PO Box 7117
Beaumaris VIC, 3193
+61-3-9018-7121
www.environmetrics.net.au

An Environmetrics Australia Technical Report

Statistical issues associated with the development of an ecosystem report card

By
David R. Fox,
PhD.,C.Stat.,P.Stat.,C.Sci.

13 December 2013

Contents

List of Figures

1. Introduction

Environmetrics Australia was engaged by the Gladstone Health Harbour Partnership (GHHP) to provide statistical advice to its Independent Scientific Panel (ISP) to assist in the identification of key issues associated with the development of a report card framework. The advice sought was general in nature and not intended to articulate the methodological detail associated with index development, report card scoring or monitoring program development.

Within the limited time available we have:

- Undertaken a review of relevant literature, web-sites, technical reports and conference presentations;
- Reviewed methodologies associated with ecosystem scorecards currently being used in Queensland;
- Accessed and analysed locally-relevant data to help highlight potential difficulties with current benchmarking and grading processes;
- Participated in the ISP meeting in Gladstone on December 9, 2013.

This report is a synthesis of those activities and forms the basis of the following set of recommendations to the GHHP ISP.

Recommendations

The following recommendations are provided to assist with the development of a report card framework and integrated monitoring program for the GHHP:

Indicator / index development

1. Adopt a staged approach such as that used for the development of the SEQ;
2. Undertake targeted investigations using *existing data* to investigate and assess:
    - the merits of various computational methods such as the CCME WQI method; and
    - the implications of equal and unequal weighting schemes.
3. Undertake validation study using group of experts and methodology provided here (or suitable alternative);
4. Using *validated* indices, apply to existing data to quantify spatial correlation structure and temporal variation.

Report card development

1. As a matter of priority, undertake a project to re-evaluate index aggregation and scoring methodologies in current use (eg. Fitzroy Basin and EHMP)
    - Investigate alternatives / modifications that better deal with distributional changes in indicators other than gross shift in *location* (eg. mean);
2. Undertake validation study using group of experts to:

(a) Assist in the development of a ‘formula’ to convert the aggregated and (possibly) weighted indices to a suitable report card grade;

(b) ‘Road-test’ this formula by applying to existing data to establish that the resulting grades: accord well with expert assessment; reflect meaningful changes in ecosystem status; and adequately reflect differences between sub-regions.

Monitoring program development

1. Adopt a high-level framework such as that suggested in the National Water Quality Management Strategy (ANZECC/ARMCANZ 2000) to assist in the identification of sample design elements.

2. Use the process outlined here (Figure 18) (or similar) to ensure the logical sequencing of additional investigative and validation studies required to inform the monitoring program design.

3. Develop field sampling and data analysis protocols on the basis of final decisions associated with: index computation; sub-region identification; and report grading ‘formula’.

2. Indices

2.1 Rationale

Policy-makers, NRM managers and the general public are all interested in the “state of environment” at multiple levels – for example, as a statement of overall condition at a single place and time or, as is more often the case, an assessment of spatial-temporal trends. Given the enormous number of parameters, metrics, and methods available to quantify specific elements of ecosystem health, it has been necessary to prioritise and aggregate this data to reduce the dimensionality of the problem in order to (a) make an overall assessment of condition and (b) reduce or eliminate conflicting assessments based on considerations of individual parameters.

The approach that has been widely adopted since the 1970s to measure and monitor ecosystem health mimics that of financial markets through the construction of *indices*. In the financial context, the primary use of an index is not to say something about an individual stock, but rather to give an overall picture of a complex financial system and to track the performance of that system over time. The same is true in an environmental setting but with the additional requirement for the constructed index or set of indices to unravel complex space-time interactions as well as measure the effectiveness of deliberate interventions designed to move the ecosystem to a preferred status. Although succinct and reasonable, it is this latter assessment that can be extremely difficult to answer accurately and coherently (Jordan and Vaas 2000).

While environmental managers, politicians, and the general public have enthusiastically embraced the report card concept and the environmental indices that underpin them, a significant challenge for environmental scientists now is to rationalise the plethora of metrics and computational procedures to avoid recreating the dimensionality problem that indices were meant to solve. The issue of multiple indices was discussed at a special session of the 2009 Coastal and Estuarine Research Federation meeting and concluded “now are *(sic)* the time to evaluate existing alternatives and identify preferred approaches, rather than spending energy developing additional indices. The challenge for the next decade is to accomplish sufficient index performance comparisons to reach scientific consensus on preferred index approaches for each biological element that managers wish to include” (Borja et al. 2009).

