

Scoring and aggregation choices alter coastal ecosystem health classifications: implications for coastal management

Enze Zhang

Beijing University of Technology

Hao Li

leslieleehao@163.com

Guangxi Academy of Sciences

Research Article

Keywords: coastal management, ecosystem health, aggregation, regulatory screening, classification uncertainty, Beibu Gulf

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11041793/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

Coastal ecosystem health assessments inform decisions about which sites require further investigation. We evaluated whether ecological classifications and screening decisions remain consistent when monitoring indicators are scored against survey distributions or environmental quality standards and their scores are combined using different methods.

Methods

We analysed water-quality and sediment-quality indicators and measurements of phytoplankton, zooplankton and macrobenthos at 51 stations in three Beibu Gulf surveys. Relative scores were calculated within individual surveys or across all surveys. Eight methods combined scores, including averaging and taking the lowest component score. Shuffling component scores among stations within each survey tested whether unrelated components could produce classification disagreement. Resampling stations estimated uncertainty in disagreement rates. Adding assumed measurement noise and shifting indicators towards poorer conditions tested classification stability and responses to simulated deterioration.

Results

Under within-survey relative scoring, 50 stations received different grades across methods; 43 were rated moderate or worse by some methods and good or excellent by others. When water-quality and sediment-quality indicators were scored against environmental standards, 15 stations were rated healthy by some of the six applicable methods and sub-healthy or unhealthy by others. In simulations, differences in the average scores assigned by the methods accounted for approximately 66% of score variance.

Conclusion

Reports should explain how scores were derived and combined, identify the lowest-scoring component and individual exceedances of environmental standards, and show whether different methods identify the same stations for further assessment.

1. Introduction

Coastal management requires monitoring results that can be related to explicit environmental objectives and interpreted consistently when deciding where further assessment is needed. Elliott et al. (2025) distinguish indicators and regulatory targets from the evidence required to demonstrate management success. Work on indicator selection for marine spatial planning likewise links the meaning of an indicator to the ecosystem component or service being assessed (von Thenen et al. 2020). A further decision arises after indicators have been selected: how to score and aggregate heterogeneous observations into a health class. This step can change the apparent priority of a station without any change in the monitoring record. Understanding that dependence is necessary when a composite class is used to communicate condition, screen sites for follow-up or compare successive assessments.

Chinese coastal assessments provide a practical setting for examining this problem. The nearshore ecosystem health guideline GB/T 42631 – 2023, referenced by the marine environmental impact assessment guideline HJ 1409–2025, combines water environment, sediment, biological quality, habitat and biological community categories into a weighted index. Its weights are 15, 10, 10, 20 and 45, respectively, with three grades separated at 43 and 80. By contrast, the one-out-all-out approach discussed for the European Water Framework Directive gives a decisive role to the lowest biological quality element (Hering et al. 2010; Borja and Rodríguez 2010). Index-based studies also use averages of assemblage-level scores (Williams et al. 2009; Zhu et al. 2023). These approaches embody different allowances for compensation, meaning the extent to which a high score can offset a low score. They need not serve the same management objective.

Aggregation sensitivity is already well documented. Minimum rules become more restrictive as the number of elements increases (Caroni et al. 2013; Latinopoulos et al. 2021), and a low class can have consequences for restoration planning (Prato et al. 2014). Classification also depends on index uncertainty and proximity to boundaries (Moe et al. 2015). Reviews describe a continuum from additive to mixed and minimum aggregation (Borja et al. 2014; Langhans et al. 2014), while applications demonstrate sensitivity to weighting and aggregation (Dobbie and Dail 2013) and disagreement between expert-assigned categories despite similar site rankings (Bay et al. 2007). Chinese estuarine and marine studies report differences between assessment methods or seasons (Chen et al. 2023; Shen et al. 2025). A provincial draft-standard explanation also questions the consequences of lowest-element classification (Jiangsu Provincial Environmental Monitoring

Centre and Nanjing Normal University 2024). Thus, demonstrating that rules yield different classes is insufficient: the management question is when those differences alter a stated screening outcome and what evidence supports that interpretation.

Three issues motivate the present analysis. First, relative positions within a survey and scores based on regulatory limits answer different questions and may identify different limiting elements. Second, data-driven weighting should be examined together with the aggregation hierarchy. Nominal weights need not equal indicator influence (Saisana et al. 2005; Becker et al. 2017; Lindén et al. 2021), despite the use of entropy and related weights in ecosystem health assessment (Wu et al. 2023; Li et al. 2024). The typicality approach of Zhao et al. (2025), for example, combines assemblages using their locations in a t-SNE embedding; its behaviour may depend on both weighting and the nesting of assemblages within a composite. Third, disagreement needs to be interpreted against benchmarks for weakly related elements and assumed measurement error. Without independent condition labels, neither agreement nor greater sensitivity to a simulated impairment establishes that a rule is ecologically correct.

We examine these issues using 51 stations with concurrent water, sediment, phytoplankton, zooplankton and macrobenthos observations from three Beibu Gulf environmental impact assessment campaigns. The study asks how scoring and aggregation affect classes and illustrative screening outcomes; whether the limiting element changes with scoring basis; how weighting and nesting affect responses to simulated impairment; and what can be inferred when a national guideline is applied to incomplete, partly protocol-mismatched observations. We compare three scoring bases, eight rules and two class schemes, use null models and uncertainty simulations, and separate the terms of a factorial score-variance decomposition. The campaigns are used as worked assessment cases, not as a comparison of bay health. The intended contribution is a transparent interpretation of classification choices for coastal management, with the boundary between screening evidence and demonstrated ecological condition made explicit.

2. Materials and methods

2.1 Data provenance

Water, sediment and biotic data come from three environmental impact assessment campaigns following the national specifications for marine monitoring (GB 17378 – 2007): Qinzhou Bay in April 2022 (39 stations with complete water-quality records, 24 sediment stations and 23 stations with all three biotic summaries), Fangchenggang in September 2022 (22, 17 and 15 stations) and Tieshan Harbour near Beihai in November 2023 (24, 13 and 13 stations) (Fig. 1; Table S2). Each campaign sampled one season of one year, so bay, season, year and laboratory batch are aliased; no between-bay statement about ecological condition is made, and the unit of inference is the station within its campaign. Analytical procedures, detection limits and quality control are summarised in Table S1. The present study uses the 51 stations at which water quality, surface-sediment quality and the three biotic summaries were all recorded on the same visit (Qinzhou Bay 23, Fangchenggang 15, Tieshan Harbour 13); Table S2 gives the station counts of each campaign by data type. The analysis concerns the behaviour and interpretation of assessment rules; station observations supply the element scores and indicator-level compliance checks.

2.2 Elements and indicators

Five elements were formed from the summaries that Chinese assessment reports print (Table S12). The water element contains dissolved inorganic nitrogen and active phosphate (nutrient subgroup), chemical oxygen demand and dissolved oxygen (oxygen subgroup) and water-column Cu, Zn, Cd, As and Hg (metal subgroup); the sediment element contains Cu, Pb, Zn, Cd, Cr, As and Hg (metal subgroup) and organic carbon, sulfide and oil (organic subgroup); the three biotic elements are phytoplankton (cell density and chlorophyll a), zooplankton (density and biomass) and macrobenthos (density and biomass). Water-column Cr, Pb and oil were excluded from relative scoring because their values were substituted at half the reporting limit at 34, 15 and 22 of the 51 stations, so that a percentile score would have measured the substitution rather than the water (Helsel 2006; George et al. 2021); the reporting limits were the same in all three campaigns, and the three indicators re-enter the regulatory basis as class assignments, because these reporting limits fall below the relevant class-I limits. Each element score is the mean of its subgroup means, so that a subgroup of seven correlated metals counts once rather than seven times.

