Ecological Indicators 113 (2020) 105834

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Review

Ecosystem health report cards: An overview of frameworks and analytical methodologies

Murray Logan[a,*], Ziyuan Hu[b,c,d], Richard Brinkman[a], Song Sun[c,d,e], Xiaoxia Sun[b,c,d,e], Britta Schaffelke[a]

[a] Australian Institute of Marine Science, PMB No 3, Townsville, Queensland 4810, Australia
[b] Jiaozhou Bay National Marine Ecosystem Research Station, Institute of Oceanology, Chinese Academy of Sciences, 7 Nanhai Road, Qingdao 266071, China
[c] Laboratory for Marine Ecology and Environmental Science, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266071, China
[d] Center for Ocean Mega Science, Chinese Academy of Sciences, Qingdao 266071, China
[e] CAS Key Laboratory of Marine Ecology and Environmental Science, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China

ARTICLE INFO

Keywords:
Indicators
Indices
Metrics
Aggregation

ABSTRACT

Ecosystem health report cards have become increasingly more important tools for communicating environmental state and assessing progress towards management goals. We provide an overview of the major analytical methods underpinning the translation of observed data into robust health indices. In particular, we outline the process of indicator selection, illustrate a variety of index metrics and describe index aggregation with consideration for weighting and the propagation of uncertainty.

1. Introduction

Ecosystems provide a variety of social, economic, cultural and environmental services yet they are also under increasing pressures from development and landuse practices. Management strategies are often implemented in an attempt to prevent further deterioration and to move the ecosystem towards a desired state (e.g. Williams et al., 2009; Newall et al., 2012; Jones et al., 2013; Thompson et al., 2014; GBR, 2014). Effective resource management in complex ecological and political environments requires adaptive decision support tools that are credible, up to date and relevant as well as broadly accessible.

Whilst technical reports and scientific papers are an effective and accepted way of disseminating important information to a specialist audience, the complexity of language and presentation conventions along with limited accessibility does severely restrict their usefulness at reaching other stakeholders beyond the scientific community (such as managers and policy makers) (Schiller et al., 2001). Report cards are becoming increasingly popular tools for assimilating, distilling and disseminating complex scientific knowledge into simpler, succinct, yet holistic assessments of a system in a way that informs a large and diverse non-technical audience (including the general public, managers and policy makers) about the current state and trajectory/progress towards achieving the desired goal(s) (Harwell et al., 1999; Dennison et al., 2007; Conner et al., 2010; Williams et al., 2010). Despite the reduction and simplification of information that is inevitably associated with condensing complex multidimensional data (Saisana et al., 2005), report cards are considered effective management tools (Connolly et al., 2013). Distinguished examples from amongst an ever expanding list include the Chesapeake Bay Report Card (USA; Williams et al., 2009), Pulse of the Bay (San Francisco, USA; SFEI, 2015), the Florida Everglades (Doren et al., 2009), South East Queensland Healthy Waterways (EHMP, 2008), the Fitzroy Basin Report Card (Queensland, Australia; Jones et al., 2013), Reef Rescue Monitoring Program (Great Barrier Reef, Australia; GBR, 2014), Tamar River/Estuary Report Card (Tasmania, Australia; Newall et al., 2012) and the Gladstone Healthy Harbour Partnership Report Card (Queensland, Australia; Gladstone Healthy Harbour Partnership, 2016; McIntosh et al., 2019).

Report cards are communication tools built on top of ecosystem monitoring and management tools which are responsible for integrating and distilling scientific understanding. To encourage and support decision-making, the Organization of Economic Cooperation and Development (OECD, 1993) and the European Environment Agency (EEA, 1995) developed the Driver-Pressure-State-Impact-Response (DPSIR) framework. Broadly speaking, this conceptual framework provides an adaptive management tool for establishing and representing cause and effect linkages (and feedbacks) of environmental problems and in so doing, highlights where the chain can be broken by management actions. Notwithstanding the criticisms discussed by Gari

* Corresponding author.
E-mail address: m.logan@aims.gov.au (M. Logan).

https://doi.org/10.1016/j.ecolind.2019.105834
Received 2 June 2019; Received in revised form 4 October 2019; Accepted 14 October 2019
Available online 19 November 2019

et al. (2015), the DPSIR framework continues to be extensively employed (albeit in a modified state in some cases) and appreciated for its flexibility and communicative power as well as its multidisciplinary and inclusive approach to stakeholder participation. Importantly, in the context of report cards, the DPSIR framework provides a structure within which to identify and communicate the indicators required to address the goals and objectives.

Since ecosystems are complex, dynamic and not directly measurable, different aspects of ecosystem health must be assessed via performance *indicators*. Furthermore, since each indicator is likely to represent a different characteristic of the ecosystem (or part thereof), the specifics of each indicator will vary according to the scientific, economic or management objectives (Harwell et al., 1999). Nevertheless, some form of overall compilation or *index* that assimilates all of these different perspectives is usually desirable. Assembling a coherent and representative set of indicators and indices from amongst the enormous number of potential candidates is a substantial challenge for monitoring and reporting programs (Kannel et al., 2007). Each of these steps must be guided by a considered framework and associated metrics.

Building on the earlier work of Harwell et al. (1999), Kuhnert et al. (2007) outlined a comprehensive ecosystem report card framework that featured four phases (establish goals, formulate conceptual model, establish measurement framework and develop metrics and reporting tools) to translate ecological objectives and outcomes between society (represented by managers and policy makers) and scientists. Specifically, a top-down pass of the phases progressively transfers societal values (goals and objectives set by managers, policy makers and scientists on behalf of society) into sets of scientifically driven indicators (measurable or monitored ecosystem attributes) from which data summaries can be progressively aggregated into easily interpreted and understood scores and grades for managers and policy makers in a bottom-up pass of the phases. As such, this framework stresses the importance of explicitly articulating and linking societal expectations/perceptions and management priorities with scientific data and knowledge.

There are a multitude of sub frameworks and computational procedures that can be used to formulate indicators and generate indicator scores and indices (Sarkar and Abbasi, 2006; Connolly et al., 2013; Raican et al., 2013; Borja et al., 2016). Most published report cards adopt a unique set of computational metrics and analytical routines and while these are typically appropriate, it can be difficult to gain an appreciation of the diversity of methods available. The purpose of the current publication is to provide a comprehensive overview of indicator formulation and combination methodologies that can be used to support report cards.

2. Indicator selection

One of the biggest challenges of report card development is the selection of appropriate indicators from amongst a potentially very large candidate pool. Since the outcomes, conclusions and implications are all dependent on the indicators selected, the selection process is one of the most influential steps and has justifiably received a great deal of attention.

As part of their ecosystem report card framework, Harwell et al. (1999) urged that the alignment of scientific information with societal goals and objectives should be the guiding principle of indicator selection. In their framework, clearly articulated societal goals and objectives (a combination of societal values and scientific knowledge, such as restored and sustainable wetland system) are translated into Essential Ecosystem Characteristics (EECs) that represent a set of generic attributes that further refine the broad goals (such as water quality, sediment quality, habitat quality, ecological processes). The EEC's are then further translated into a set of scientifically informed indicators that are measured to indicate the status of trends or states associated with the EEC's.

Numerous studies have focused on providing more formal, objective criteria for indicator selection (Dauvin et al., 2008; Emerson et al., 2012; Flint et al., 2012; James et al., 2012). Most can be broadly encapsulated by Dauvin et al. (2008)'s contextual implementation of the Doran (1981)'s SMART (Simple, Measurable, Achievable, Realistic, and Time limited) principle. A 'good' indicator should be representative, easily interpreted, broadly comparable, sensitive to change and have a reference or guideline value. To be 'useful', an indicator must be approved by international consensus, be well grounded and documented, have a reasonable cost/benefit ratio and ideally have adequate historical and on-going spatial-temporal coverage. Flint et al. (2012) and James et al. (2012) further developed numerical scoring systems to help evaluate indicators objectively. Nevertheless, Neary (2012) warned against the potential to manipulate an index by saturating it with inappropriate, irrelevant or biased indicators and whilst recommending that an index comprise of at least seven indicators, they did advocate that the type of indicator is more important than the number of indicators.

Since final outcomes are likely to be highly influenced by indicator choice, the robustness and sensitivity of both indicators and final outcomes to changes in ecosystem health should be understood if not formally investigated as part of the indicator selection process (Dobbie and Dail, 2013). This can, for example, involve sensitivity analyses:

- simulating changes in the underlying data of different magnitudes and estimating the resulting sensitivity (percentage or probability of change) expressed by the indicator
- estimating the effect of past perturbations on the indicator hindcasted from historical data

Alternatively, indicators can be weighted or selected on the basis of entropy. Entropy is a measure of system disorder or uncertainty (Shannon, 1948) and in a statistical context can be a measure of the quantity of *information* contained in stochastic data. The concept of entropy draws an important distinction between data (the actual observations) and information (knowledge required to produce the data). And whilst both data and information are measured in units of storage (called bits as they typically pertain to booleans), typically fewer bits are required to store information.