2.2 Objectives

In discussing uses and requirements of indices and later, report cards it is useful to make a distinction between an environmental *indicator* and an environmental *index*. For this purpose we have adopted (with minor modification) the definitions used by the Center for International Earth Science Information Network (de Sherbinin et al. 2013).

> **Environmental indicators** are metrics derived from observation used to identify indirect drivers of environmental problems (eg. population growth), direct pressures on the environment (eg. overfishing), environmental condition (eg. contaminant concentrations), broader impacts of environmental condition (eg. health outcomes), or effectiveness of policy responses.
>
> **Environmental index** is a dimensionless number obtained by aggregating a number of environmental indicators. While not a requirement, the aggregation process usually utilises some form of weighted averaging.

The uses of both indicators and indices are many and varied although the main ones identified in de Sherbinin et al. (2013) based on Failing and Gregory (2003) include:

a. To discriminate among competing hypotheses (for scientific exploration);
b. To structure understanding of issues and conceptualize solutions;
c. To track performance as determined by results-based management;
d. To discriminate among alternative policies either for specific decisions or general policy directions; and
e. To inform general users (public, stakeholders, community).

The usefulness and hence uptake of an index depends on the degree of 'resonance' with the target audience. Thus a scientist will use and respond strongly to index type (a) while an NRM manager will most likely be interested in a type (c) index. Hezri and Dovers (2006) contend that indicator resonance is a function of *content* and *legitimacy*. Briefly, *content* is associated with the validity, reliability, and timeliness while *legitimacy* concerns the degree to which indicators incorporate alternative viewpoints, are consistent with dominant political and social norms, and are constructed in a fair and transparent manner (de Sherbinin et al. 2013).

While much has been written about indicators and indices, the ultimate requirements are:

1. To answer the question "How good or how bad are current conditions"?
2. To generate an accurate and holistic evaluation of spatial and temporal trends.
3. To utilise methods that are scientifically valid and easily understood by professionals and the public.

(adapted from Kaurish and Younos, 2007)

The so-called 'information pyramid" shown in Figure 1 captures the main elements of index construction, the data foundations and relationships with audience and message complexity.

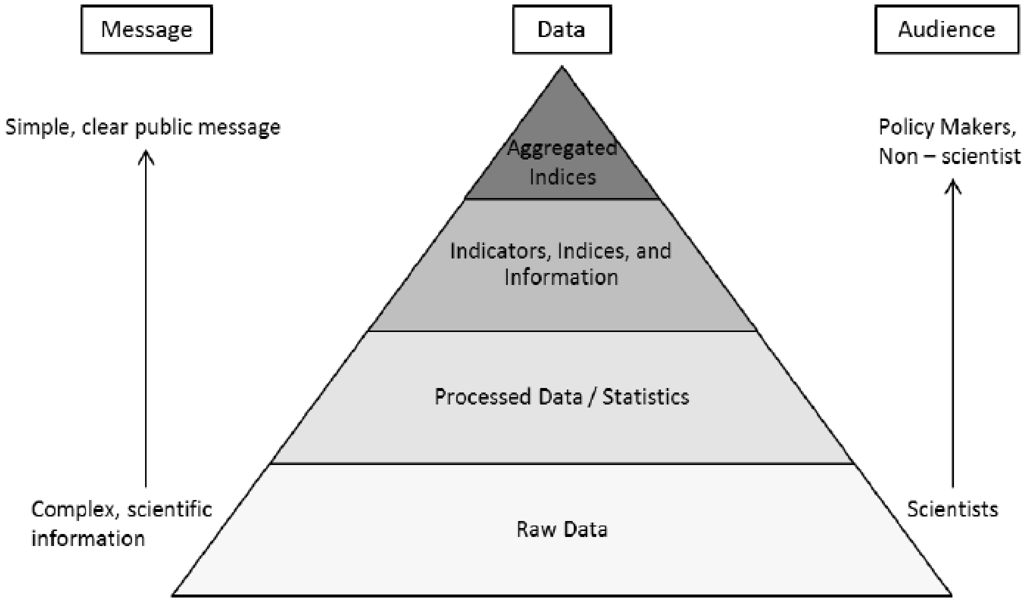

Message
Data
Audience
Simple, clear public message
Complex, scientific information
Aggregated Indices
Indicators, Indices, and Information
Processed Data / Statistics
Raw Data
Policy Makers, Non – scientist
Scientists

Figure 1. The 'information pyramid' showing interactions between indices and complexity of messages for different audiences. *(Source: Hijuelos and Reed 2013).*

We next consider factors affecting the choice of a particular index and the data needed to compute it.