Indicator direction follows guideline convention rather than empirical screening, because no reference stations exist in these campaigns. Contaminant and nutrient indicators are scored lower-is-better and dissolved oxygen higher-is-better. Phytoplankton density, chlorophyll a, zooplankton density and macrobenthos density are scored unimodally as the deviation from the campaign median, because GB/T 42631 – 2023 treats both depletion and bloom or enrichment states as departures from its reference values; biomass is scored higher-is-better. Two alternative direction sets are reported as sensitivity analyses (Supplementary Methods).

2.3 Scoring bases

All indicators except dissolved oxygen were log₁₀ transformed. On the relative basis an indicator was scored by the ratio method of index practice, $s = (x - P5)/(P95 - P5)$ for higher-is-better indicators and $s = (P95 - x)/(P95 - P5)$ for lower-is-better indicators, clipped to [0, 1], where

P5 and P95 are the fifth and 95th percentiles computed within the campaign (main analysis) or over the pooled 51 stations (sensitivity). Unimodal indicators were scored $s = 1 - |x - m|/D95$, where m is the campaign median and $D95$ the 95th percentile of the absolute deviation, computed on the same basis. On the regulatory basis each water indicator was assigned its class under GB 3097 – 1997 (classes I to IV, or worse) and scored 1, 0.75, 0.5, 0.25 or 0, and each sediment indicator its class under GB 18668 – 2002 (classes I to III, or worse) and scored 1, 0.67, 0.33 or 0. Regulatory-based element scores remain averages of mapped indicator classes, not official compliance classifications. The regulatory basis is a two-element configuration of water and sediment, because the biotic thresholds of GB/T 42631 – 2023 are defined for a different sampling protocol from the one the campaigns used and are examined separately (Section 2.7).

2.4 Combination rules

Eight rules represent different allowances for compensation among element scores (Table 1; Langhans et al. 2014). Five form an averaging family: the equal-weight mean, an adapted guideline-weighted mean, the geometric mean, centroid typicality weighting and t-SNE typicality weighting. The adapted guideline rule renormalises the surveyed-category weights of GB/T 42631 – 2023 (water 15, sediment 10, biological community 45), allocating the community weight among its five available indicators. It is a sensitivity rule, not the complete guideline implementation examined in Section 2.7. Centroid weights use max-min normalised closeness to the element centroid in indicator-score space, then normalise across elements within each station.

Table 1

Combination rules applied to the element scores E_1 to E_k of a station, covering aggregation forms discussed by Langhans et al. (2014). W denotes station-specific weights that sum to one.

Rule	Definition	Family	Practice represented
Worst element	$\min(E)$	non-compensatory	minimum operator applied to element scores; analogous operator, not a complete regulatory procedure
Mixed	$0.5 \text{ mean}(E) + 0.5 \min(E)$	partly compensatory	mixed additive-minimum aggregation
Capped	$\text{mean}(E)$, but at most the top of the moderate class if any element is poor or worse	conditional	illustrative conditional restriction on compensation
Geometric	$\exp(\text{mean}(\ln \max(E, 0.001)))$	weakly compensatory	multiplicative aggregation
t-SNE typicality	$(E_{\text{water}} + E_{\text{sediment}} + \sum_a W_a E_a)/3$, W_a from the closeness of the station-assemblage row to the origin of a t-SNE embedding, normalised across the three assemblages	compensatory, data-driven	replica of Zhao et al. (2025)
Guideline-weighted	$(15 E_{\text{water}} + 10 E_{\text{sediment}} + 45 [E_{\text{phyto}}/5 + 2 E_{\text{zoo}}/5 + 2 E_{\text{benthos}}/5]) / 70$	compensatory, fixed	GB/T 42631 – 2023 weights, renormalised to the three surveyed categories
Equal	$\text{mean}(E)$	compensatory	equal-weight composite for comparison
Centroid typicality	$\sum_e W_e E_e$, W_e from the max-min normalised closeness of the station's indicator profile to the element centroid, normalised across elements	compensatory, data-driven	deterministic analogue of typicality weighting
Entropy (Supplement)	$\sum_e w_e E_e$, w_e from the entropy of the element score distribution	compensatory, data-driven	entropy weighting in Chinese ecosystem-health assessment

The t-SNE implementation adapts Zhao et al. (2025). The 153 station-assemblage rows are described by log density and log biomass or chlorophyll a, standardised within assemblage, and embedded in two dimensions (perplexity 10, principal-component initialisation). Closeness to the origin is max-min normalised, then normalised across the three assemblages within each station. Their weighted biotic composite is averaged with water and sediment, giving three top-level elements. Three additional rules use a minimum or cap: the mixed rule averages the mean and minimum; the capped rule limits the integrated grade to moderate when any element is poor or worse; and the worst-element rule takes the minimum. The cap is illustrative, and the minimum acts on averaged element scores; neither implements the full regulatory procedures that motivate it. Entropy weighting is reported separately. The Supplement specifies weights, nesting and sensitivity settings.

2.5 Grading and benchmark

Integrated scores were graded on a common scale for all rules. On the relative basis the benchmark was the 95th percentile of the equal-weight score, which keeps the convention that the best observed condition scores one while removing the circularity of benchmarking each rule to its own distribution; on the regulatory basis the benchmark is one. Five classes were cut at 0.2, 0.4, 0.6 and 0.8 of the benchmark (severe, poor, moderate, good, excellent), as in index practice, and three classes at 0.43 and 0.80 of the benchmark (unhealthy, sub-healthy,

healthy), the proportions of the 43 and 80 cut-points of GB/T 42631 – 2023. The five-class labels are analytical categories; the relative three-class labels also do not constitute a guideline-compliant health assessment. Two benchmark alternatives were run as sensitivity analyses: the theoretical maximum of one, and the per-rule 95th percentile used by Zhao et al. (2025).

For a decision-oriented summary, we additionally counted whether a station crossed an illustrative screening threshold under all, some or none of the applicable rules. The threshold was moderate or worse under the within-campaign five-class scheme (eight rules), and sub-healthy or worse under the regulatory-based three-class scheme (six rules). These are separate screening scenarios, not equivalent legal or ecological thresholds. Counts across rules describe the selected rule set and are not probabilities of impairment. This descriptive re-tabulation uses the existing station classifications without refitting any model.

2.6 Agreement statistics, null models and simulations

Rule dependence was summarised against the equal-weight rule by the share of stations receiving the same class, by linear and quadratic weighted kappa (Cohen 1968) and by the Spearman correlation of scores, and across all rules by Krippendorff's alpha with the ordinal metric (Hayes and Krippendorff 2007) and Kendall's W. The flip share, the share of stations whose class differs between at least two rules, is reported for the full rule set, within the averaging family and between the equal-weight rule and the non-compensatory rules. A decisive change is a range of benchmark-normalised scores across rules larger than one class width, which separates flips produced by proximity to a boundary from divergence of the scores. Proportions carry Wilson or bootstrap intervals. The limiting element of a station is the element with the lowest class, ties shared.

The principal within-campaign flip share was referenced to an independent-element null (indicator scores of each element permuted across stations within campaign, 300 permutations) and a concordant null (every element score replaced by the station mean plus Gaussian noise of standard deviation 0 to 0.25). Sampling uncertainty came from a bootstrap stratified by campaign (2000 resamples). Measurement error was propagated by 400 Monte Carlo draws with lognormal noise of coefficient of variation 0.10 for metals and organics, 0.15 for nutrients and 0.30 for biotic summaries; the share of stations whose class under a rule differed from baseline describes instability under the assumed error scenario. The error coefficients are scenario assumptions, not campaign-specific estimates. The t-SNE weights were refitted in the first 100 draws and held fixed in the remaining 300; the reported overall instability and modal vote shares use all 400 draws. Because no validation set exists, the sensitivity of each rule to impairment was simulated by shifting all indicators of one element towards impairment by 0.5, 1.0 or 1.5 pooled standard deviations at ten random stations (120 replicates per element and shift), recording sensitivity (planted stations whose class worsened) and specificity (unplanted stations whose class was unchanged). A factorial simulation used 80 replicates with measurement perturbation and random selection of ten impaired stations per element and replicate. For each station-element combination, means and variances were computed within each rule-by-impairment-state cell. Equal weighting of the two impairment states and the eight rules partitioned the benchmark-normalised score variance into the rule main effect, impairment-state main effect, rule-by-state interaction and mean within-cell variance. The state main effect describes the average response across rules; it does not require identical responses by every rule. Within-cell variation includes the assumed measurement noise and any effects of sample-dependent reweighting as the impaired station set changes. Component shares were averaged across station-element combinations. This decomposition describes simulated scores, not observed ecological change or the fraction of classification errors. Seed and perplexity dependence of the t-SNE rule, alternative biotic directions and configurations, and a species-level check at the 13 Tieshan Harbour stations were run as sensitivity analyses. Full specifications of all simulations are given in the Supplementary Methods (Online Resource 1). Analyses were run in Python 3.14 with pandas, NumPy, scikit-learn and Matplotlib.