The classic application of entropy measures the minimum information required to represent the coding of sequences of symbols. For example, if we had two sequences (S1 = {A,A,B,B,C,C} and S2 = {A,A,A,A,B,C}) and expressed them as probabilities (P(S1) = {A:1/3, B:1/3, C:1/3} and P(S2) = {A:2/3, B:1/6, C:1/6}), we see that when the probabilities are even (e.g. S1) there are few patterns (information) from which to leverage predictability and thus disorder (entropy is relatively high). By contrast, greater variety in the probabilities (e.g. S2) results in greater ability to leverage patterns and thus lower overall entropy. This concept is captured in Shannon (1948)'s, entropy of data ($H(X)$):

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i)$$

where $P(x_i)$ is the probability of the i^{th} value of x, calculated as $x_i / \sum_i^n = 1 x_i$.

Environmental indicator data can be normalized and standardized (see Section 4.3) so as to be expressed as relative probabilities and onto which the concept of entropy can be applied (e.g. Zou et al., 2006). For m indicators, each with n observations, the entropy of the i^{th} indicator is described by:

$$H_i = -\frac{1}{log_2 n} \sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where $\frac{1}{log_2 n}$ normalizes the entropy to the range of [0,1] and $P(x_i)\log_2 P(x_i) = 0$ when $P(x_i) = 0$.

The entropy values can then be converted into weights (w_i) according to:

$$w_i = \frac{1 - H_i}{m - \sum_{i=1}^{m} H_i}$$

Take for example, three variables (candidate indicators), each with four normalized and standardized values (V1 = {1/2, 1/2, 1/2, 1/2}, V2 = {1/2, 0, 1/2, 0} and V3 = {1, 0, 0, 0}) which translate into probabilities of V1 = {1/4, 1/4, 1/4, 1/4}, V2 = {1/2, 0, 1/2, 0} and V3 = {1, 0, 0, 0}. It is clear that V1 is invariant and V3 the most varying and these values translate into entropies (V1 = 1.0, V2 = 0.5 and V3 = 0.0) and entropy weights of (V1 = 0.0, V2 = 0.067, V3 = 0.133). V3 will be weighted twice as much as V2 and V1 will be ignored.

Entropy weights provides an objective means by which the relative influence of different indicators, sites or any other units can be controlled across aggregations based on historical variability. However, the use of entropy weights assumes that the data used to generate the weights is going to be representative of the system into the future, an assumption that is likely be difficult to justify in systems that are expected to be undergoing change (arguably most systems for which a report card is in use).

3. Hierarchies

Across the literature and published report cards, there are substantial inconsistencies in what are considered 'indicators'. Often this is largely due to contextual differences. As stressed above, indicators should align closely with report card objectives. Yet in the more broad ecosystem report card frameworks, such indicators are often too general to be measurable. For example, indicators of environmental health might be comprised of water quality, aspects of habitat quality and some representation of the state of the biota. In such cases, these 'indicators' might be further sub-divided into progressively more specific and directly measureable properties. Correspondingly, an indicator of water quality might comprise sub-indicators of nutrients, productivity and water clarity which in turn might be represented by more specific measures such as total nitrogen, total phosphorus, chlorophyll-a and secchi depth (see Fig. 1).

Similarly, whilst report cards are typically presented at large spatial and temporal scales (e.g zones/regions and annually), the data are collected at smaller scales (e.g site and monthly) to ensure more thorough representativeness. Examples of indicator, spatial and temporal hierachies are presented in Fig. 1. The resulting report card design could be represented by multiple hierarchical structures in which sub-indicators (etc) are nested within indicators, spatial scales are nested from entire regions, sub-regions or zones down to individual sites or sampling units and time nested from years to months or even days.

One of the strengths of such a hierarchical report card framework is that the inherent inbuilt redundancy allows for the addition, deletion or exchange of lower order items (such as sites and actual measured variables) with minimum disruption to the actual report indicators. That is, the indicator is relatively robust to some degree of internal makeup. Furthermore, by abstracting away the fine details of an indicator, similar indicators from different report cards (each potentially comprising different sampling designs) are more directly comparably. For example, in different report cards that include water quality, a water quality indicator of 'water clarity' might comprise different Measures (e.g. suspended solids, NTU, Secchi depth etc) collected from different sources (e.g. satellite, in situ loggers or hand samples), yet provided each of these water clarity indicators are well calibrated, it should be possible to compare state and trend across the report cards.

4. Ecosystem health indices

Each individual indicator (or sub-indicator) addresses a different aspect of the state of an ecosystem. Hence, even a modest number of (sub) indicators will yield multiple perspectives on ecosystem health. Capturing the essence of the ecosystem health or an indicator thereof, necessitates integrating (aggregating) each of these perspectives together into a single *index*. There are numerous methods that have been applied to index aggregation, the most popular of which are itemized by Fox (2013) and described and evaluated in the context of water quality indices by either Walsh and Wheeler (2012) (from the perspective of cost benefit analyses) or Whittaker et al. (2012).

Multivariate statistical methods

Motivated by the need to integrate multiple disparately scaled ecological variables together in the absence of any normalizing information (such as benchmarks, see Section 4.1), a variety of predominantly multivariate analyses have been used in the generation of ecosystem health indices. However, Whittaker et al. (2012) cautioned that since the incorporated weights are all exclusively informed by the statistical properties of the constituent indicator data, if these statistical properties did not coincide with expert knowledge of the relative importance of the indicators, then the resulting indices are likely to be poor.

As an alternative, Whittaker et al. (2012) suggest the Malmquist index. The computational details of the Marlmquist index are rather complex and since this method does not appear to have been adopted by any report card, we will restrict our description to just a brief overview. Whittaker et al. (2012)'s proposed version of the Malmquist index calculates pairwise ratios of indicator distances from a multivariate benchmark curve. The benchmark curve (a form of indifference curve), is a multivariate curve defined by the lower boundary of a convex hull of all indicator values and is thus derived entirely from the observed data. Using simulated data with manufactured statistical complications (heterogeneity and temporal autocorrelation), Whittaker et al. (2012) demonstrated that the Malmquist index out-performs indices based on principal components analysis and suggested other statistical methods would have similar shortcomings.

4.1. Benchmarks – References, baselines, triggers, thresholds and guidelines

The absolute value of an indicator is rarely a meaningful representations of ecosystem health. Nor are the statistical properties of a time series necessarily a good basis for normalizing indicators or representing the objectives. What constitutes a 'good' or 'poor' level is likely to vary according to indicator, the ecosystem (e.g. freshwater, estuarine or marine), geographical and temporal (e.g. pre-industrial or current, seasonal) context as well as our understanding of the long-term ranges of ecosystem condition (including amplitudes of responses to disturbances). Ecological managers have long recognized the need to express ecosystem ratings as standardized scores and in terms that are accessible to policy makers and the general public. Whilst initial applications focused on normalizing observed measures against subjective rating curves to yield dimensionless index values on the scale of [0,1] that could be readiby combined into a single understandable score or rating (e.g. Miller et al., 1986), more recent studies have explored formulations that compare observed measures to baseline, reference, objectives or guideline values (collectively, benchmarks) (e.g. CCME, 2001; Hurley et al., 2012; Jones et al., 2013).

Benchmarks are typically either reference or baseline conditions (sites or historic data representing relatively low disturbance 'healthy' conditions), threshold, ecotoxicology or guideline values (e.g. derived from historical quantiles or ecotoxicology). Thresholds, triggers and guideline values are typically peer reviewed and ecologically meaningful, they are context specific and cannot necessarily be scaled up or applied elsewhere.