2.3 Criteria for index and data selection

Dauvin et al. (2008) invoked the SMART principle (Simple, Measurable, Achievable, Realistic, and Time limited) to help guide the selection of indicators for the development of a report card for the Seine estuary. This principle reflects considerations of *quality* and *usefulness*. A ‘good’ indicator must:

- be representative;
- be appropriate to the time span and spatial scale of the characterized phenomenon;
- easy to interpret;
- comparable across multiple jurisdictions;
- able to show the principal changes in space and time; and
- have a reference or threshold value.

while to be useful, the indicator must:

- be approved by expert consensus;
- be well grounded and well documented; and
- have a reasonable cost/benefit ratio.

The Yale Center for Environmental Law and Policy (YCELP) and the Center for Earth Information Science Information Network (CIESIN) at Columbia University have developed the Environmental Sustainability Index (ESI) in response to a claimed inability of many quantitative metrics to effectively demonstrate improved environmental performance (Emerson et al. 2012). Criteria used to select indicators were based on considerations of:

- *relevance* (indicator is widely applicable);
- *performance* (indicator responds to and reflects altered environmental status);
- *credibility* (indicator has been peer-reviewed); and
- *completeness* (data used to construct the indicator has adequate historical and on-going spatial-temporal coverage).

A total of 16 criteria was used to select indicators for the Ecosystem Health Index and report card for the Fitzroy Basin and these were grouped on 4 dimensions covering *data*; *interpretation and communication*; *relevance*; and *practicality and timeliness* (Box 1).

In the context of river ecosystem health, Bunn et al. (2010) suggest that index development should also include aspects of organisation (eg. biodiversity, species composition); vigour (eg. rates of production, nutrient cycling); and resilience.

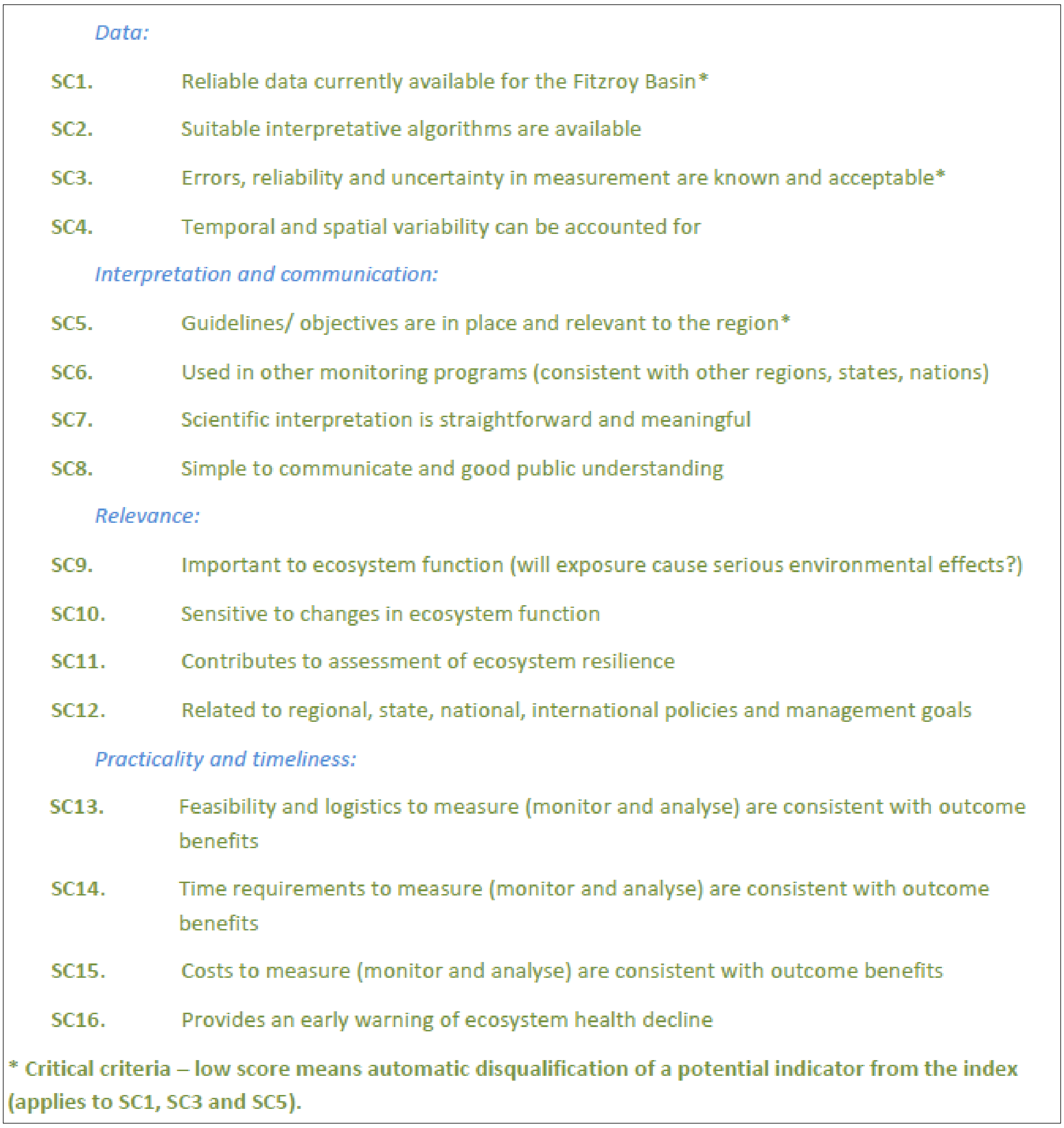
For the Fitzroy freshwater catchments and estuary, Flint et al. (2012) recommended that a variety of indicators be selected that reflect the status of:

