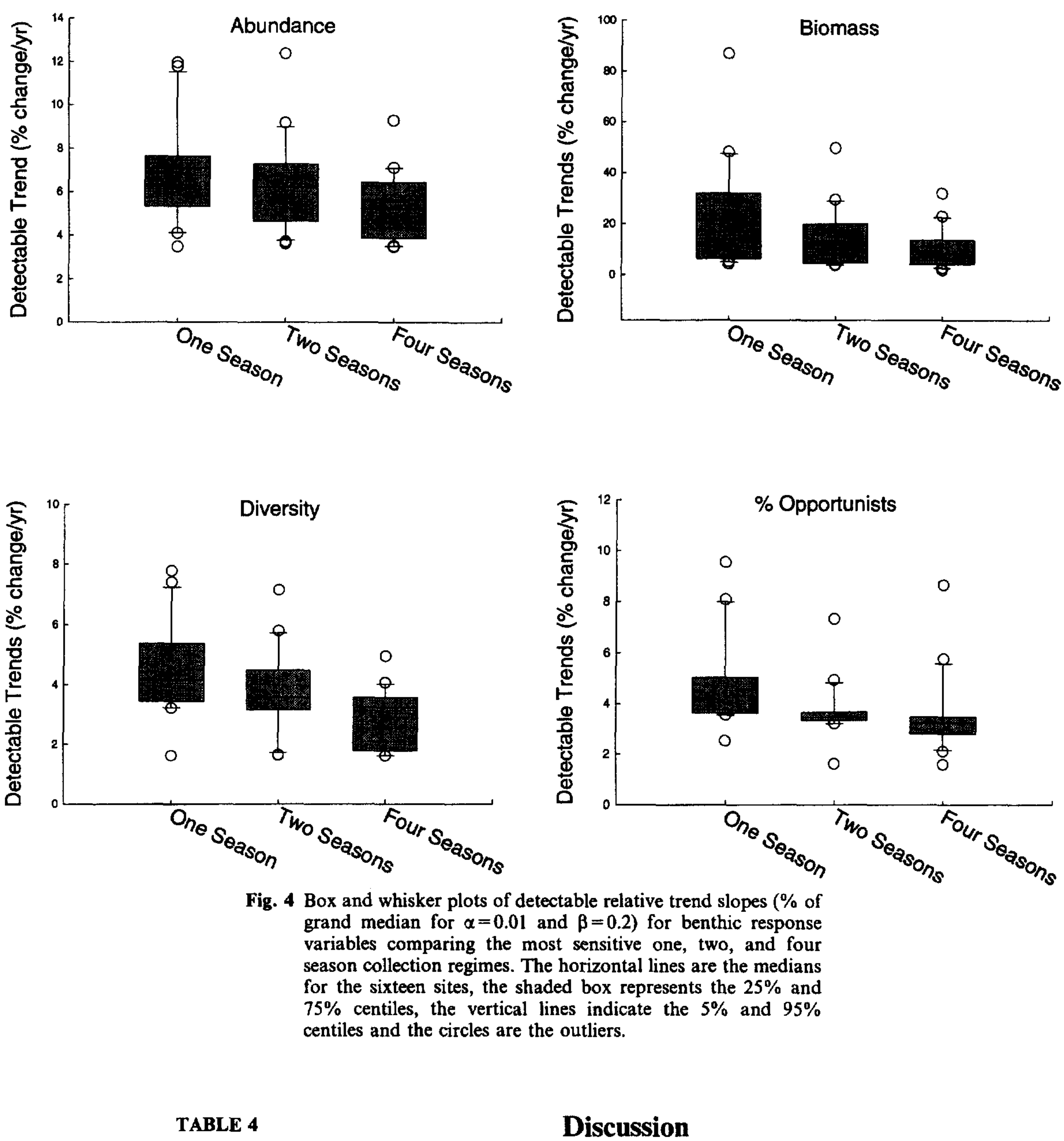

Fig. 4 Box and whisker plots of detectable relative trend slopes (% of grand median for $\alpha=0.01$ and $\beta=0.2$) for benthic response variables comparing the most sensitive one, two, and four season collection regimes. The horizontal lines are the medians for the sixteen sites, the shaded box represents the 25% and 75% centiles, the vertical lines indicate the 5% and 95% centiles and the circles are the outliers.