2.7 Conditional application of the national guideline

We also evaluated the scoring arithmetic and order of operations prescribed by GB/T 42631 – 2023, conditional on the available inputs and the applicability of its reference values. For estuaries and bays it forms five category indices with weights summing to one hundred: water environment (15; the share of the assessment area in sea-water class I or II), sediment environment (10; organic carbon and sulfide, no metal indicator), biological quality (10; tissue residues), habitat (20; the annual change of coastal wetland area and of sediment grain size) and biological community (45; six indicators, of which the campaigns hold five, phytoplankton density, zooplankton density and biomass and macrobenthos density and biomass). For the biological community the guideline averages each indicator over the stations of the area, classes that mean against its regional seasonal reference value, assigns 45, 30 or 15 points and averages over the indicators; for the sediment category it classes each station, assigns 10, 5 or 1 and averages over stations and indicators. Because the sixth community indicator, biological quality and habitat were not surveyed, the comprehensive index is reported as an interval that brackets every admissible value of the missing items, which retains the prescribed weights while making the missing-input uncertainty explicit. The water indicator is defined as a share of area; we report the share of stations and, as a proxy for area, the share of grid cells assigned to their nearest station. The area approximation is a sensitivity check, not a validated estimate of the guideline-defined assessment area. Protocol differences prevent treating this exercise as an unqualified health diagnosis. Reference values, class bands, the seasonal convention and the sampling-gear basis of the reference values are given in the Supplementary Methods and Table S13.

3. Results

3.1 Weak correlations among element scores

On the within-campaign relative basis the five element scores showed weak pairwise rank correlations: Spearman coefficients ranged from – 0.32 (water against zooplankton) to 0.12 (sediment against macrobenthos), and on the pooled basis from – 0.15 to 0.36. Classed on the common benchmark, the elements of a station were two or more classes apart at 42 of 51 stations on the within-campaign basis and 41 on the pooled basis (Fig. 2a). On the regulatory basis the water and sediment elements were also uncorrelated (ρ 0.03), with limited variation in sediment scores: the sediment element was in the top class at every station and the water element at 36 of 51. Entropy weights followed the dispersion of the element scores rather than any property of the bays: 0.15, 0.33, 0.21, 0.12 and 0.18 for water, sediment, phytoplankton, zooplankton and macrobenthos on the within-campaign basis, and 0.97 for water against 0.03 for sediment on the regulatory basis, where variation in water scores dominated that in sediment scores (Table S5).

3.2 Dependence of the station class on the combination rule

Under within-campaign relative scoring, 50 of 51 stations received different five-class labels across the eight rules (0.98; Wilson interval 0.90–1.00; stratified bootstrap 0.94–1.00), and 40 differed by at least two classes (Fig. 2b; Table 2). Using the guideline cut-point proportions for three analytical classes, 47 changed class and 12 spanned healthy to unhealthy. Across-rule agreement was limited (Krippendorff's α 0.37 and 0.38 for five and three classes), while rankings were more concordant (Kendall's W 0.81; score correlations with equal weights 0.67–0.95). At 96% of stations, the between-rule score range exceeded one class width. Varying cut-points or the number of classes retained a flip share of 0.96–1.00; benchmarking each rule to its own distribution reduced it to 0.82.

Table 2

Agreement of each rule with the equal-weight rule on the within-campaign relative basis (51 stations, common benchmark 0.63), illustrative screening counts, and class instability under measurement error (share of stations whose five-class label differed from baseline over 400 Monte Carlo draws, with t-SNE weights refitted in the first 100 and fixed in the remaining 300; modal-class vote share, median over stations).

Rule	Same class, five classes	Quadratic kappa, five classes	Quadratic kappa, three classes	Spearman rho of scores	Mean score difference	Stations moderate or worse	Stations sub-healthy or worse	Instability under measurement error	Modal vote share
Equal	1.00	1.00	1.00	1.00	0.000	4	27	0.20	0.82
Guideline-weighted	0.80	0.73	0.69	0.95	+ 0.006	2	25	0.20	0.81
Geometric	0.67	0.70	0.62	0.91	– 0.051	10	36	0.28	0.74
Mixed	0.18	0.37	0.30	0.85	– 0.129	27	47	0.24	0.80
Capped	0.59	0.53	0.80	0.84	– 0.040	25	32	0.18	0.89
Worst element	0.02	0.14	0.21	0.67	– 0.258	43	50	0.31	0.71
Centroid typicality	0.73	0.60	0.65	0.85	– 0.002	1	28	0.27	0.78
t-SNE typicality	0.71	0.70	0.64	0.79	– 0.019	10	30	0.17	0.93

Among the five averaging rules, 34 stations changed class and five changed by at least two classes. Comparing equal weights with the minimum-based and capped rules increased these counts to 50 and 35. Moderate-or-worse screening counts ranged from one for centroid weighting to 43 for the worst-element rule, compared with four for equal weights (Fig. 3a). Three-class and pooled-basis results are provided in Table 2 and Table S3. The high flip share occurred in all campaigns (22/23, 15/15 and 13/13 stations), without implying comparable bay condition.

Regulatory-based scoring produced fewer differences. Across the six applicable rules, 15 stations changed five-class label (0.29; Wilson interval 0.19–0.43): 14 by one class and one by two. Fourteen were in Qinzhou Bay and one in Tieshan Harbour. No rule placed a station below moderate. Under three classes, sub-healthy-or-worse counts were 15 for the worst-element rule, 13 for the mixed rule, three for the geometric mean and one each for equal weights and the cap (Fig. 3c). Indicator-level water checks were more restrictive than these averaged scores: 23 stations failed class II and 33 failed class I on at least one indicator.

3.3 The limiting element depends on the scoring basis

On the within-campaign basis, the lowest element class was water at 16.4 stations and sediment at 16.4, with ties shared (bootstrap intervals 10.8–21.8 and 10.9–22.3). The remaining counts were phytoplankton 8.2, macrobenthos 5.7 and zooplankton 4.4 (Fig. 4; Table S4). Under regulatory-based scoring, water alone limited 15 stations and tied with sediment at 36; sediment remained in the top element class throughout. Nutrients accounted for the water shortfall: dissolved inorganic nitrogen and active phosphate were each in class IV or worse at 17 stations. All water metals were in class I except Hg, which was in class II at nine stations.

The contrast persisted with the element set held to water and sediment. Thirty-six stations changed class across the six rules under relative scoring, compared with 15 under regulatory-based scoring. Equal-weight classes agreed between these bases at only 12 stations (quadratic weighted kappa 0.02). These results concern scoring choices within stations, not differences in condition between campaigns.

3.4 Guideline scores conditional on available inputs and reference values

Using the prescribed reference values and order of operations, the five available biological-community indicators yielded 18.0 points at Qinzhou Bay and 15.0 at Fangchenggang and Tieshan Harbour, within the unhealthy band (Table S13). Fourteen of the 15 campaign-level indicator means were in class III. Macrobenthos density at Qinzhou Bay was the exception: two oyster-culture stations raised its mean into class II, although all 23 stations were in class III when classified individually. Bracketing the missing sixth indicator gave community indices of 17.5–22.5, 15.0–20.0 and 15.0–20.0. The sediment category scored 10 points in each campaign. Station-based water compliance shares were 26.1%, 73.3% and 84.6%; approximate area weighting increased them to 62.5%, 76.5% and 93.6% and changed the Qinzhou Bay water grade.