- physical and chemical parameters
- nutrients;
- toxicants; and
- ecology (ecosystem processes; habitat; invertebrates; fish).

while indicators for the marine areas were chosen so as to align with the Reef Water Quality Protection Plan and include:

- water quality (Chla; TSS);
- seagrass (abundance; reproductive effort; nutrient status);
- corals (cover; composition; macroalgae cover; juvenile density)

The issue of *how many indicators* to include in the development of an overall index of ecosystem health appears to be an open-ended question with relatively few published articles devoted to or even tackling this issue. A consultant's review of the Canadian Water Quality Index noted that there was little uniformity in approaches to indicator selection and that "it is possible to manipulate the outcome of the Water Quality Index by including large numbers of parameters for which there is no exceedence of guidelines" (Neary 2012). In the context of the Canadian Water Quality Index (CWQI), Neary (2012) recommended a minimum of <u>seven parameters</u> be used at each site and a minimum of <u>six samples</u> be included for each index period while noting that the *type* of parameters selected is more influential than the *number* of parameters although there are no guidelines for selecting an 'optimal' combination of parameters.



Data:

SC1. Reliable data currently available for the Fitzroy Basin*

SC2. Suitable interpretative algorithms are available

SC3. Errors, reliability and uncertainty in measurement are known and acceptable*

SC4. Temporal and spatial variability can be accounted for

Interpretation and communication:

SC5. Guidelines/ objectives are in place and relevant to the region*

SC6. Used in other monitoring programs (consistent with other regions, states, nations)

SC7. Scientific interpretation is straightforward and meaningful

SC8. Simple to communicate and good public understanding

Relevance:

SC9. Important to ecosystem function (will exposure cause serious environmental effects?)

SC10. Sensitive to changes in ecosystem function

SC11. Contributes to assessment of ecosystem resilience

SC12. Related to regional, state, national, international policies and management goals

Practicality and timeliness:

SC13. Feasibility and logistics to measure (monitor and analyse) are consistent with outcome benefits

SC14. Time requirements to measure (monitor and analyse) are consistent with outcome benefits

SC15. Costs to measure (monitor and analyse) are consistent with outcome benefits

SC16. Provides an early warning of ecosystem health decline

*** Critical criteria – low score means automatic disqualification of a potential indicator from the index (applies to SC1, SC3 and SC5).**

Box 1. Indicator selection criteria used in developing EHIs for the Fitzroy Basin. *(Source: Flint et al. 2012).*

There is no agreed or standardised *process* or *mechanism* by which candidate metrics are identified; screened; and selected for follow-up investigation. Borja and Dauer (2008) suggest a sequential process although this has no feedback or decision-points (Figure 2) while a more complex, two-phase process was adopted by the SEQ Healthy Waterways Partnership to select freshwater indicators. Candidate indicators, monitoring protocols and a classification taxonomy were identified in Phase I while short-listed indicators were trialled in a Phase II field study (Figure 3).