Whilst a 'distance to benchmark' approach does provides some level

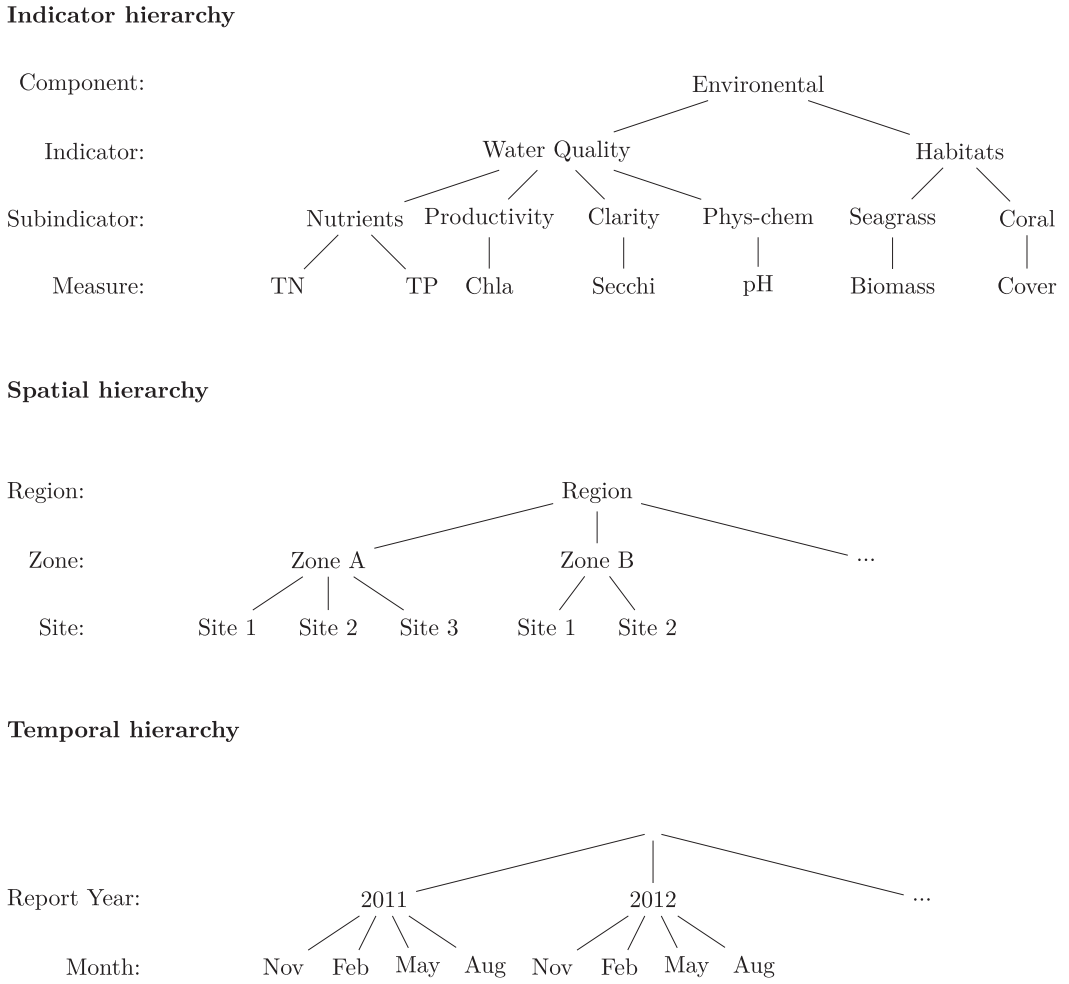


Fig. 1. Example spatial and temporal hierarchies, illustrating the cascade from higher to lower scales and abstraction levels. Note, from the perspectives of aggregation calculations, these hierarchies are upside down.

of standardization (Connolly et al., 2013), to be useful, not only should there be some form of homogenization in what the benchmark condition represents, the polarity of the distance (direction of the response) should be well understood (Hijuelos and Reed, 2013) and the magnitude needs to be meaningful with regard to range of ecosystem condition values, that for example may vary along environmental or disturbance gradients. That is, there should be some consistency in what it means to be above or below a benchmark, and indeed what it means to be a certain distance from a benchmark. Ideally, benchmarks should also be locally relevant (Connolly et al., 2013) and consider seasonal variability (Hallett et al., 2012; Coates et al., 2007). Indeed, in a review of the methodologies used to set benchmarks, Borja et al. (2012) demonstrated the importance of setting appropriate benchmarks from which to assess ecosystem quality by directly linking the inability of indices to detect impacts in ecosystems to inappropriate reference conditions.

It is also important that benchmarks align with objectives in order to ensure indicators are appropriate. For example, if an objective is to maintain sustainable stocks of a particular species of fish, a benchmarks that reflect either historical numbers or the numbers present at low pressure sites do not necessarily represent the level of sustainability.

Connolly et al. (2013) reviewed the use of report cards for monitoring ecosystem health and tabulated the general properties of a range of index-generating methods employed across a many different monitoring programs. Rather than duplicate that information here, the current intention is to provide more specific details about the algorithms used across those programs.

4.2. Unifying indices

The Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI, CCME, 2001) incorporates comparisons to baseline based on *scope* (F1: proportion of indicators that have one or more failures to meet objectives), *frequency* (F2: proportion of all comparisons failing to meet objectives) and *amplitude* (F3: the normalized degree to which failed comparisons exceed objectives).

$$
\begin{aligned}
F_1 &= 100.\left(\frac{\textit{Number of failed indicators}}{\textit{Total number of indicators}}\right)\\
F_2 &= 100.\left(\frac{\textit{Number of failed comparisons}}{\textit{Total number of comparisons}}\right)\\
F_3 &= \frac{100.E}{1+E}\\
E &= \sum_{i=1}^{n} e_i/n\\
e_i &= z_i.\left[\left(\frac{x_i}{benchmark_i}\right)^{\lambda_i} - 1\right]\\
z_i &= \begin{cases} 1 & \text{if ith comparison fails}\\ 0 & \text{otherwise}\end{cases}\\
\lambda_1 &= \begin{cases} 1 & \text{If} < benchmark_i = \text{fail}\\ -1 & \text{If} > benchmark_i = \text{fail}\end{cases}\\
CCMEWQI &= 100 - \left(\frac{\sqrt{F_1^2+F_2^2+F_3^2}}{1.732}\right)
\end{aligned}
$$

where n is the number of comparisons.

Whilst the CCME WQI serves its purpose in the context to which it is

applied, without first aggregating, it is unlikely to be a useful metric for any indices involving remote sensing data or indeed any situation with a reasonable large amount of data or indicators. One-third of the weighting of the metric is calculated on the proportion of indicators that one or more failures. The more observations are collected, the more likely at least one of them will exceed the benchmark resulting in an indicator fail. Hence, this one-third will quickly approach a constant of 1 thereby reducing overall sensitivity. In addition, the one-third of the method that weights on amplitude only does so with respect to failure - there is no degree of how well the data recedes the benchmark. Finally, unifying indices have very limited scope for propagating any uncertainty.

4.3. Hierarchical indices

The CCME WQI unifies all indicators into a single index as part of the calculations. However, most other indices are calculated at the level of the individual indicators which can then be progressively aggregated together to form higher level indices. One of the challenging aspects of combining multiple indicators together is that to do so, all indices must be expressed on a common scale which encapsulates the full spectrum of possible states.

4.3.1. Distribution based indices

One way to achieve standardization is to define numerical boundaries for both condition categories and the associated scoring scale for each indicator (see Fig. 2). Thus for a particular indicator, values are converted into scores via simple linear interpolation. Whilst this methodology allows a simple and flexible conversion between observed data and indicator scores, it does necessitate careful consideration of condition boundaries. Nevertheless, boundaries could be based on historical quantiles or other distributional properties. Similarly, standardized scores can be calculated as simple z-scores from historical distributions.

4.3.2. Comparison to benchmark/thresold/reference

There are numerous ways to formulate indicator scores based on deviations from a benchmark (see Table 1). The Binary method expresses a comparison to benchmark values on a binary compliance scale (1: complies with benchmark, 0: fails to comply) and whilst simple to perform and understand, this method results in indices that have the potential to be either under or overly sensitive (depending on how far observed values typically are from the benchmark). For example, at one extreme (when values are close to benchmark), slight changes yield dramatic fluctuations in scores. However, when values are substantially above or below the benchmark, even modest improvements or deterioration will be undetected. This rapid 'switching' behaviour is depicted by the stepped response curve.

In the State of the Great Lakes Report (EPA/EC, 1995), greater resolution is achieved via a panel of experts who classify each of six health indicators (aquatic community health, human health, habitat, contaminants, nutrients and economy) into four categories: poor, mixed/deteriorating, mixed/improving, good/restored. Similar expert rating or multi-category exceedance grading systems are employed in other report cards (e.g Tamar estuary Report Card; Attard et al. (2012)) and whilst probably reasonably accurate, they are nonetheless highly dependent on the availability of a reasonably stable panel of independent experts over time.

There are numerous examples of indices that attempt to capture both distance and direction of benchmark comparisons so as to distinguish degrees of compliance/non-compliance.

The Benchmark and Worst Case Scenario method (see Table 1) employed by the Fitzroy Basin Report Card (Jones et al., 2013) reflects the degree of failure by scaling the difference between the observed values and benchmarks (20th or 80th percentile of long term data for values above and below the benchmark respectively) to the Worst Case Scenario values (10th or 90th percentiles respectively). The associated response curve demonstrates a linear decline in Score with increasing distance from the benchmark.

The Modified Amplitude method calculates the distance to benchmark on a logarithmic (base 2) scale. The base 2 logarithm represents ratios on a symmetric scale such that value that are twice and half the benchmark yield scores of the same magnitude (yet apposing signs), and has some inbuilt capacity to accommodate skewed data. The Modified Amplitude response curve illustrates how this method can be simultaneously relatively insensitive to slight fluctuations around the benchmark as well as sensitive to changes further away from the benchmark.

The Modified Amplitude method can be scaled by capping and rescaling to a specific range. The formulation employed by the Gladstone Healthy Harbour Partnership (?) caps and scales the Scores to the range of [0,1] (equivalent to double and half the value of the benchmark).