Bracketing all missing inputs yielded comprehensive-index intervals of 38.5–67.5, 41.0–70.0 and 46.0–75.0. All upper bounds were below the healthy threshold of 80. Qinzhou Bay and Fangchenggang nevertheless spanned the unhealthy/sub-healthy boundary of 43; Tieshan Harbour was sub-healthy throughout its interval. Approximate area weighting shifted Qinzhou Bay to 43.5–72.5. These classifications remain conditional on reference applicability and spatial support. The non-prescribed station-level aggregation is reported separately in Table S10.

Biotic departures from the references were broadly consistent in direction across campaigns, but their cause was unresolved. Two campaigns used phytoplankton bottle samples against net-based references, and zooplankton biomass used a different net type from its reference. Low zooplankton density and macrobenthic summaries were not explained by the documented gear differences. Table S13 and the Supplementary Results retain the indicator-level diagnostics. Using winter references for the late-November Tieshan Harbour campaign did not change the conclusion.

3.5 Null benchmarks and instability under assumed measurement error

The observed flip share on the within-campaign basis (0.98) lay within the interval of the independent-element null (0.97, 95% interval 0.92 to 1.00). In the concordant null the flip share rose from 0.02 with identical element scores to 0.46, 0.78, 0.88, 0.93 and 0.94 at element-score dispersions of 0.05, 0.10, 0.15, 0.20 and 0.25 (Fig. S1); the observed dispersion was 0.20. The observed disagreement was consistent with that generated by weakly related scores on the specified class scale; this comparison does not establish statistical independence or ecological equivalence of the elements. The assumed measurement-error scenario changed the five-class label of 17% (t-SNE) to 31% (worst element) of stations under a single rule (equal weights 20%, guideline-weighted 20%), with median modal-class vote shares of 0.71 to 0.93 (Fig. 2c), and the separate fixed-weight and refitted-weight t-SNE summaries agreed after rounding. The share of stations changing class within the averaging family (0.67) was about three times the instability of the equal-weight rule under measurement error, and the share changing when minimum-based or capped rules were included (0.98) was about five times that value. These descriptive comparisons use different sources of variation and are not signal-to-noise estimates. Re-estimating the scoring constants in each bootstrap resample widened the interval of the flip share only slightly (0.92 to 1.00 against 0.94 to 1.00).

3.6 Responses to simulated impairment

For a one-standard-deviation shift of all indicators in one element at ten stations, the mean fraction of planted stations downgraded was 0.80 for the geometric mean, 0.68 for the mixed rule, 0.61 for the minimum, 0.50 for equal weights, 0.48 for adapted guideline weighting, 0.45 for the cap, 0.43 for centroid weighting and 0.40 for t-SNE weighting (Table 3; Fig. 5a). The geometric mean remained first at shifts of 0.5 and 1.5 standard deviations. Unplanted stations retained their classes under fixed-weight rules; the corresponding fractions were 0.98 for t-SNE and 0.93 for centroid weighting.

Table 3

Planted-impairment simulation on the within-campaign basis: share of planted stations (ten per replicate) whose five-class label worsened, averaged over the five impaired elements and 120 replicates, for shifts of 0.5, 1.0 and 1.5 pooled standard deviations of the log10 indicators; specificity at 1.0 standard deviation as the share of unplanted stations whose class was unchanged and as the share not downgraded; and sensitivity by impaired element at 1.0 standard deviation. The three-element mean is a rule with the nesting of the t-SNE rule and fixed weights. Rules are ordered by sensitivity at 1.0 standard deviation.

Rule	0.5 SD	1.0 SD	1.5 SD	Specificity, class unchanged	Specificity, not downgraded	Water	Sediment	Phytoplankton	Zooplankton	Macrobenthos
Geometric	0.44	0.80	0.92	1.00	1.00	0.75	0.78	0.97	0.86	0.65
Mixed	0.37	0.68	0.83	1.00	1.00	0.64	0.62	0.87	0.83	0.44
Worst element	0.38	0.61	0.74	1.00	1.00	0.65	0.61	0.70	0.72	0.38
Equal	0.24	0.50	0.59	1.00	1.00	0.46	0.38	0.65	0.63	0.40
Guideline-weighted	0.25	0.48	0.61	1.00	1.00	0.49	0.26	0.32	0.76	0.59
Capped	0.26	0.45	0.52	1.00	1.00	0.55	0.41	0.56	0.52	0.24
Centroid typicality	0.28	0.43	0.48	0.93	0.93	0.34	0.47	0.42	0.47	0.47
Three-element mean (nesting of the t-SNE rule)	0.24	0.42	0.54	1.00	1.00	0.71	0.65	0.29	0.25	0.22
t-SNE typicality	0.25	0.40	0.48	0.98	0.99	0.66	0.60	0.27	0.27	0.19

Responses depended on the altered element (Fig. 5b). The t-SNE rule downgraded 19–27% of stations with a biotic shift, versus 60–66% with a water or sediment shift. A fixed-weight rule with the same three-element nesting gave similar ranges (22–29% and 65–71%), implicating the hierarchy as well as the weights. Across 30 embeddings, approximately 10% of stations changed class, primarily with perplexity (Table S7). Single-indicator shifts produced smaller responses, with the geometric mean again first (Table S9). Alternative directions and configurations retained high class-change shares (Table S6); collapsing assemblages increased agreement. The species-level check at 13 stations did not validate the summary elements (Table S8). These are responses to prescribed perturbations, not field-validated detection rates.

3.7 Components of simulated score variance

In the main factorial scenario, the mean variance shares of benchmark-normalised scores were 66.0% for the rule main effect, 13.5% for within-cell variability, 15.6% for the impairment-state main effect and 4.8% for the rule-by-state interaction (Fig. 5c; percentages may not sum to 100 because of rounding). The rule main effect represents average offsets across both impairment states. At the noise-free baseline, between-rule score variance was positively associated with the dispersion of the five element scores (Spearman 0.80, R^2 0.60). This supports an association between element disagreement and rule sensitivity, but does not identify the entire rule component as baseline inconsistency. The impairment-state main effect was the average response across rules, not a separate observation of ecological degradation.

Rule main-effect shares ranged from 64% to 67% across impaired elements. The impairment-state component was largest for zooplankton and phytoplankton (19% and 17%), and within-cell variability was largest for macrobenthos (17%). The worst-element and geometric rules together accounted for 52% of the interaction component, compared with 13% each for t-SNE and centroid weighting and 22% for the remaining four rules (Table S11). With five-class labels treated as ordered numerical codes, the corresponding partition was approximately 60%, 20%, 14% and 6%; this ordinal-code calculation is a sensitivity check. Across the smaller and larger impairment shifts and the halved and doubled error coefficients, the rule main effect remained the largest component and the interaction the smallest (Table S11). Their magnitudes are conditional on the simulated shift, noise model, scoring constants and selected rules.

3.8 Agreement at illustrative screening thresholds

Changing a class did not always change the binary screening outcome. On the within-campaign relative basis, 50 stations changed five-class label, but 43 crossed the threshold of moderate or worse depending on the rule. Eight stations remained above that threshold under all eight rules, and no station was below it under all eight (Table 4). Thus, every station flagged by at least one rule had a rule-dependent screening outcome in this scenario. On the regulatory basis, 15 stations crossed the threshold of sub-healthy or worse across the six applicable rules, 36 were unflagged by all six, and none was flagged by all six. Station-level counts and limiting elements are provided in Table S14 (Online

Resource 1) and the companion CSV (Online Resource 2). The two scenarios use different scoring bases, element sets and screening thresholds; their counts describe within-scenario agreement and should not be interpreted as equivalent measures of ecological impairment.