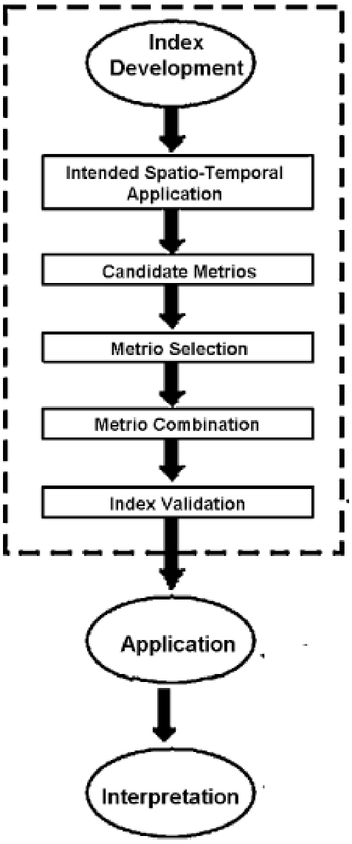


Figure 2. Steps involved in developing and using an EHI. *(adapted from Borja and Dauer, 2008)*

2.4 Data collection and statistical QA/QC

The integrity of a performance measure is highly reliant upon the accuracy and reliability of the data used to derive the measure (Hijuelos and Reed 2013). This presents a significant challenge for the construction of environmental health indices which invariably rely on gathering data from multiple sources having varying and possibly at times, unknown standards of collection, QA/QC, and analytical procedures. A general set of data quality attributes attributable to Maggino and Zumbo (2012) and used in the development of the Louisiana Report Card (Hijuelos and Reed 2013) is reproduced in Box 2.