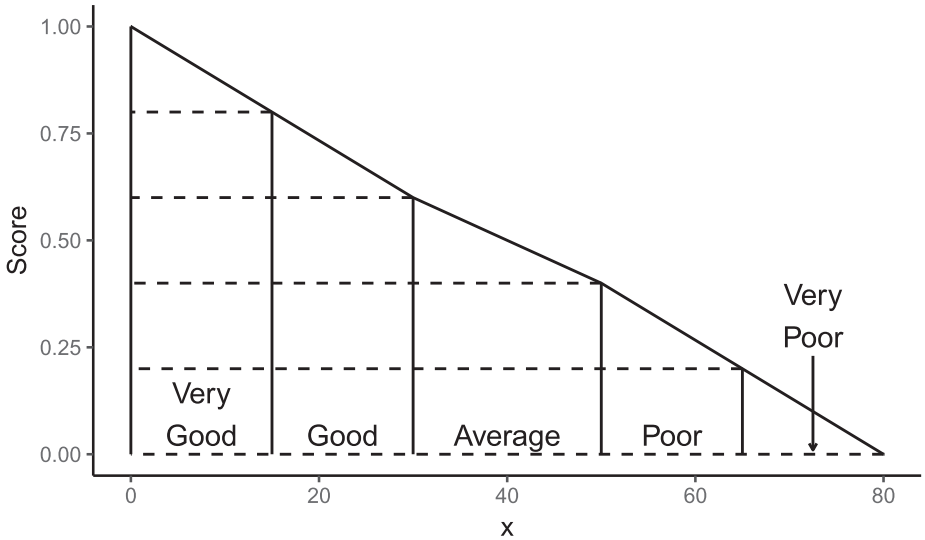
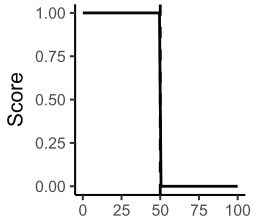
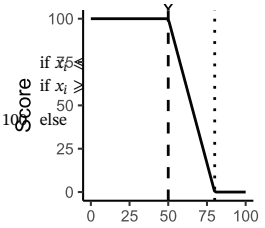
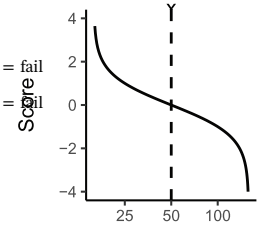
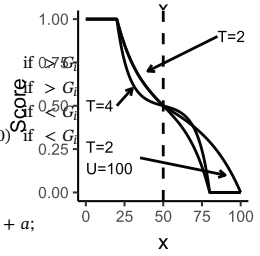
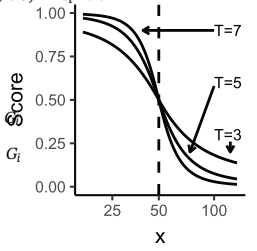


Fig. 2. Observed value to score conversion chart. The solid vertical lines represent the values of the observed data associated with the boundaries of each condition class and the dashed horizontal lines represent the associated score boundaries.

Table 1
Formulations and example response curves for a variety of indicator scoring methods that compare observed values (x_i) to associated benchmark, guidelines or references values (G_i and dashed line). WCS_i (dotted line) represents Worst Case Scenario values as a means to scale the severity of non-compliant (relative to guidelines) observations. In the Hyperbolic Sine (HS) Amplitude method, L_i and U_i represent lower and upper capping bounds and T determines the degree of curvature. In the Logistic Amplitude method, T is a tuning parameter that controls the logistic rate (steepness at the inflection point). For the purpose of example, the benchmark, Worst Case Scenario, lower and upper caps were set to 50,75,20 and 80 respectively.

Method	Formulation	Response curve
Binary	$Score_i = \begin{cases} 1 & \text{if } x_i \leqslant G_i \\ 0 & \text{if } x_i \text{ else} \end{cases}$	
Benchmark and WCS	$Score_i = \begin{cases} 100 & \text{if } x_i \leqslant G_i \\ 0 & \text{if } x_i \geqslant WCS_i \\ \left[1.0 - \frac{x_i - G_i}{WCS_i - G_i}\right].\ 100 & \text{else} \end{cases}$	
Modified Amplitude	$Score_i = \begin{cases} log_2(\frac{x_i}{G_i})^{-1} & \text{if } > G_i = \text{fail} \\ log_2(\frac{x_i}{G_i})^{1} & \text{if } < G_i = \text{fail} \end{cases}$	
HS Amplitude	$s_i = \begin{cases} f(g(x_i), 0, g(L_i), 0, T) & \text{if } > G_i \\ f(g(x_i), g(U_i), 0, -T, 0) & \text{if } > G_i \\ f(-g(x_i), 0, g(L_i), 0, T) & \text{if } < G_i \\ f(-g(x_i), g(U_i), 0, -T, 0) & \text{if } < G_i \end{cases}$ $g(x) = G_i - x;$ $f(x, l, u, a, b) = (b-a)\frac{x-u}{l-u} + a;$ $h(x) = (e^x - e^{-x})/2$	
Logistic Amplitude	$s_i = \begin{cases} -T & \text{if } s_i < -T \\ T & \text{if } s_i > T \\ s_i & \text{else} \end{cases}$ $Score_i = \begin{cases} f(h(s_i), 0, h(T), 0.5, 1) & \text{if } s_i \geqslant 0 \\ f(h(s_i), h(-T), 0, 0, 0.5) & \text{if } s_i < 0 \end{cases}$ $\lambda_i = \begin{cases} -1 & \text{if } > G_i = \text{fail} \\ 1 & \text{if } < G_i = \text{fail} \end{cases}$ $R_i = \begin{cases} -1 \times \frac{G_i}{x_i} - 1 & \text{if } x_i \geqslant G_i \\ \frac{x_i}{G_i} - 1 & \text{if } x_i < G_i \end{cases}$ $Score_i = \frac{1}{1 + e^{\lambda R_i T}}$	

$$Scaled.\ Score_i = \begin{cases} 0 & \text{If } Score_i < -1 \\ 1 & \text{If } Score_i > 1 \\ \frac{1}{2}.\ (Score_i + 1) & \text{else} \end{cases}$$

Asymmetric rescaling across the range ([min, max] e.g. based on historical quantiles) to a new range ([0,1]) can be achieved by:

$$Scaled.\ Score_i = \begin{cases} 1 & \text{if } Score_i > max \\ (1 - 0.5)\frac{Score_i}{max} + 0.5 & \text{if } Score_i \geqslant 0 \\ (0.5)\frac{Score_i - min}{max - min} & \text{if } Score_i < 0 \\ 0 & \text{if } Score_i < min \end{cases}$$

A similar shaped response curve can be achieved via a hyperbolic sine (Table 1), arcsine or inverse hyperbolic tan function. In such formulations, the curvature of the response curve can be adjusted to alter the relative sensitivity close and far from the benchmark and alter the symmetry of this sensitivity differently depending on whether values are above or below the benchmark.

Contrastingly, the Logistic Amplitude method operates on a logit scale such that it is very sensitive to slight fluctuations close to the benchmark and becomes progressively less sensitive with increasing distance. This method is also automatically scaled to the range [0,1]. The steepness of the Logistic Amplitude response can also be controlled by a tuning parameter (T).

Applying a logit function ($1/(1 + e^{Score_i.T})$) to the unscaled Monified Amplitude Scores will also have the effect of rescaling the [− ∞, ∞] range into [0,1] and also offers the potential to adjust the sharpness of the response curve via a tuning (T) parameter (similar to the Logistic Amplitude method, see Table 1).

In an attempt to better represent the health of a complex dynamic ecosystem by a single metric, the Ecosystem Health Monitoring Program (EHMP; which provides health assessments for a large number of marine, estuarine and freshwater ecosystems in south-east Queensland) calculate indicator scores by estimating the proportion of the water body that is compliant (EHMP, 2008; Dobbie and Clifford, 2015). More specifically, spatial interpolation functions (either locally weighted regression or robust quantile regression) are used to yield a median curve (or surface) for each indicator from which a compliance function (C; akin to Binary comparison to benchmark) can be integrated through the entire spatial domain (S).

$$Score_i = \frac{1}{|S|}\int_s C(x_i)\delta x$$

4.4. Modelled indices

In the absence of any reference, benchmark or guideline information, indices can be estimated from statistical models. For example, mixed effects models in which major spatial and temporal elements (such as site and year) are fitted as random effects on longitudinal data, can be used to estimate spatio-temporal changes from the average conditions. Such deviations from the grand mean can then be expressed as probabilities drawn from a Gaussian distribution with a mean of zero and standard deviation equal to the standard deviation of the combined random effects. Probabilities drawn from a Gaussian distribution will follow a sigmoidal response curve similar to that of a logistic curve and thus, sensitivity is greatest close to the grand mean.