Table 4

Agreement at illustrative screening thresholds among the 51 stations. Mixed agreement means that at least one rule flags the station and at least one does not. The scenarios have different scoring bases, element sets and thresholds and are not equivalent regulatory or ecological assessments. Counts across rules are descriptive, not probabilities of impairment.

Scoring scenario	Screening threshold	Rules	All flag	Mixed agreement	None flag
Within-campaign relative, five elements	Moderate or worse, five classes	8	0	43	8
Regulatory-based, water and sediment	Sub-healthy or worse, three classes	6	0	15	36

4. Discussion

4.1 Classification disagreement and its management meaning

Aggregation choices affected both the reported class and the screening outcome. Fifty of 51 stations changed five-class label, while 43 crossed the moderate-or-worse screening threshold under some rules but not others. Distinguishing these outcomes connects classification sensitivity to its management consequence: a shift from excellent to good need not trigger the same follow-up as a shift from good to moderate. Reporting screening agreement makes this dependence explicit, while the selected rules remain alternative assessment choices rather than independent votes on ecological condition.

The high class-change share was consistent with the independent-element null. Disagreement is therefore expected when weakly correlated elements are combined and converted to discrete classes, extending previous findings for aquatic status and expert sediment assessment (Caroni et al. 2013; Moe et al. 2015; Langhans et al. 2014; Bay et al. 2007). Relatively high ranking concordance alongside limited class agreement further explains why similar station rankings can support different management categories.

In the factorial simulations, average rule offsets accounted for the largest variance share, while the rule-by-impairment interaction was the smallest component. A rule that assigns lower classes overall is thus not necessarily more responsive to an additional impairment. The 17% to 31% within-rule class instability under assumed measurement error adds a complementary measure of robustness. Together, these results distinguish the restrictiveness of a rule, its response to change and the stability of its classifications.

4.2 Relative prioritisation and regulatory screening answer different questions

Scoring basis changed the limiting element even when the element set was held constant. Water and sediment contributed similarly under relative scoring, whereas nutrient-related water shortfalls limited the 15 stations below the top element class under regulatory-based scoring. Relative scores describe position within the sampled distribution; regulatory-based scores map observations against specified limits. The appropriate basis therefore depends on whether the purpose is relative prioritisation or regulatory screening.

Within-element averaging also affected interpretation. Although sediment scores remained in the top class, a sediment indicator exceeded class I at nine stations. Similarly, 23 stations failed a water class-II limit on at least one indicator, compared with 15 flagged by the worst-element rule after averaging. Composite scores should therefore be accompanied by individual exceedances, with the indicator and applicable threshold identified. The nutrient result directs follow-up towards the observed water-quality shortfalls while preserving the separate evidence on sediment and biota.

This distinction connects monitoring indicators to explicit management targets (Elliott et al. 2025). Relative scoring can support the prioritisation of additional monitoring, whereas compliance assessment requires the applicable regulatory procedure. The allowance for compensation among elements should be chosen in advance according to the assessment purpose and the consequences of missed or unnecessary follow-up.

4.3 Weighting, nesting and responses to simulated impairment

Entropy weighting assigned approximately a third of the relative weight to sediment but 97% of the regulatory-based weight to water, illustrating how data-derived weights reflect score dispersion rather than ecological importance (Becker et al. 2017; Lindén et al. 2021). Centroid weighting produced fewer moderate-or-worse stations than equal weights, yet its mean response to planted impairment was only moderately smaller. Typicality weighting therefore showed no consistent leniency across these measures, although sample-dependent reweighting changed some unplanted stations' classes.

Aggregation hierarchy explained much of the contrast between biotic and abiotic responses. The t-SNE implementation downgraded 19% to 27% of stations with simulated biotic impairment, compared with 60% to 66% for water or sediment; a fixed-weight mean with the same

nesting showed a similar pattern. Approximately one station in ten also changed class across embedding settings. These settings merit sensitivity checks because t-SNE geometry has no direct ecological interpretation (van der Maaten and Hinton 2008; Kobak and Linderman 2021). The adaptation of Zhao et al. (2025) examined here tests the aggregation design using different descriptors.

The geometric mean responded most often to planted impairment, whereas the minimum could remain unchanged until the altered element became the lowest. This distinction helps relate compensation choices to the intended response of an assessment (Munda and Nardo 2009; Langhans et al. 2014). The comparison concerns prescribed perturbations and does not establish field detection performance or a universally preferred rule.

4.4 Matching guideline calculations to monitoring inputs

Bracketing missing inputs retained the national guideline's prescribed weights and produced interpretable index intervals. All upper bounds were below the healthy threshold of 80, conditional on the reference assignments and spatial approximation. The practical value is to show which classifications the available inputs support and where further observations are needed, without assigning assumed values to unsurveyed categories.

Reference compatibility is central to that interpretation. Phytoplankton bottle counts from two campaigns and zooplankton biomass obtained with a different net type were compared with references based on other protocols. Other low biotic values were not explained by the documented gear differences. Regional assessments using different metrics and sampling designs provide context (Shen et al. 2025; Zeng et al. 2025), but resolving these departures requires matching observations to appropriate reference methods. Table S13 provides the indicator-level diagnostics.

Spatial support and category coverage also affect the meaning of the final class. Approximate area weighting changed the Qinzhou Bay water grade, demonstrating that station shares and areal proportions are not interchangeable. The guideline's sediment category includes organic carbon and sulfide but no metals. Assessment reports should therefore specify the spatial unit and pressures represented by each category.

4.5 Reporting assessment choices for coastal management

The results support reporting composite classes with four accompanying items: the scoring basis and individual regulatory exceedances; the limiting element and distance from the class boundary; screening agreement across a justified rule set; and uncertainty at the threshold where follow-up would change. This connects environmental targets to monitoring outputs (Elliott et al. 2025) and gives a rule-dependent flag a clear interpretation: the screening outcome depends on an assessment choice that should be explained.

For stations near a screening boundary, checking analytical precision, sample representativeness and scoring constants helps distinguish stable from uncertain classifications. Modal-class vote shares summarise stability under the stated error scenario (Carstensen 2007); they are not calibrated probabilities of ecological health. Additional monitoring can then target the unresolved indicator or protocol mismatch. Recording indicators, references, weights, nesting, cut-points and missing inputs also supports comparison between assessments without confusing a methodological change with environmental change.

The study compares assessment choices within stations and campaigns; the single visits confound bay, season, year and laboratory batch. Relative results depend on the sampled distribution, and bootstrap intervals do not account for spatial dependence. Independent ecological validation and prospective management outcomes were unavailable. The proposed reporting approach therefore provides a basis for testing how transparent classification choices inform subsequent coastal management decisions.

5. Conclusions

Scoring and aggregation choices changed the interpretation of the same coastal monitoring records. Under within-campaign relative scoring, 50 of 51 stations changed class across eight rules, and 43 changed whether they crossed an illustrative screening threshold. Under regulatory-based scoring of water and sediment, 15 stations had a rule-dependent screening outcome across six rules, with nutrient-related water-quality shortfalls providing the limiting element. Neither the composite classes nor agreement among rules constitutes an independent compliance or ecological-health diagnosis.

The simulations showed that average rule offsets accounted for more score variance than the interaction between rule and simulated impairment, while assumed measurement error produced appreciable within-rule instability. The guideline calculation also remained conditional on missing observations, spatial support and sampling-reference compatibility. For coastal management, composite classes should therefore be accompanied by the scoring basis, individual regulatory exceedances, limiting elements, screening agreement and uncertainty. This preserves a clear distinction between relative prioritisation, regulatory screening and health assessment, and identifies the additional evidence needed before a classification is used to justify action.

Declarations

Author Contribution

E.Z. and H.L. conceived the study and acquired funding. E.Z. developed the methodology and analysis software, performed the data analyses, prepared the figures, and wrote the original manuscript. H.L. conducted the investigation, curated the data, supervised the research, and reviewed and edited the manuscript.

Data Availability

The station-level element scores and the analysis scripts are available from the corresponding author on reasonable request. The underlying monitoring data are held by the commissioning agencies and are not publicly archived.