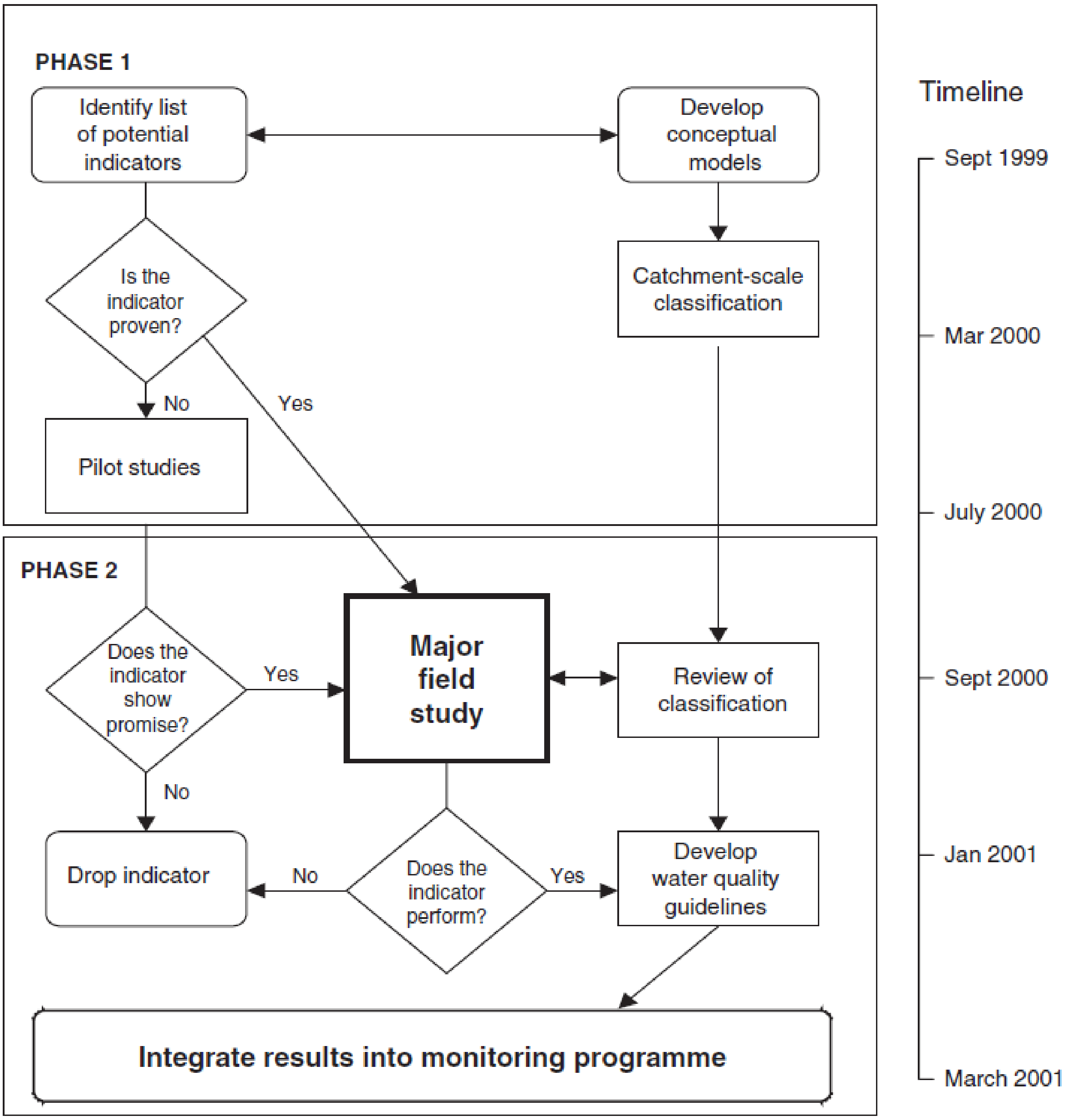

PHASE 1
Identify list of potential indicators
Develop conceptual models
Is the indicator proven?
Catchment-scale classification
No
Yes
Pilot studies
PHASE 2
Does the indicator show promise?
Yes
Major field study
Review of classification
No
Drop indicator
No
Does the indicator perform?
Yes
Develop water quality guidelines
Integrate results into monitoring programme
Timeline
Sept 1999
Mar 2000
July 2000
Sept 2000
Jan 2001
March 2001

Figure 3. Schematic showing steps involved in indicator selection for the SEQ Regional Water Quality Management Strategy. *(Source: Bunn et al. 2010).*

1. **Methodological Soundness**

- Internationally accepted standards, guidelines, or good practices should be employed for data collection efforts.
- Performance measures should be based upon data sources and statistical techniques that are regularly assessed and validated to ensure accuracy and reliability of measurements. The accuracy of an estimate involves analyzing the total error associated with the estimate: sampling error and measurement error.

2. **Integrity**

- The principle of objectivity in the collection, compilation, and dissemination of data, statistics, and results should be adhered to ensure professionalism in statistical policies and practices, transparency, and ethical standards.

3. **Serviceability**

- Data users and their expectations should be identified in order to adequately meet their needs.
- Data should be timely with respect to the length of time between its availability and the event it describes.
- Data should be regularly analyzed in order to record differences and disparities between units, groups, geographical areas and so on, by employing the available information as much as possible.

4. **Accessibility**

- Presentations and documentations concerning data and metadata should be clearly accessible.
- Data should be easily findable, accessible, useable, analyzable, and interpretable in order to gain users' confidence.

Box 2. Data quality assurance criteria. *(Source: Hijuelos and Reed 2013)*

An often overlooked (or poorly executed) aspect of the data collection and processing function is the area of *statistical* QA/QC. This is quite distinct from the more familiar analytical and laboratory QA/QC procedures and refers to the set of statistical activities associated with the identification and treatment of missing and/or 'aberrant' observations and other data 'cleansing' activities. By aberrant we mean any data value that is in some way 'unusual'. This unusualness can arise in many and varied ways with the most common being attributable to 'outliers'. It is important to note however that while all outliers are unusual the converse is not necessarily true. It is beyond the scope of this document to go into more detail and this will be the subject of future discussions as the monitoring program takes shape. For the time being, we shall simply flag that attention needs to be paid to the generic set of activities that define statistical QA/QC which include:

- Treatment of missing data (including the use of models and data imputation techniques);
- Methods for detecting aberrant observations – particularly in a multivariate context;
- Data transformations (eg. to stabilise variance; restore normality); and
- Statistical calibration and error detection.

On the last dot-point above, we note that for the recently completed Western Basin Dredging and Disposal Project (WBDDP) *Environmetrics Australia* developed a number of tools and algorithms to process large quantities of turbidity and PAR data generated from telemetered in situ loggers. Our Anomalous Data Detection Macro (*ADDM*) has been in continuous use for over the past year and is used by Vision Environment Queensland to screen large volumes of data generated from dual instruments and to flag asynchronous periods and identify the unreliable instrument.

2.5 Computational and statistical aspects

Methods for computing an index are discussed in section 2.5.2 below. In this section, we focus on quantitative methods and statistical procedures in the development stage to ensure the resulting index satisfies the objectives outlined in section 2.2. These activities will necessarily be dictated by the specific circumstances, data, intended application and a number of other factors and we are thus unable to be prescriptive about recommended statistical methodologies. Nevertheless, as a guide it is imperative that the data processing activities used to develop and validate an index be:

- **Targeted** – avoid the ‘shotgun’ approach where numerous statistical methods are trialled in order to see what works best. In making this recommendation we acknowledge that an element of relatively unstructured *Exploratory Data Analysis* (EDA) is required in the early stages, however as development proceeds, more targeted statistical analysis that is informed by working hypotheses are required;
- **Relevant** – the statistical method(s) might be targeted to addressing specific issues but those issues may not be relevant to the derivation and validation of an index *in the current context and setting* (a Type III error);
- **fit-for-purpose** - ensure any statistical methods used represent the best tool for the job. Avoid compromise approaches and the application of ‘standard’ or ‘text-book’ statistical methods. Most environmental data violate the intrinsic assumptions underlying the legitimate use of a particular technique. These violations include: non-normality; heteroscedasicity; over and under dispersion; spatial and autocorrelation.
- **address key management issues** – ensure that the program of statistical analysis is driven by a requirement to address specific technical matters that are ultimately linked to the implementation of the index and management issues. Unless specifically part of a research project designed to advance the state of knowledge about index construction and validation, avoid curiosity driven research.
- **scientifically credible** – the development of indices (environmental and otherwise) has become somewhat of a cottage industry resulting in a plethora of metrics and approaches – not all of which have passed scientific scrutiny. Indices used to measure, monitor, and manage the environment should be peer-reviewed before adoption.
- **statistically defensible** – An index might be scientifically credible to the extent that it has the *potential* to do what it claims, but unless its statistical properties are understood (for example is the signal-to-noise ratio sufficiently high that it can differentiate between background variation and a putative impact?) then attaching *significance* (statistical or otherwise) to resulting values will be problematic.

Statistical methods have played a crucial role in index development. A review of 824 published papers by Whittaker et al. (2012) found the most popular techniques used in the development of environmental indices included (in descending order of frequency of usage):

- principal component analysis,
- cluster analysis;
- canonical correspondence analysis / correspondence analysis;
- Analysis of Variance;
- Artificial Neural Networks;
- Fuzzy sets;
- Factor Analysis;
- Discriminant analysis;
- SARIMA (seasonal autoregressive integrated moving average model);
- multiple regression;
- MRPP (Multiresponse Permutation Procedure);
- MDS (Multidimensional scaling); and
- MCA (Multicriteria analysis).

Their review concluded with a salutary warning that *"generally speaking, it can be expected that the calculation of a WQI using a statistical approach will provide a poor WQI if the statistical properties of water quality constituents do not happen to coincide with a knowledgeable evaluation of importance"* (emphasis added). We believe this is a fundamental issue and one that has not been fully addressed in the rush to publish and promote scorecard evaluations and results.

2.5.1 Candidate metrics

As alluded to in the previous section, considerable effort has focussed on metric development and relatively little on metric refinement. Borja at al. (2009) observed that "as the concept of indices has gained acceptance, there has been a proliferation of index approaches" suggesting that what is now required is "to unify approaches that provide managers with the simple answers they need to use ecological condition information effectively and efficiently". Their strategy for achieving this relied on:

- Reducing the array of indices by identifying the index approaches that are most widely successful;
- Establishing minimum criteria for index validation;
- undertaking comparative calibration experiments to achieve uniform assessment scales across geographies and habitats; and
- integrating indices across ecosystem elements.

The remainder of this section provides details of some commonly used indices for assessing ecosystem health.

Weighted sub-index methods

In the context of water quality, Harbans (2011) outlined the following index development process based on a weighting of individual parameters (sub-indices):

1. Identify water quality parameters of interest and their ranges of acceptability for the intended uses of the water body;
2. Compare the measured value with the subjective rating curve and arriving at a dimensionless sub index value (0-1) for each parameter;
3. Define the weighing factor and/or heuristics for each parameter to be considered while building an overall WQI;
4. Select an algorithm and computing the WQI with the available data and assumptions.

The final water quality index (*WQI*) may be represented by the generic formula given by equation 1.

$$WQI_p(x_1,\ldots,x_n)=\begin{cases}\left(\sum_{i=1}^{n} w_i\, x_i^p\right)^{\frac{1}{p}} & p\neq 0\\ \prod_{i=1}^{n} x_i^{w_i} & p=0\end{cases} \tag{1}$$

where the x_i are the sub-index values; the w_i are a set of positive weights that sum to unity; and *p* is an exponent.

Baseline comparative methods

Indices are developed by comparing observations to benchmark values rather than normalising them as in the weighted sub-index approach. Selection of benchmark values is arbitrary yet clearly influential in this process. This method has been used widely and is the basis for the Canadian Water Quality Index (CWQI), the United Nations Environment Program (UNEP) Global Environmental Monitoring System (GEMS), the Fitzroy Partnership's report card and SEQ Healthy waterways report card.

We outline the computational procedure for the Canadian and local approaches.