4.5. Indicator score aggregation

Multiple indicator scores are typically aggregated into a single *Index* so as to reflect the overall system condition or state. There are numerous ways to aggregate data and often a report card comprises data from a variety of sources (indicator types) as well as spatial and temporal scales. Importantly, given the complexity of natural ecosystems,

Table 2
Fabricated illustration of the discrepancies between total means (i.e. Global index) generated from row means (Zone mean scores) and column means (Global Indicator mean scores).

	Indicators			
Zone	Indicator 1	Indicator 2	Indicator 3	Index
1	0.5	0.2	0.3	0.333
2	0.6			0.600
3	0.6	0.4	0.3	0.433
4	0.7		0.4	0.550
5	0.5	0.3		0.400
Global index	0.580	0.300	0.333	X

If X (mean) is calculated from the 5 row means = 0.463.
If X (mean) is calculated from the three column means = 0.404.

indices of system health are greatly enhanced when interpreted in the context of concurrent estimates of confidence, uncertainty or variability (Dobbie and Clifford, 2015).

4.5.1. Hierarchies and aggregation

Indices are usually calculated at the level of individual observations (e.g site and monthly) for each variable (e.g. total nitrogen), yet report cards are typically presented at coarser spatial and temporal scales (e.g zones/regions and annually) and more general indicator level. Thus it is necessary to aggregate the indices across numerous hierarchies (refer again to Fig. 1).

Although a hierarchical design does offer substantial redundancy and power advantages, it also introduces additional complexities concerning how to combine items that differ in spatio-temporal coverage and how to propagate uncertainty throughout the hierarchy. Consider an example in which the data consist of a number of different indicators measured from a number of sites throughout a number of regions. This represents a double aggregation hierarchy. When sampling effort (or sample size) differs across the hierarchy, the order in which data are aggregated is important. Table 2 provides a simple illustration of the complexity introduced by unbalanced designs and multiple aggregation hierarchies. Hence it is necessary to decide whether the Global index directly reflects the constituent regions or the indicators and aggregate accordingly. The more hierarchical levels, the more complex these decisions become.

4.5.2. Simple deterministic means

The most common deterministic aggregations formulations are summarized in Table 3.

The majority of published indices aggregate indicator scores as simple arithmetic means since these are easy to calculate, understand and interpret. Alternatively, a multiplicative (geometric) mean is useful when formulating a summary from multiple items with substantially different distributional properties. For example, if we were aggregating over three different indicators, two of which were on a [0,1] scale and the other on a [0,100] scale, the later indicator would dominate (eclipse) the arithmetic mean whereas the former (a measure of central tendency) ensures more equal contributions by normalizes the ranges.

Table 3
Formulations for various index aggregation methods. x_i represents the *ith* out of a total of n (sub) indicators and w_i represents the associated unit weights such that $\sum_{i=1}^{n} w_i = 1$.

Method	Formulation
Arithemetic mean	$Index = \frac{\sum_{i=1}^{n} x_i}{n}$
Geometric mean	$Index = (\prod_{i=1}^{n} x_i)^{1/n}$
Weighted arithmetic mean	$Index = \sum_{i=1}^{n} x_i . w_i$
Weighted geometric mean	$Index = \prod_{i=1}^{n} x_i^{w_i}$
Harmonic mean	$Index = \sqrt{\frac{n}{\sum_{i=1}^{n}(1/x_i)}}$
Minimum operator	$Index = min(x_i)$

Weighted aggregations offer a means to assign greater or lesser influence on individual indicators based on the relative ecological importance (on the basis of non-compliance implications) or reliability of individual indicators. Numerous weighting methods have been used for index calculation, most of which involve specialist ecosystem knowledge in conjunction with details about the sampling conditions and QAQC assessments. One major downside of weightings is that modifications to the design (introduction, modification or removal of indicators) potentially necessitate a reevaluation and specification of weights in order to maintain the integrity of the overall compilation (Flint et al., 2012).

Harmonic means are apparently more sensitive to the most impaired indicator (Dojlido et al., 1994) thereby mitigating eclipsing impacts. However, harmonic means are also known to potentially result in ambiguous or inconsistent outcomes (Swamee and Tyagi, 2000). For some situations in which a poor performance of any one of the indicators is considered an indication of poor health (such as seagrass biomass, richness and composition), Smith (1990) advocates the use of minimum operator aggregation. Nevertheless, its total insensitivity to the full suite of available indicators renders it unsuitable for many other purposes (such as water quality).

Median and quantile based aggregations are possible. However, the sparsity of input data along with the use of course grained indicator score algorithms (such as Binomial method), results in indices that are simultaneously under and over sensitive. For example, consider the situation where the scores for ten indicators were {0,0,0,0,0,0,0,0,1,1}. The median is 0. If one (or even two) of the scores of 0 improved to a one, the median would remain 0. If however, three of the zeros became ones, the median would jump to 0.5, after which any additional ones will result in medians of 1.

4.5.3. Incorporating uncertainty

Whilst any of the above algorithms may be adequately suited for the purpose of generating point estimates of an index, they do restrict the ability to propagate estimates of uncertainty across the hierarchy. For arithmetic means of simple hierarchies, it is possible to propagate variance via first (or second) order Taylor expansion of the full covariance matrix (Wang and Iyer, 2005; Mekid and Vaja, 2008).

$$\sigma_{Index}^2 = \sum_{i=1}^{n} j_i^2 \sigma_i^2 + 2 \sum_{\substack{i=1 \\ i \neq k}}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} j_i j_k \sigma_{ik} \quad (1)$$

where j_i are the first order derivatives of the $p \times n$ gradient matrix.

Although relatively simple, this method assumes that the indices are normality distributed. This condition unlikely to be satisfied, particularly given that most are bound by [0,1] and is likely to yield inflated variance estimates. Moreover, the complexity of the calculations increases substantially with each additional level of aggregation. Analogous algorithms are theoretically possible to accompany the other deterministic point estimates.

To facilitate uncertainty estimates into the EHMP, Dobbie and Clifford (2015) proposed a technique that utilizes Monte Carlo simulations drawn from the estimates of median and its association precision. Thus, rather than integrating over a single interpolated curve or surface, integration occurs over a large number of simulated interpolations thereby allowing estimates of variance in addition to median. Similarly, the NEAT (Nested Environmental Status Assessment Tool,http://www.devotes-project.eu) marine assessment tool estimated uncertainty during aggregations via Monte Carlo simulations drawn from normal distributions (based on means and standard errors). Unfortunately, due to limitations of estimating precision (standard error) over multiple aggregations and reliance on strong distributional

assuptions, the Monte Carlo technique does not lend itself so well to a complex hierarchical aggregation schedule.

Bayesian Network

Bayesian Networks (BN; Pearl, 1988) are probabilistic (rather than deterministic) graphical models that represent a set of variables and their inter-dependencies via a directed acyclic graph (DAG). By leveraging Bayesian (conditional) probability theory, BN's offer a relatively simple and elegant way to aggregate proportional (binned) distributions for both quantitative and qualitative data. Estimates of location (e.g. $\bar{y}$) are calculated as the sum of bin midpoints (b_i, e.g. {0.1,0.3,0.5,0.7,0.9}) weighted by their conditional probabilities (c_i). Similarly, estimates of scale are calculated as the sum of the squared differences between bins and the mean weighted by conditional probabilities.

$$\bar{y} = \sum_{i=1}^{n} b_i . c_i \tag{2}$$

$$\sigma_y^2 = \sum_{i=1}^{n} (b_i - \bar{y})^2 . c_i \tag{3}$$

The above formulae could also be modified to yield summaries based on any of the above deterministic means (weighted, geometric, harmonic means etc).

BN's are intuitively interpreted and are considered relatively robust and offer good predictive capacity for the design imbalance and small sample sizes that typically characterize monitoring data. Furthermore, Bayesian probability theory also allows the incorporation of subjective weights and assessments thereby facilitating expert opinion interventions. Johnson et al. (2016) recently demonstrated the potential of applying a BN to the Gladstone Healthy Harbour Partnership Report Card.

In applications where data comprise largely of quantitative data on a continuous scale, the necessity of binning has the potential to substantially reduce the resolution or sensitivity of the resulting index - particularly towards the extremes of 0 and 1 (since the binning phase shrinks the data range away from the Score extremes of 0 and 1 to the extremes imposed by the bin midpoints).

Bootstrapping

Bootstrapping is a simulation process that involves repeated sampling (with replacement) of one or more distributions and offers an alternative approach to propagating uncertainty throughout an aggregation hierarchy. Rather than aggregating the individual distribution averages (as in simple aggregation), bootstrapp aggregation combine the averages of a large number of bootstrapp distributions, each resulting from drawing a single observation from each of the original distributions (see Fig. 3). Hence the result of the aggregation is a distribution (rather than a point estimate) from which measures of uncertainty, such as range, quantiles and confidence intervals can be simply calculated.