References

1. Bay SM, Berry W, Chapman PM, Fairey R, Gries T, Long E, MacDonald D, Weisberg SB (2007) Evaluating consistency of best professional judgment in the application of a multiple lines of evidence sediment quality triad. *Integr Environ Assess Manag* 3:491–497 https://doi.org/10.1897/ieam_2007-002.1
2. Becker W, Saisana M, Paruolo P, Vandecasteele I (2017) Weights and importance in composite indicators: Closing the gap. *Ecol Indic* 80:12–22 <https://doi.org/10.1016/j.ecolind.2017.03.056>
3. Borja A, Prins TC, Simbhora N, Andersen JH, Berg T, Marques JC, Neto JM, Papadopoulou N, Reker J, Teixeira H, Uusitalo L (2014) Tales from a thousand and one ways to integrate marine ecosystem components when assessing the environmental status. *Front Mar Sci* 1:72 <https://doi.org/10.3389/fmars.2014.00072>
4. Borja A, Rodríguez JG (2010) Problems associated with the 'one-out, all-out' principle, when using multiple ecosystem components in assessing the ecological status of marine waters. *Mar Pollut Bull* 60:1143–1146 <https://doi.org/10.1016/j.marpolbul.2010.06.026>
5. Caroni R, van de Bund W, Clarke RT, Johnson RK (2013) Combination of multiple biological quality elements into waterbody assessment of surface waters. *Hydrobiologia* 704:437–451 <https://doi.org/10.1007/s10750-012-1274-y>
6. Carstensen J (2007) Statistical principles for ecological status classification of Water Framework Directive monitoring data. *Mar Pollut Bull* 55:3–15 <https://doi.org/10.1016/j.marpolbul.2006.08.016>
7. Chen M, Wei L, Ding L, Zheng L, Huang Q (2023) Key issues in the estuarine ecosystem health assessment of the Yangtze River. *Adv Environ Prot* 13:1247–1253 (in Chinese) <https://doi.org/10.12677/aep.2023.136149>
8. Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 70:213–220 <https://doi.org/10.1037/h0026256>
9. Dobbie MJ, Dail D (2013) Robustness and sensitivity of weighting and aggregation in constructing composite indices. *Ecol Indic* 29:270–277 <https://doi.org/10.1016/j.ecolind.2012.12.025>
10. Elliott M, Borja Á, Cormier R (2025) Managing marine resources sustainably – But how do we know when marine management has been successful? *Ocean Coast Manag* 265:107623 <https://doi.org/10.1016/j.ocecoaman.2025.107623>
11. GB 17378-2007 (2007) The specification for marine monitoring, parts 3 to 7. Standardization Administration of China, Beijing (in Chinese).
12. GB 18668-2002 (2002) Marine sediment quality. Standardization Administration of China, Beijing (in Chinese).
13. GB 3097-1997 (1997) Sea water quality standard. State Environmental Protection Administration, Beijing (in Chinese).
14. GB/T 42631-2023 (2023) Guideline for marine ecosystem health assessment in nearshore area. Standardization Administration of China, Beijing (in Chinese).
15. George BJ, Gains-Germain L, Broms K, Black K, Furman M, Hays MD, Thomas KW, Simmons JE (2021) Censoring trace-level environmental data: Statistical analysis considerations to limit bias. *Environ Sci Technol* 55:3786–3795 <https://doi.org/10.1021/acs.est.0c02256>
16. Hayes AF, Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 1:77–89 <https://doi.org/10.1080/19312450709336664>
17. Helsel DR (2006) Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* 65:2434–2439 <https://doi.org/10.1016/j.chemosphere.2006.04.051>
18. Hering D, Borja A, Carstensen J, Carvalho L, Elliott M, Feld CK, Heiskanen A-S, Johnson RK, Moe J, Pont D, Solheim AL, van de Bund W (2010) The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. *Sci Total Environ* 408:4007–4019 <https://doi.org/10.1016/j.scitotenv.2010.05.031>

19. HJ 1409-2025 (2025) Technical guidelines for environmental impact assessment: Marine ecology and environment. Ministry of Ecology and Environment, Beijing (in Chinese).
20. Jiangsu Provincial Environmental Monitoring Centre and Nanjing Normal University (2024) Explanatory note on the draft technical specification for ecological quality assessment of coastal waters. Department of Ecology and Environment of Jiangsu Province, Nanjing (in Chinese) https://sthjt.jiangsu.gov.cn/art/2024/9/26/art_89507_11371164.html Accessed 13 September 2026
21. Kobak D, Linderman GC (2021) Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol* 39:156–157 <https://doi.org/10.1038/s41587-020-00809-z>
22. Langhans SD, Reichert P, Schuwirth N (2014) The method matters: A guide for indicator aggregation in ecological assessments. *Ecol Indic* 45:494–507 <https://doi.org/10.1016/j.ecolind.2014.05.014>
23. Latinopoulos D, Spiliotis M, Ntislidou C, Kagalou I, Bobori D, Tsiaoussi V, Lazaridou M (2021) "One out–all out" principle in the Water Framework Directive 2000—A new approach with fuzzy method on an example of Greek lakes. *Water* 13:1776 <https://doi.org/10.3390/w13131776>
24. Li Y, Huang Y, Zhang X, Jiang Y, Jing X, Deng Y, Li R, Jiang D, Mu J (2024) Ecosystem health assessment: A PSR analysis combining AHP and EW methods for Sansha Bay, China. *Front Mar Sci* 11:1396026 <https://doi.org/10.3389/fmars.2024.1396026>
25. Lindén D, Cinelli M, Spada M, Becker W, Gasser P, Burgherr P (2021) A framework based on statistical analysis and stakeholders' preferences to inform weighting in composite indicators. *Environ Model Softw* 145:105208 <https://doi.org/10.1016/j.envsoft.2021.105208>
26. Moe SJ, Lyche Solheim A, Soszka H, Gołub M, Hutorowicz A, Kolada A, Picińska-Fałtynowicz J, Białokoz W (2015) Integrated assessment of ecological status and misclassification of lakes: The role of uncertainty and index combination rules. *Ecol Indic* 48:605–615 <https://doi.org/10.1016/j.ecolind.2014.08.018>
27. Munda G, Nardo M (2009) Noncompensatory/nonlinear composite indicators for ranking countries: A defensible setting. *Appl Econ* 41:1513–1523 <https://doi.org/10.1080/00036840601019364>
28. Prato S, La Valle P, De Luca E, Lattanzi L, Migliore G, Morgana JG, Munari C, Nicoletti L, Izzo G, Mistri M (2014) The "one-out, all-out" principle entails the risk of imposing unnecessary restoration costs: A study case in two Mediterranean coastal lakes. *Mar Pollut Bull* 80:30–40 <https://doi.org/10.1016/j.marpolbul.2014.01.054>
29. Saisana M, Saltelli A, Tarantola S (2005) Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *J R Stat Soc Ser A Stat Soc* 168:307–323 <https://doi.org/10.1111/j.1467-985X.2005.00350.x>
30. Shen F, Li M, Pang B, Chen Y, Li H (2025) Ecological health assessment of the Beibu Gulf based on a zooplankton index of biotic integrity. *Mar Environ Sci* 44:481–488 (in Chinese with English abstract) <https://www.nmemc.org.cn/hyhjxj/lwlb/2025/2025v4/202509/P020250917518654323754.pdf>
31. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9:2579–2605 <https://jmlr.org/papers/v9/vandermaaten08a.html>
32. von Thenen M, Frederiksen P, Hansen HS, Schiele K (2020) A structured indicator pool to operationalize expert-based ecosystem service assessments for marine spatial planning. *Ocean Coast Manag* 187:105071 <https://doi.org/10.1016/j.ocecoaman.2019.105071>
33. Williams M, Longstaff B, Buchanan C, Llansó R, Dennison W (2009) Development and evaluation of a spatially-explicit index of Chesapeake Bay health. *Mar Pollut Bull* 59:14–25 <https://doi.org/10.1016/j.marpolbul.2008.11.018>
34. Wu W, Hu J, Song D (2023) Assessment of marine ecosystem health and its key influencing factors in Laizhou Bay, China. *Front Mar Sci* 10:1115896 <https://doi.org/10.3389/fmars.2023.1115896>
35. Zeng H, Dong J, Liu W, Li B, Wang X (2025) Macrobenthic community structure and ecological quality assessment in the coastal waters of the northern Beibu Gulf. *Mar Environ Sci* 44:26–34 (in Chinese with English abstract) <https://www.nmemc.org.cn/hyhjxj/lwlb/2025/2025v1/202502/P020250228498424976799.pdf>
36. Zhao Z, Zhu L, Zheng W, Jiang J, Wang Y, Xu L (2025) Development of a multicommunity index of biotic integrity for ecological health assessments of water: A case study of a mountain watershed in eastern China. *Ecol Indic* 176:113737 <https://doi.org/10.1016/j.ecolind.2025.113737>
37. Zhu H, Zhang YZ, Peng YC, Shi BC, Liu T, Dong HB, Wang Y, Ren YC, Xi YL (2023) Assessing the ecological health of the Qingyi River Basin using multi-community indices of biotic integrity. *Ecol Indic* 156:111160 <https://doi.org/10.1016/j.ecolind.2023.111160>