TABLE 4

Median percent difference in power for trend detection between the two-season (spring–summer) and four-season sampling regimes, and between single-season (summer) and four-season sampling regimes.

	Two-season	Single-season
Abundance	10	27
Biomass	44	84
Diversity	9	48
Percent of abundance as opportunistic taxa	6	9

least in Winter (Table 5). The direction of stressed-reference site differences was consistent throughout the year for all variables except abundance. Abundance was lower at reference sites than degraded sites in spring, but higher in other seasons. The high percentage of opportunist taxa observed at degraded sites in spring suggest the recruitment of opportunistic taxa that fail to survive beyond initial settlement.

Discussion

Benthic macrofaunal communities are monitored for various purposes, the most frequent of which is status and trend assessment. Other objectives may include site-specific impact assessment, identification of causes for regional degradation, quantification of process rates for ecosystem model parameterization, or quantification of prey availability to higher trophic levels. Our results suggest that the choice of seasonal sampling strategy may depend most on the importance of these other objectives, since the power for trend detection is affected to only a small degree by the seasonal sampling strategy.

For trends assessments, our analyses suggest that sampling in all four seasons provides the greatest degree of sensitivity. For each of the four response variables we examined, power for trend detection was consistently higher for four-season sampling than for one or two season regimes. Sampling in multiple seasons may also

TABLE 5
Mean benthic community response at reference (R) and stressed (S) sites in three habitats. Asterisks indicate significant differences in response.

Attribute	Site type	Spring	Summer	Autumn	Winter
Low Mesohaline Mud Habitat					
Shannon-Wiener Index	S	1.81	2.00	2.56	1.63
	R	2.24	2.64*	2.87	2.22
Abundance (no. m^{-2})	S	8904	2505	1333	5352
	R	2679*	1876	1759	3269
Biomass (g m^{-2})	S	4.3	1.1	0.2	1.2
	R	9.5	11.1*	15.4*	18.8
Opportunist abundance (%)	S	61.8	86.6	83.1	28.8
	R	63.2	64.1*	60.1*	54.6
High Mesohaline Mud Habitat					
Shannon-Wiener Index	S	1.96	1.31	2.12	2.27
	R	2.40*	2.73*	2.27	2.46
Abundance (no. m^{-2})	S	3577	772	918	2615
	R	2680	1141*	2395*	4109*
Biomass (g m^{-2})	S	1.8	0.2	0.5	1.0
	R	27.7*	12.2*	11.2*	37.8*
Opportunist abundance (%)	S	82.2	78.1	81.2	67.7
	R	43.7*	58.0*	55.6*	55.9
Polyhaline Mud Habitat					
Shannon-Wiener Index	S	1.73	0.94	2.00	2.24
	R	3.88*	3.67*	3.13*	3.08*
Abundance (no. m^{-2})	S	4517	891	829	1346
	R	3861	2648*	2600*	3714*
Biomass (g m^{-2})	S	0.8	0.02	0.1	0.2
	R	13.5*	6.94*	8.1*	13.0*
Opportunist abundance (%)	S	86.5	68.7	78.6	75.3
	R	36.4*	34.5*	48.8*	62.1*

be advantageous for gathering other information such as identifying recruitment patterns or measuring productivity.

Although four season sampling provides the most powerful trend analysis, our analyses showed that it is not necessary for many trends analyses. We found that some two seasons designs (e.g. the spring–summer regime) are only slightly less powerful and trend patterns were consistent across seasons at each site. We never observed contradictory trend-season interactions for any of 64 combinations of four response variables and 16 sampling sites. Therefore, the choice of seasonal sampling regime is largely a tradeoff between power, cost, and the importance of other objectives. The additional power gained by sampling in four seasons was small, except for the biomass measure, and may not be worth the cost of additional collection cruises.