Any averaging or statistical summarising method, including those highlighted in Table 3 can be applied during the bootstrapp aggregation process. Indeed, different statistical methods can be applied at different levels of the aggregation if appropriate. Furthermore, it is possible to incorporate qualitative data (such as ratings) into a boostrapp aggregation by enumerating the categories before bootstrapping.

As a stochastic process, repeated calculations will yield different outcomes. Nevertheless, the more bootstrapp samples are collected, the more accurately the resulting aggregated distribution will reflect the true distribution of aggregated values.

Beta approximation

Whilst the bootstrapp aggregation approach described above does offer a robust way to combine data across scales and sources, for large data sets, it does impose large computational and storage burdens. For such cases (large data such as remote sensing), index distributions can be approximated by beta distributions. The beta distribution is defined on the interval [0,1] and is parameterized by two positive shape parameters (α, β) according to the following:

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

A beta function can manifest as many different shapes and as all of these are described by just two shape parameters. Therefore, rather than store all the bootstrapped values for each distribution, we can alternatively approximate each distribution by a beta and store only the defining shape parameters of each distribution. When combining, rather than randomly sample 10,000 stored values of each distribution, we simple resample 10,000 random draws from each beta distribution[1]. The combined distribution can then be approximated by a beta distribution and so on (Fig. 4).

4.5.4. Weighting

In a review of marine ecological assessment tools, Borja et al. (2016) enphasized the importance of being able to assign weights at different spatial, temporal or indicator scales in order to reflect the relative importance or confidence in different information. Weights can be incorporated into an aggregation hierarchy by applying weighted versions of the averaging functions, or additionally in the case of bootstrapp aggregation, by using weights in the resampling routine. In addition to allowing expert driven weights, it is possible to weight according to entropy (based on historical measures of relative variety, see Section 2) or relative to spatial areas during spatial aggregations such that larger areas have proportionally more influence.

4.6. Certainty rating

Incorporating an estimate of scale (variance) into a certainty or confidence rating necessitates re-scaling the estimates into a standard scale. In particular, whereas a scale parameter of high magnitude indicates lower degrees of certainty, for a certainty rating to be useful for end users, larger numbers should probably represent higher degrees of certainty. Thus, the scaling process should also reverse the scale. Furthermore, variance is dependent on the magnitude of the values.

In order to re-scale a scale estimate into a certainty rating, it is necessary to establish the range of values possible for the scale estimate. Whilst the minimum is simple enough (it will typically be 0), determining the maximum is a little more challenging depending on the aggregation algorithm (bootstrapping, Bayesian Network etc). One of the advantages in utilizing proportional distributions (such as is the case for a Bayesian Network or a re-sampled bootstrapp distribution) is that the scale parameter for the single worst case scenario can be devised (once the worst case scenario has been determined) independent of sample sizes or weightings. In most situations this is going to be when the distribution comprises equal mass at (and only at) each of the two extremes (for example, values of just 0 and 1).

The confidence rating discussed above is purely an objective metric derived from the variance in the aggregation hierarchy. It largely fails to incorporate issues such as missing data, outliers and limit of detection - the influences of which on a confidence rating are necessarily subjective. A full confidence rating would combine objective variance components with additional subjective considerations such as climatic and disturbance information, and the perceived influence of missing, limit of detection and outlying data. Hence, the statistical variance would form just one component in the confidence rating system.

The bootstrapp aggregation method provides a mechanism for estimating variance from which to build such an expert considered Confidence Rating system.

[1] Unfortunately there is no closed-form general formula for the sum of multiple independent beta distributions.

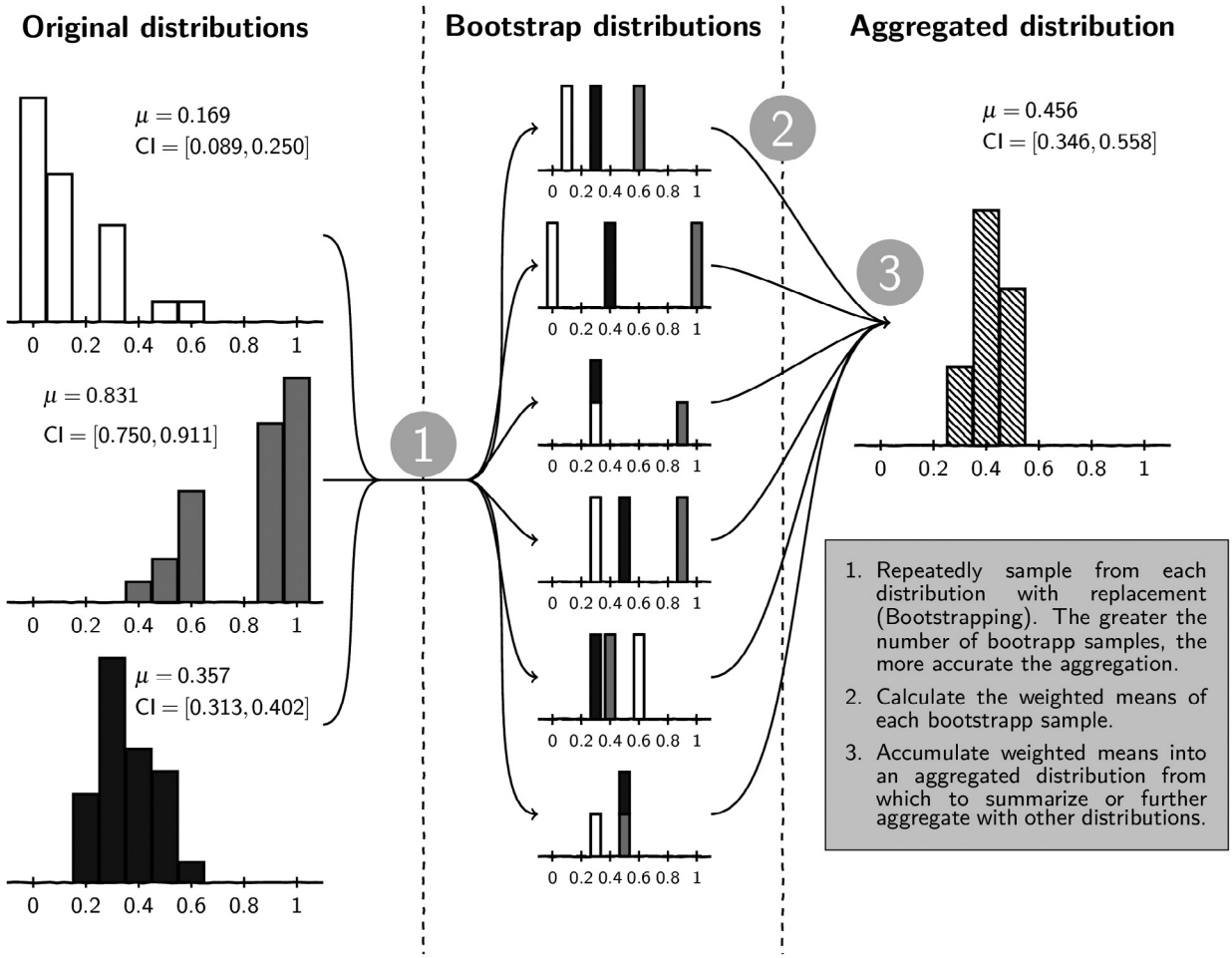


Fig. 3. Illustration of Bootstrapped aggregation of three distributions. The fills of the three original distributions differ in lightness to assist in traching their contribution to the Bootstrapp distributions. Original and Aggregated distributions are summarized by their mean (μ) and 90% confidence interval (CI).

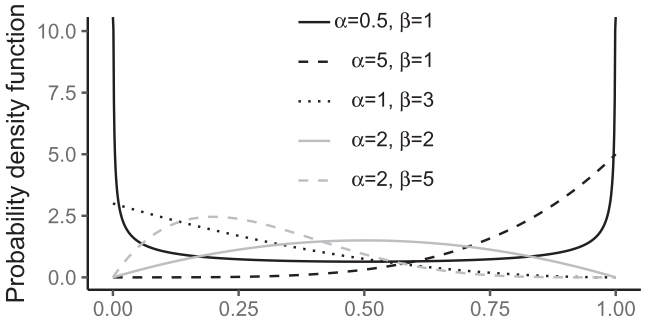


Fig. 4. Beta probability densities.

5. Grades

Indices are typically converted into a single five-point alphanumeric Grade (A-E) or Compliance rating scale along with an associated 'traffic light' like color scheme using a grade conversion charts (see Fig. 5). The grade boundaries are usually determined by experts to ensure that the range of indices represented by each grade classification is congruent with community interpretation of a letter grade report card.