Figures

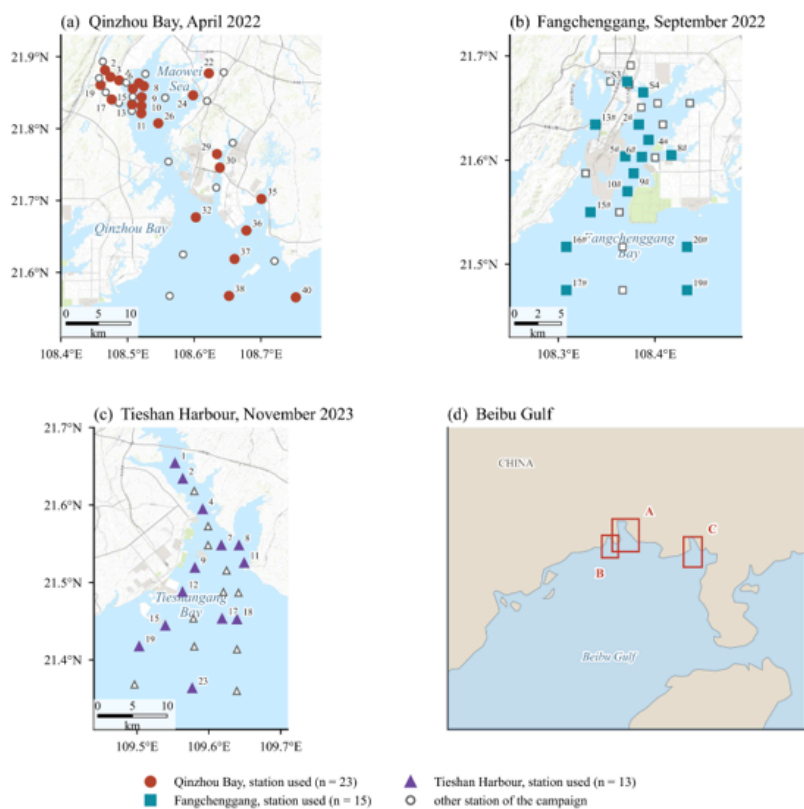


Figure 1

Study area. Stations of the three environmental impact assessment campaigns in (a) Qinzhou Bay, April 2022 (the inner basin is the Maowei Sea), (b) Fangchenggang, September 2022 and (c) Tieshan Harbour, November 2023; filled symbols are the 51 stations with complete water, sediment and biotic records used here, open symbols the other stations of each campaign. (d) Location of the three areas in the Beibu Gulf

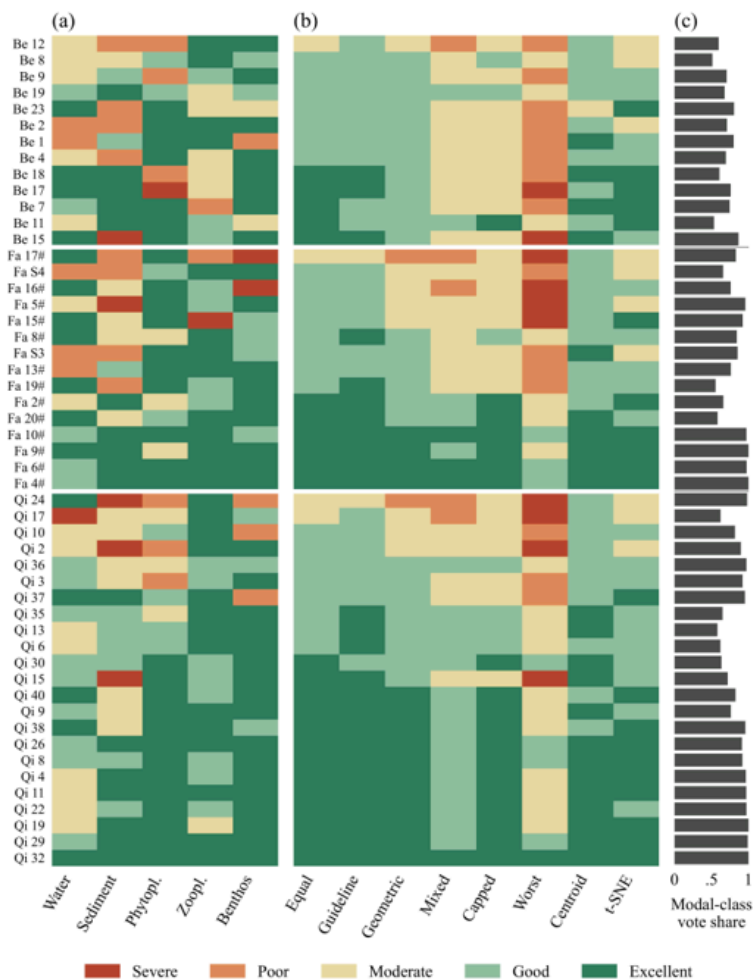


Figure 2 Station classes on the within-campaign relative basis (common benchmark 0.63). (a) Class of each element; (b) integrated class under the eight combination rules; (c) modal-class vote share of the equal-weight rule under measurement error. Rows are stations grouped by campaign (Qi, Qinzhou Bay; Fa, Fangchenggang; Be, Tieshan Harbour) and ordered by the equal-weight score

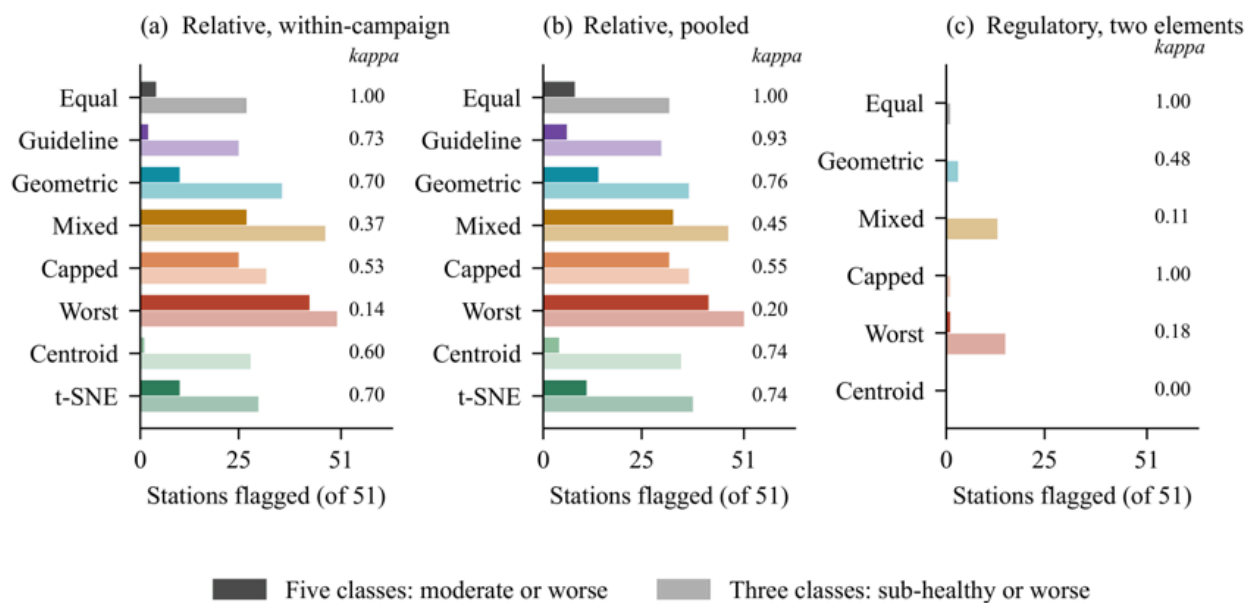


Figure 3 Relative, within-campaign kappa, Relative, pooled kappa, and Regulatory, two elements kappa