For programmes that are limited to a single season, we found that summer is the best season to sample. Summer provided the greatest power for trend detection, although power differences among seasons were mostly limited to the abundance measure. Differences among reference and degraded sites were also greatest in the summer. Recent efforts to develop biocriteria assumed that summer sampling would provide the greatest opportunity for distinguishing degraded from reference sites (Engle *et al.*, 1994; Weisberg *et al.*, 1997) and our results confirm the validity of this assumption. The differences are probably greatest in summer because it includes the period of greatest hypoxic stress and precedes the fall recruitment period when partial recovery might occur. Summer is also the time of highest metabolic activity for poikilotherms and therefore the effects of sediment contamination would be most severe.

Sampling in two seasons, particularly in spring and summer, provides a compromise between the relative advantages of single season sampling (to maximize spatial coverage for status assessment) and year-round sampling (to maximize temporal coverage for trend assessment). Sampling in two seasons increases the power for trend detection substantially over single season sampling while reducing collection costs to 50% of the four season option. Two season sampling also provides a broader context for interpretation of trends than single season sampling, should that be an important objective. For instance, sampling in spring allows differentiating whether low summer abundance is due to poor spring recruitment or poor survival after recruitment.

It is likely that there would be little improvement in power for quantifying long-term trends in benthic condition by sampling more often than four times a year, the highest frequency explored in this study. In Chesapeake Bay, many seasonal differences are due to changes in species with one well-defined recruitment pulse and, except for a few species, year-to-year fluctuations are small (Holland *et al.*, 1987). Buchanan and Warwick (1974) and McCall (1977) reported similar patterns for the Northumberland coast and Long Island

Sound. In most temperate estuaries, the seasons are *1.* a period of high abundance and low biomass of benthic communities while most single-recruitment-pulse species recruit (Spring), *2.* a period of low abundance when juveniles of many anadromous fish species crop benthic populations and when some areas are also reduced in abundance due to anoxic or hypoxic events (Summer), *3.* a period of recovery when there is another, smaller, spurt of benthic recruitment and immigration, replacing some summer losses (Autumn), and *4.* a relatively stable period of slow growth and low predation (Winter). Because inter-season transitions are relatively sharp, it is likely that increasing intra-season sampling frequency results only in more precise characterization of the season within which sampling is repeated; rather than increasing the power to detect inter-annual trends. In fact, increasing intra-season sampling frequency would tend to increase the probability of autocorrelation (serial correlation) in the data sets, in turn, increasing the potential for Type I or α errors in trend analyses.

Previous authors have addressed the tradeoffs between spatial coverage, precision of point estimates and power for trend detection (Saila *et al.*, 1976; Kingston and Riddle, 1989). We addressed a complementary element and found that a high degree of flexibility is available in the choice and amount of effort needed for adequate temporal coverage. Based on these results and supplementary power analyses, the Chesapeake Bay benthic monitoring program has been modified to increase spatial coverage, while decreasing temporal coverage, with minimal and acceptable loss of power for trend detection at individual sites.

These conclusions will be likely to have wide applicability, probably including areas outside Chesapeake Bay, because the results were consistent among stations in a range of different habitats. No contradictory slopes and few significant seasonal interactions were observed in any habitat (Table 1) and patterns of relative strength of trend among the four variables were similar for all stations regardless of habitat (Tables 2–5). The stations were located in a range of salinity, sediment silt-clay content, and low dissolved oxygen exposure habitats. One way to test the applicability of these results would be to repeat these analyses on similar data from other geographic areas.

We gratefully acknowledge helpful discussions with J. Volstad at the outset of this project. Portions of this work were funded by the Maryland Department of Natural Resources, the Virginia Department of Environmental Quality, and the U.S. Environmental Protection Agency. We thank our contract officers, R. Magnien, R. Hoffman and J. K. Summers for their support.

Aller, R. C. (1982) The effects of macrobenthos on chemical properties of marine sediments and overlying water. In *Animal-Sediment Relations*, eds P. L. McCall and M. J. S. Tevesz, pp. 53–102. Plenum Press, New York.

Bilyard, G. R. (1987) The value of benthic infauna in marine pollution monitoring studies. *Marine Pollution Bulletin* **18**, 581–585.

Boesch, D. F. and Rosenberg, R. (1981) Response to stress in marine benthic communities. In *Stress Effects on Natural Communities*, eds G. W. Barrett and R. Rosenberg, pp. 179–200. John Wiley, London.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statisticians Society Series B* **26**, 211–252.