The grade conversion charts adopted by the AIMS inshore water quality Marine Monitoring Program (MMP: Lønborg et al., 2016) and the Gladstone Healthy Harbour Partnership (Gladstone Healthy Harbour Partnership, 2016) both define two levels (Poor and Very Poor) below the threshold values and three above (Satisfactory, Good and Very Good). The threshold is purposely placed at the boundary of two grades so as to ease the distinction between 'pass' and 'fail'. The major difference between these two charts is that whereas the AIMS MMP report card grade conversion chart partitions the three better than threshold categories equally, the Gladstone Healthy Harbour Partnership report card grade conversion chart employs numerically simpler boundary cutoffs around the 'B' grade (although this does result in arbitrarily unequal category sizes).

By contrast, the MidCoast Council (formally Great Lakes Council) Waterway and Catchment Report (MidCoast Council, 2016) uses grade boundaries based on historical score distribution quantiles associated with definitions of what proportion of total observations (sites) are considered 'Excellent' (A), 'Good' (B), 'Fair' (C), 'Poor' (D) and 'Very Poor' (Fig. 5d). For example, the 'Very Poor' grade was defined as the worst 5% of sites across the entire State of New South Wales and the lowest 5% of sites has a maximum score of 0.4. This approach recognizes the non-linear spread of scores resulting from their particular metrics and attempts to ensure that grades are intuitively interpretable (a grade of A means the site is in Excellent condition). Nevertheless, it does necessitate availability of historical data and as well as a very specific and agreed upon set of a priori condition definitions.

It is far less clear how estimates of uncertainty can be incorporated into such a grading scheme in a manner that will be intuitive to non-technical audiences. That said, statistical uncertainty is just one of many sources of uncertainty that should be captured in a confidence or certainty rating. Hence any expectations of presenting the full extent of uncertainty in a quantitative manner may well be unrealistic.

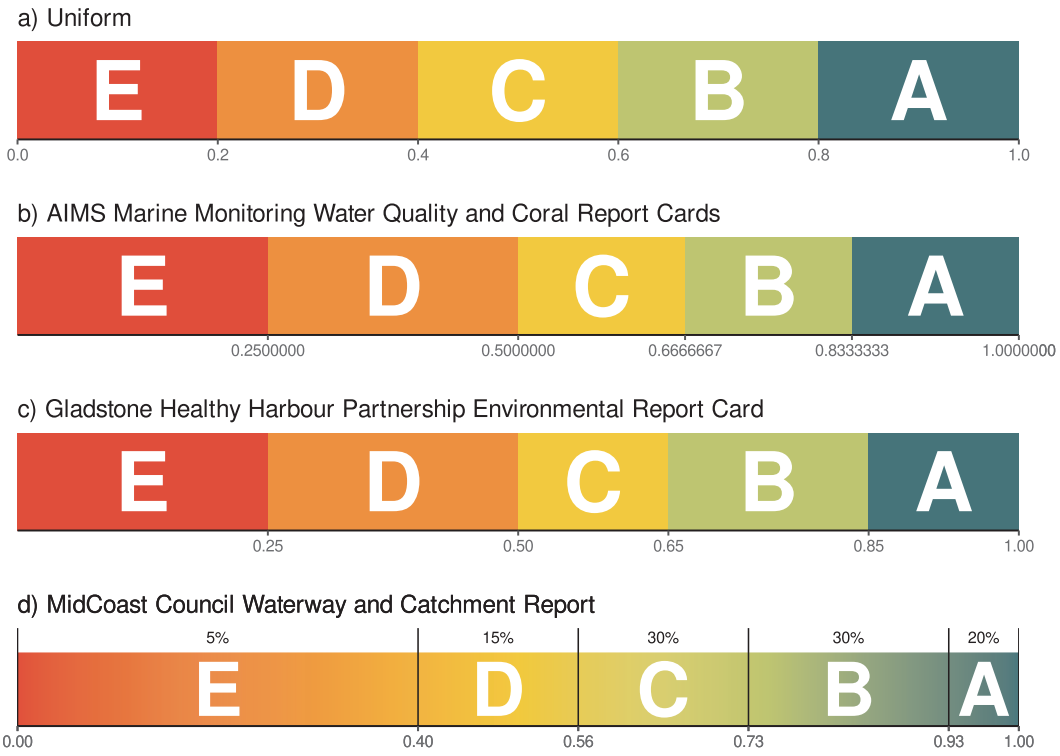


Fig. 5. Examples of score to grade conversion charts. In each case, the scale along the base defines the grade boundaries. For d) the scale along the top represents historical score distribution quantiles.

Supplimentary materials

There is an R package hosted on github (https://github.com/AIMS/reportcards) to provide code in support of many of the analytical aspects discussed here.

Acknowledgements

This project was funded from the Key Program for International S&T Cooperation Projects: Sino-Australian Center for Healthy Coasts (No. 2016YFE0101500), as well as the Australia China Science Research Fund grant ACSRF 48162, Australian Government Department of Industry, Innovation and Science. We would also like to acknowledge the invaluable inputs of David Fox, Uthpala Pinto, John Rolfe, John Kirkwood, Angus Thompson, Cedric Robillot and Lyndon Llewellyn as well as the two anonymous reviewers whose constructive comments greatly improved the manuscript.

References

Attard, M., Thompson, M., Kelly, R., Locatelli, A., 2012. Tamar Estuary Ecosystem Health Assessment Program Monitoring Report 2012. Report to NRM North Tamar Estuary and Esk Rivers Program. Technical Report.

Borja, A., Dauer, D.M., Gremare, A., 2012. The importance of setting targets and reference conditions in assessing marine ecosystem quality. Ecol. Ind. 12, 1–7.

Borja, A., Elliott, M., Andersen, J.H., Berg, T., Carstensen, J., Halpern, B.S., Heiskanen, A.S., Korpinen, S., Lowndes, J.S.S., Martin, G., Rodriguez-Ezpeleta, N., 2016. Overview of integrative assessment of marine systems: The ecosystem approach in practice. Front. Marine Sci. 3, 20. URl: https://www.frontiersin.org/article/10.3389/fmars.2016.00020.

CCME, 2001. Canadian water quality guidelines for the protection of aquatic life: CCME Water Quality Index 1.0, Technical Report. Technical Report. CCME, Winnipeg, Manitoba, Canada.

Coates, S., Waugh, A., Anwar, A., Robson, M., 2007. Efficacy of multimetric fish index as an analysis tool for the transitional fish component of the water framework directive. Mar. Pollut. Bull. 55, 225–240. https://doi.org/10.1016/J.MARPOLBUL.2006.08.029.

Conner, C., Dennison, W., Thomas, J., 2010. Communication strategy: packaging and delivering the message for maximum impact. IAN Press, Cambridge, MD chapter Integrating and Applying Science: A practical handbook for effective coastal ecosystem assessment. pp. 45–58.

Connolly, R.M., Bunn, S., Campbell, M., Escher, B., Hunter, J., Maxwell, P., Page, T., Richmond, S., Rissik, D., Roiko, A., Smart, J., Teasdale, P., 2013. Review of the use of report cards for monitoring ecosystem and waterway health. Report to: Gladstone Healthy Harbour Partnership. Technical Report. Queensland, Australia.

Dauvin, J.C., Fisson, C., Garnier, J., Lafite, R., Ruellet, T., Billen, G., Deloffre, J., Verney, R., 2008. A report card and quality indicators for the seine estuary: from scientific approach to operational tool. Mar. Pollut. Bull. 57, 187–201.

Dennison, V.C., Lookingbill, T.R., Carruthers, T.J.B., Hawkey, J.M., Carter, S.L., 2007. An eye-opening approach to developing and communicating integrated environmental assessments. Front. Ecol. Environ. 5, 307–314.

Dobbie, M.J., Clifford, D., 2015. Quantifying uncertainty in environmental indices: an application to an estuarine health index. Marine Freshwater Res. 66, 95–105.

Dobbie, M.J., Dail, D., 2013. Robustness and sensitivity of weighting and aggregation in constructing composite indices. Ecol. Ind. 29, 270–277.

Dojlido, J., Raniszewski, J., Woyciechowska, J., 1994. Water quality index applied to rivers in the vistula river basin in poland. Environ. Monit. Assess. 33, 33–42.

Doran, G.T., 1981. There's a s.m.a.r.t. way to write management's goals and objectives. Management Review (AMA FORUM) 70, 35–36.

Doren, R.F., Trexler, J.C., Gottlieb, A.D., Harwell, M.C., 2009. Ecological indicators for system-wide assessment of the greater everglades ecosystem restoration program. Ecol. Ind. S2–S16. https://doi.org/10.1016/j.ecolind.2008.08.009.

EEA, 1995. Europe's Environment: the Dobris Assessment. Technical Report. European Environmental Agency. Copenhagen.

EHMP, 2008. Ecosystem Health Monitoring Program 2006–07 Annual Technical Report. Technical Report. South East Queensland Healthy Waterways Partnership: Brisbane.

Emerson, J., Hsu, A., Levy, M., de Sherbinin, A., Mara, V., Esty, D., Jaiteh, M., 2012. Environmental performance index and pilot trend environmental performance index. Yale Center for Environmental Law and Policy, New Haven Technical Report.