Stations flagged by each rule as moderate or worse (five classes, dark bars) and sub-healthy or worse (three classes, light bars) under the three scoring bases, with the quadratic weighted kappa of each rule against the equal-weight rule. The guideline-weighted and t-SNE typicality rules require the biotic elements and are undefined for the two-element regulatory configuration, so (c) shows the six applicable rules in the same order as (a) and (b)

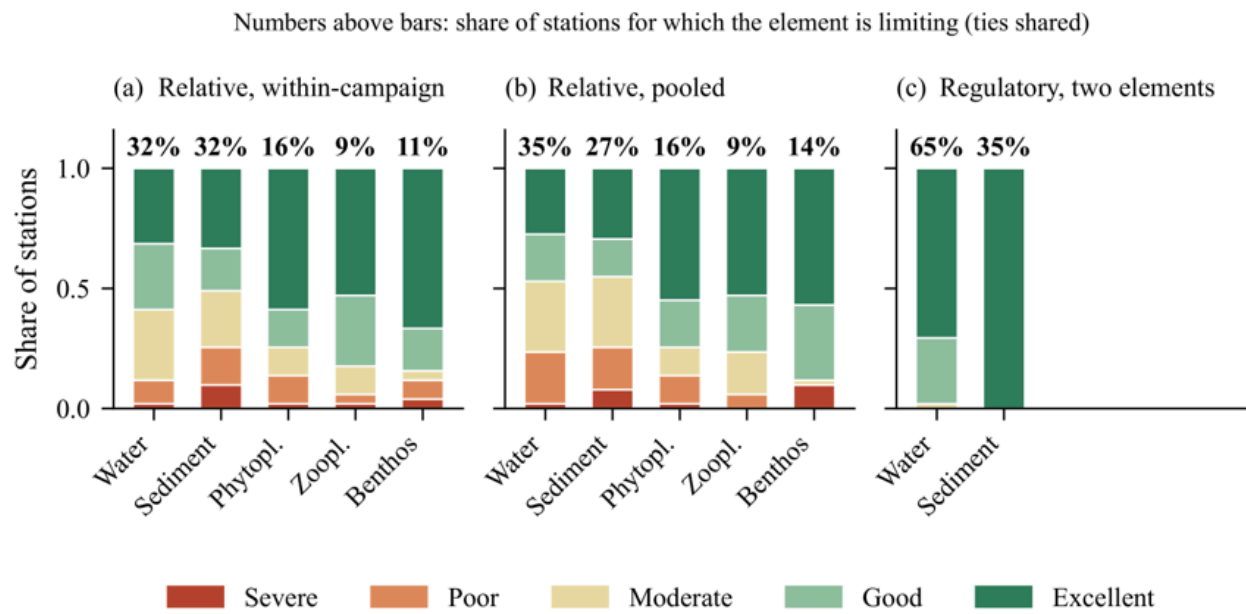


Figure 4

Distribution of element classes and share of stations for which each element is limiting (numbers above the bars, ties shared) under (a) within-campaign relative scoring, (b) pooled relative scoring and (c) regulatory classes, where only the water and sediment elements are available

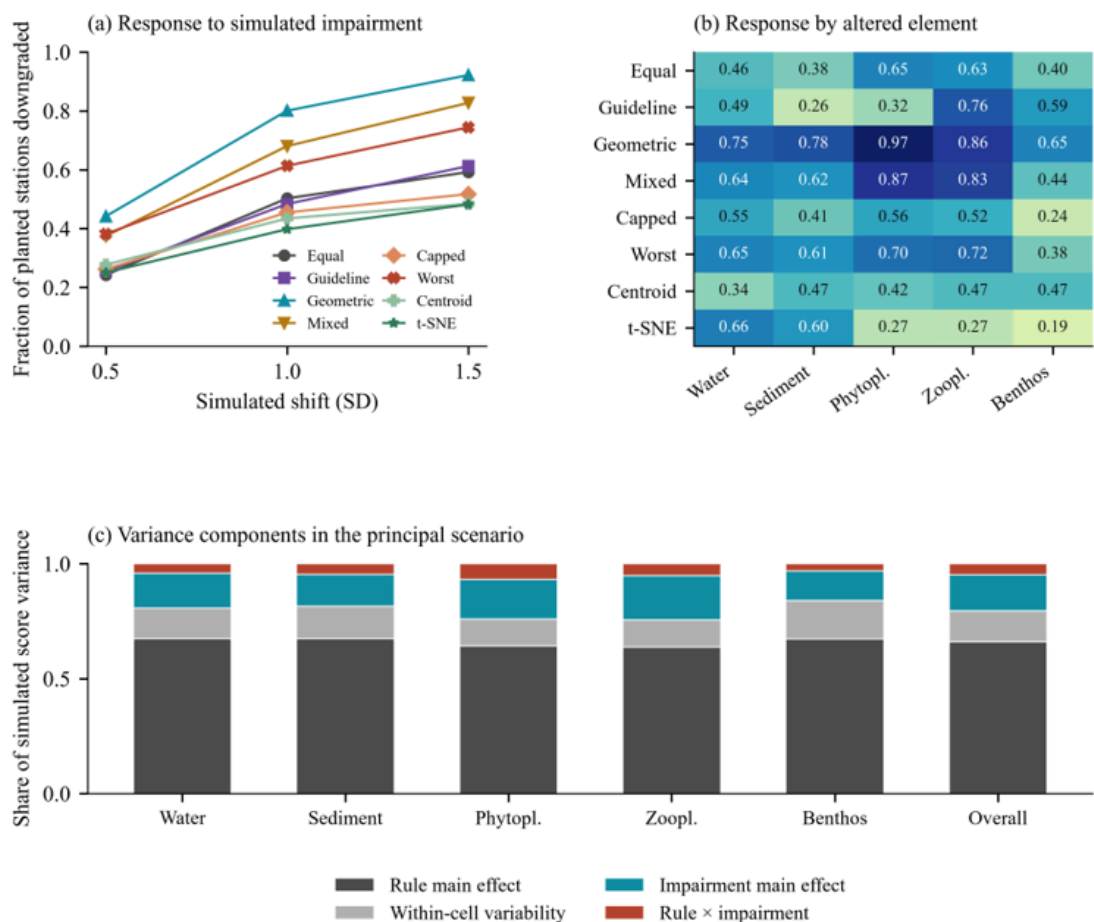


Figure 5

Responses to simulated impairment and components of simulated score variance. (a) Share of planted stations downgraded by each rule as a function of the size of the planted shift, averaged over the five impaired elements; (b) the same share at one standard deviation by impaired element; (c) share of the variance of the benchmark-normalised score at a station, over rules, measurement-error draws and planted states, attributed to the rule main effect, within-cell variability, impairment-state main effect and rule-by-state interaction, by impaired element and overall (shift 1.0 standard deviation)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [02supplementarymaterialJCC20260915.docx](#)