Buchanan, J. B. and Warwick, R. M. (1974) An estimate of benthic macrofaunal production in the offshore muc of the Northumberland coast. *Journal of the Marine Biological Association U.K.* **54**, 197–222.

Capra, J. R. and Elster, R. S. (1971) Notes on generating multivariate data with desired means, variances, and covariances. *Educational and Psychological Measurement* **31**, 749–752.

Dauer, D. M. (1993) Biological criteria, environmental health and estuarine macrobenthic community structure. *Marine Pollution Bulletin* **26**, 249–257.

Dauer, D. M. and Alden, R. W. (1995) Long-term trends in the macrobenthos and water quality of the lower Chesapeake Bay (1985–1991). *Marine Pollution Bulletin* **30**, 840–850.

Dauer, D. M., Rodi, A. J. Jr. and Ranasinghe, J. A. (1992) Effects if low dissolved oxygen events on the macrobenthos of the lower Chesapeake Bay. *Estuaries* **15**, 384–391.

Dauer, D. M., Tourtellotte, G. H. and Ewing, R. M. (1982) Oyster shells and artificial worm tubes: the role of refuges in structuring benthic infaunal communities. *Internationale Revue der gesamtum Hydrobiologie Hydrogr.* **67**, 661–677.

Diaz, R. J. and Rosenberg, R. (1995) Marine benthic hypoxia: A review of its ecological effects and the behavioral responses of benthic macrofauna. *Oceanography and Marine Biology Annual Review* **33**, 245–303.

Engle, V. D., Summers, J. K. and Gaston, G. R. (1994) A benthic index of environmental condition of Gulf of Mexico estuaries. *Estuaries* **17**, 372–384.

Gilbert, R. O. (1987) *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold Co., New York.

Gray, J. S., Ascan, M., Carr, M. R., Clarke, K. R., Green, R. H., Pearson, T. H., Rosenberg, R. and Warwick, R. M. (1988) Analysis of community attributes of the benthic macrofauna of Frierfjord/Langesundfjord and in a mesocosm experiment. *Marine Ecology Progress Series* **46**, 151–165.

Hargrave, B. T. and Theil, H. (1983) Assessment of pollution-induced changes in benthic community structure. *Marine Pollution Bulletin* **14**, 41–46.

Hartley, J. P. (1982) Methods for monitoring offshore macrobenthos. *Marine Pollution Bulletin* **13**, 150–154.

Holland, A. F., Shaughnessey, A. and Heigel, M. H. (1987) Long-term variation in mesohaline Chesapeake Bay benthos: spatial and temporal patterns. *Estuaries* **10**, 227–245.

Kingston, P. F. and Riddle, M. J. (1989) Cost-effectiveness of benthic faunal monitoring. *Marine Pollution Bulletin* **20**, 490–496.

Long, E. R., McDonald, D. D., Smith, S. L. and Calder, F. D. (1995) Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. *Environmental Management* **19**, 81–95.

McCall, P. L. (1977) Community patterns and adaptive strategies of the infaunal benthos of Long Island Sound. *Journal of Marine Research* **35**(2), 221–266.

Paul, J. F., Scott, K. J., Holland, A. F., Weisberg, S. B., Summers, J. K. and Robertson, A. (1992) The estuarine component of the U.S. EPA's Environmental Monitoring and Assessment Program. *Chemistry and Ecology* **7**, 93–116.

Pearson, T. H. and Rosenberg, R. (1978) Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology Annual Review* **16**, 229–311.

Pihl, L., Baden, S. P. and Diaz, R. J. (1991) Effects of periodic hypoxia on distribution of demersal fish and crustaceans. *Marine Biology* **108**, 349–360.

Ranasinghe, J. A., Weisberg, S. B., Frithsen, J. B., Dauer, D. M., Schaffner, L. C. and Diaz, R. J. (1993) Chesapeake Bay Benthic Community Restoration Goals. Prepared for the US EPA Chesapeake Bay Program Office, Annapolis, MD.

Saila, S. B., Pikanowski, R. A. and Vaughan, D. S. (1976) Optimum allocation strategies for sampling benthos in the New York Bight. *Estuarine and Coastal Marine Science* **4**, 119–128.