EPA/EC, 1995. State of the Great Lakes 1995. Washington (DC). Technical Report. Environmental Protection Agency and Environment Canada.

Flint, N., Rolfe, J., Jones, C., Sellens, C., Rose, A., Fabbro, L., 2012. Technical review for the development of an ecosystem health index and report card for the Fitzroy Partnership for river health. Part A: Review of ecosystem health indicators for the Fitzroy Basin. Centre for Environmental Management, Central Queensland University Technical Report.

Fox, D., 2013. Statistical issues associated with the development of an ecosystem report card. Client report to Gladstone Healthy Harbour Partnership. Technical Report. Environmetrics Australia.

Gari, S.R., Newton, A., Icely, J.D., 2015. A review of the application and evolution of the DPSIR framework with an emphasis on coastal social-ecological systems. Ocean Coastal Manage. 103, 63–77. https://doi.org/10.1016/j.ocecoaman.2014.11.013.. URLhttp://www.sciencedirect.com/science/article/pii/S0964569114003652.

GBR, 2014. Great Barrier Reef Report Card. Technical Report. Australian and Queensland Governments, Australia. Australian and Queensland Governments, Australia. URLhttps://www.reefplan.qld.gov.au/tracking-progress/reef-report-card.

Gladstone Healthy Harbour Partnership, 2016. Technical Report, Gladstone Harbour Report Card 2016, GHHP Technical Report No.3. Technical Report. Gladstone Healthy Harbour Partnership, Gladstone.
Hallett, C.S., Valesini, F.J., Clarke, K.R., Hesp, S.A., Hoeksema, S.D., 2012. Development and validation of fish-based, multimetric indices for assessing the ecological health of western australian estuaries. Estuar. Coast. Shelf Sci. 104–105, 102–113.
Harwell, M., Myers, V., Young, T., Bartuska, A., Gassman, N., Gentile, J.H., Appelbaum, S., Barko, J., Causey, B., Johnson, C., McLean, A., Smola, R., Templet, P., Tosini, S., 1999. A framework for an ecosystem integrity report card: examples from south florida show how an ecosystem report card links societal values and scientific information. Bioscience 49, 543–556. URL: http://www.jstor.org/stable/10.1525/bisi.1999.49.7.543.
Hijuelos, A., Reed, D., 2013. Methodology for Producing a Coastal Louisiana Report Card, September 13, 2013. The Water Institute of the Gulf Technical Report.
Hurley, T., Sadiq, R., Mazumder, A., 2012. Adaptation and evaluation of the canadian council of ministers of the environment water quality index (CCME WQI) for use as an effective tool to characterize drinking source water quality. Water Resour. 46, 3544–3552.
James, C.A., Kershner, J., Samhouri, J., O'Neil, S., Levin, P.S., 2012. A methodology for evaluating and ranking water quantity indicators in support of ecosystem-based management. Environ. Manage. 49, 703–719.
Johnson, S., Logan, M., Fox, D., Kirkwood, J.U.P., Mengersen, K., 2016. Environmental decision-making using bayesian networks: creating an environmental report card. Appl. Stochastic Models Bus. Ind. https://doi.org/10.1002/asmb.2190.
Jones, C., Flint, Rolfe, J., Sellens, C., Fabbro, L., 2013. Technical review for the development of an ecosystem health index and report card for the Fitzroy Partnership for river health. Part B: Methodology and data analysis to support an ecosystem health index and report card for the Fitzroy Basin. Centre for Environmental Management, Central Queensland University Technical Report.
Kannel, P.R., Lee, S., Lee, Y.S., Kanel, S.R., Khan, S.P., 2007. Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment. Environ. Monit. Assess. 132, 93–110.
Kuhnert, P., Bartley, R., Peterson, E., Browne, M., Harch, B., Steven, A., Gibbs, M., Henderson, A., Brando, V., 2007. Conceptual and statistical framework for a water quality component of an integrated report card for the Great Barrier Reef catchments. Reef and Rainforest Research Centre Limited, Cairns Unpublished report to the Marine and Tropical Sciences Research Facility. Technical Report.
Lønborg, C., Devlin, M., Brinkman, R., Costello, P., da Silva, E., Davidson, J., Gunn, K., Logan, M., Petus, C., Schaffelke, B., Skuza, M., Tonin, H., Tracey, D., Wright, M., Zagorskis, I., 2016. Reef Rescue Marine Monitoring Program. Annual Report of AIMS and JCU Activities 2014 to 2015–Inshore water quality monitoring. Report for the Great Barrier Reef Marine Park Authority. Technical Report. Australian Institute of Marine Science and JCU TropWATER. Townsville. 170pp.
McIntosh, E.J., Rolfe, J., Pinto, U., Kirkwood, J., Greenlee, M., Poiner, I.R., 2019. Designing report cards for aquatic health with a whole-of-system approach: gladstone harbour in the great barrier reef. Ecol. Ind. 102, 623–632. https://doi.org/10.1016/j.ecolind.2019.03.012.. URLhttp://www.sciencedirect.com/science/article/pii/S1470160X19301931.
Mekid, X., Vaja, D., 2008. Propagation of uncertainty: Expressions of second and third order uncertainty with third and forth moments. Measurement 41, 600–609.
MidCoast Council, 2016. MidCoast Council 2016 Waterway and Catchment Report. Technical Report. MidCoast Council, Natural Systems and Estuaries Section.
Miller, W.W., Joung, H.M., Mahannah, C.N., Garret, J.R., 1986. Identification of water quality differences in nevada through index application. J. Environ. Qual. 15, 265–272.
Neary, B.P., 2012. A sensitivity analysis of the Canadian Water Quality Index. A report for CCME prepared by Gartner Lee Limited, Ontario, Canada. Technical Report.
Newall, P., Tiller, D., Lloyd, L.N., 2012. Technical Report for Freshwater Monitoring Framework and Report Card for the Tamar Estuary and Esk Rivers Program. Report to NRM North. Technical Report. Lloyd Environmental Pty Ltd, Syndal, Victoria, Australia.
OECD, 1993. OECD Core Set of Indicators for Environmental Performance Reviews. Technical Report. Organization for Economic Cooperation and Development. Paris, France.
Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.
Raican, S.M., Wang, Y.G., Harch, B., 2013. Water Quality Indices from Unbalanced Spatio-temoral Monitoring Designs. Nova Publishers, New York, NY chapter Water quality: indicators, human impact and environmental health. pp. 1–30.
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. Series A (General) 168, 307–323.
Sarkar, C., Abbasi, A., 2006. Qualidex – a new software for generating water quality indices. Environ. Monit. Assess. 119, 201–231.
Schiller, A., Hunsaker, C.T., Kane, M.A., Wolfe, A.K., Dale, V.H., Suter, G.W., Russell, C.S., Pion, G., Jensen, M.H., Konar, V.C., 2001. Communicating ecological indicators to decision makers and the public. Conserv. Ecol. 5, 19. URLhttp://www.consecol.org/vol5/iss1/art19/.
SFEI, 2015. The Pulse of the Bay: The State of Bay Water Quality, 2015 and 2065. San Francisco Estuary Institute Technical Report.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1145/584091.584093.
Smith, D.G., 1990. A better water quality indexing system for rivers and streams. Water Res.
Swamee, P.K., Tyagi, A., 2000. Describing water quality with aggregate index. J. Environ. Eng. 126, 451–455.
Thompson, A., Lønborg, C., Costello, P., Davidson, J., Logan, M., Furnas, M., Gunn, K., Liddy, M., Skuza, M., Uthicke, S., Wright, M., Zagorskis, I., Schaffelke, B., 2014. Marine Monitoring Program. Annual Report of AIMS Activities 2013 to 2014 Inshore water quality and coral reef monitoring. Report for the Great Barrier Reef Marine Park Authority. Technical Report. Australian Institute of Marine Science. Townsville. 143pp.
Walsh, P., Wheeler, W., 2012. Water quality index aggregation and cost benefit analysis. Working Paper # 12–05. Technical Report. U.S. Environmental Protection Agency, National Center for Environmental Economics. Washington, DC.
Wang, C.M., Iyer, H.K., 2005. On higher-order corrections for propagating uncertainties. Metrologia 42, 406–410.
Whittaker, G., Lautenbach, S., Volk, M., 2012. What is a good index? Problems with statistically based indicators and the malmquist index as alternative. In: Seppelt, R., Voinov, A.A., Lange, S., Bankamp, D. (Eds.), International Environmental Modelling and Software Society (iEMSs) 2012 International Congress on Environmental Modelling and Software Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany.
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.
Williams, M.R., Longstaff, B.J., Wicks, E.C., Carruthers, T.J.B., Florkowski, L.N., 2010. Report Cards: Integrating Indicators into Report Cards. IAN Press, Cambridge, Maryland chapter Integrating and applying science: a practical handbook for effective coastal ecosystem assessment.
Zou, Z.H., Yun, Y., Sun, J.N., 2006. Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. J. Environ. Sci. (China) 18, 1020–1023.