Swartz, R. C. and Lee H., II (1980) Biological processes affecting the distribution of pollutants in marine sediments. Part I. Accumulation, trophic transfer, biodegradation and migration. In *Contami-*

nants and Sediments, ed. R. A. Baker, Vol. 2 Analysis, Chemistry and Biology, pp. 533–554. Ann Arbor Science, Ann Arbor, MI.

Van Belle, G. and Hughes, J. P. (1984) Nonparametric tests for trend in water quality. *Water Resources Research* **20**(1), 127–136.

Warwick, R. H., Platt, H. M., Clarke, K. R., Agard, J. and Gobin, J. (1990) Analysis of macrobenthic and meiobenthic community structure in relation to pollution and disturbance in Hamilton Harbor, Bermuda. *Journal of Experimental Marine Biology and Ecology* **138**, 119–142.

Weisberg, S. B., Ranasinghe, J. A., Dauer, D. M., Schaffner, L. C., Diaz, R. J. and Frithsen, J. B. (1997) An estuarine Benthic Index of Biotic Integrity (B-IBI) for Chesapeake Bay. *Estuaries* (in press).

Weston, D. P. (1990) Quantitative examination of macrobenthic community changes along an organic enrichment gradient. *Marine Ecology Progress Series* **61**, 233–244.

Wilson, J. G. and Jeffrey, D. W. (1994) Benthic biological pollution indices in estuaries. In *Biomonitoring of Coastal Waters and Estuaries*, ed. J. M. Kramer, pp. 311–327. CRC Press, Boca Raton, FL.

Appendix: Minimum Detectable Impact Power Estimates

The MDI method estimates the power of the tests for the data being evaluated at various levels of trend designated by $\Theta_0, \Theta_1, \Theta_2, \ldots \Theta_K$. For the distribution F_Θ, the hypotheses for the test statistics are H_0: $\Theta = \Theta_0$ vs H_A: $\Theta > \Theta_0$. Let C be the rejection region of the test. Let P_i (unknown probability) be the power at Θ_i for i $=0,1, \ldots$ K and we want to find p_i, the estimates of these values which are based upon simulations.

Let N equal the number of simulated runs for any given set of conditions (N $=350$ for the present study, see below) and N_i for i $=0,1, \ldots$ K are the number of runs for which H_0 was rejected at the corresponding trend level (Θ_i). Let $x^{(0)}$ be a simulation of a data set from $F\Theta_0$ (e.g. detrended six year time series) and let $\delta(x^{(0)})$ be its test statistic (e.g. the Z statistic from the Seasonal Kendall test).

If $\delta(x^{(0)})\varepsilon C$, then n_0 is incremented by 1 and another data set is simulated for testing (as long as $\Sigma n_i < N$).

If $\delta(x^{(0)})C$ then $x^{(0)}$ is shifted to $x^{(1)}$ by introducing the next increment of trend (Θ_1) and retesting. The test statistic is re-evaluated and if $\delta(x^{(1)})\varepsilon C$ then n_1 is incremented by 1 and the next simulation started. If $\delta(x^{(1)})C$ then $x^{(1)}$ is shifted to $x^{(i)}$ iteratively by adding increments of trend and retesting until $\delta(x^{(i)})\varepsilon C$ (and n_i is incremented by 1) or until $i = K$ (n_k is incremented by 1) which is a maximum value established to ensure that a test and data set combination with little or no power for trend detection does not run indefinitely or that the introduced trends do not increase to unrealistic magnitudes. The increment of trend in the present study was 1% of the grand median and the K value was 1000, so the maximum annual trend was 1000%.

The power for each level of trend Θ_i i $=0,1,2 \ldots$ K-1 is determined as:

$$\hat{p}_i = (\sum_{j=1}^{i} n_j)/N$$

and confidence limits for probability curves are estimated as:

$$\hat{p} \pm Z_{\alpha/2}\sqrt{pq/N},$$

where p = probability of rejecting the null hypothesis (1-β), q = probability of not rejecting the null hypothesis (β), N = the number of simulation runs for the analysis, and Z is the value from a normal deviate distribution associated with a probability of $\alpha/2$.