CCME WQI

The Canadian Council of Ministers of the Environment (CCME) WQI is an objective-based index that compares measured water quality with guideline values using the concepts of *scope* (percentage of indicators not meeting the relevant water quality objective), *frequency* (percentage of comparisons where the guideline was not met), and *amplitude* (a normalised measure of the extent to which failed comparisons deviated from the guideline). The computational formula is given by equation 2.

$$WQI^{(CCME)}=100-\frac{\sqrt{F_1^2+F_2^2+F_3^2}}{1.732} \tag{2}$$

where *F₁* is the scope; *F₂* the frequency; and *F₃* the amplitude. Computational formulae for *F₁*, *F₂*, and *F₃* are given by equations 3(a), 3(b), and 3(c) respectively.

$$F_1 = 100 \cdot \left(\frac{\text{number of failed indicators}}{\text{total number of indicators}} \right) \tag{3a}$$

$$F_2 = 100 \cdot \left(\frac{\text{number of failed comparisons}}{\text{total number of comparisons}} \right) \tag{3b}$$

$$F_3 = \frac{100 \cdot E}{1 + E} \tag{3c}$$

where in equation (3c) $E = \frac{\sum_{i=i}^{k} e_i}{k}$; *k* is the total number of comparisons; and e_i is:

$e_i = z_i \cdot \left[\left(\frac{x_i}{O_i} \right)^{-\lambda_i} - 1 \right]$; with *O_i* the *i*th objective value; *x_i* the *i*th comparison result; and

$$z_i = \begin{cases} 1 & \text{if } i^{th} \text{ comparison results in fail} \\ 0 & \text{otherwise} \end{cases} \quad ; \; \lambda_i = \begin{cases} 1 & \text{where } i^{th} \text{ comparison fails if } x_i < O_i \\ 0 & \text{where } i^{th} \text{ comparison fails if } x_i > O_i \end{cases}$$

Hurley et al. (2012) investigated a weighted version of equation 2 and concluded this provided no additional benefits although did recommend modifications to *F₂* (equation 3b) and *F₃* (equation 3c) to help overcome bias introduced by differing number of comparisons for different indicators.

Conceptually, equation 2 is a measure of distance in an imaginary ‘objective exceedence’ space (Figure 4) which results in an index which is 0 or close to 0 for very poor water quality, and close to 100 for excellent water quality (Canadian Council of Ministers of the Environment 2001). The narrative for the CCME WQI is given in Box 3.

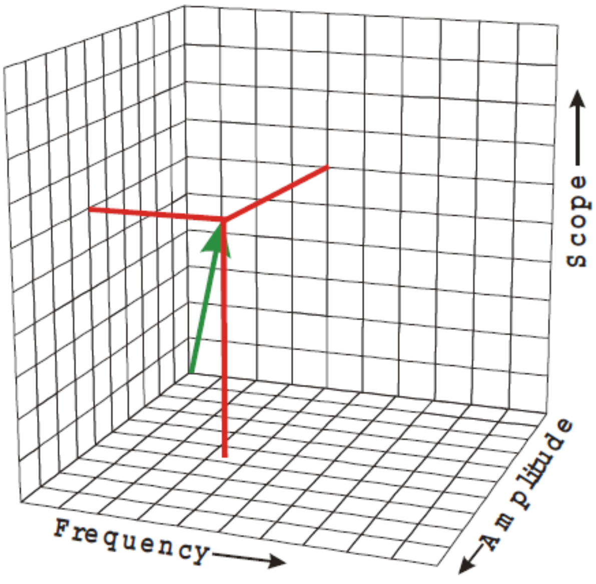


Figure 4. Representation of CCME WQI in three-dimensional exceedence space. *(from Canadian Council of Ministers of the Environment 2001).*

Excellent: (CCME WQI Value 95-100) – water quality is protected with a virtual absence of threat or impairment; conditions very close to natural or pristine levels. These index values can only be obtained if all measurements are within objectives virtually all of the time.

Good: (CCME WQI Value 80-94) – water quality is protected with only a minor degree of threat or impairment; conditions rarely depart from natural or desirable levels.

Fair: (CCME WQI Value 65-79) – water quality is usually protected but occasionally threatened or impaired; conditions sometimes depart from natural or desirable levels.

Marginal: (CCME WQI Value 45-64) – water quality is frequently threatened or impaired; conditions often depart from natural or desirable levels.

Poor: (CCME WQI Value 0-44) – water quality is almost always threatened or impaired; conditions usually depart from natural or desirable levels.

Box 3. Interpretation of the CCME WQI *(from Canadian Council of Ministers of the Environment 2001).*

We note that one of the advantages of the CCME WQI is that it is 'not tripped up' by below detection limit readings for concentration data since such readings will not constitute a 'failure' (assuming the limit of detection is always < objective value).

Local methods

The Fitzroy Partnership and the South East Queensland Healthy Waterways are among a number of agencies that compute indices on the basis of a comparison with a benchmark / guideline. The method (for a contaminant concentration) is summarised by equation 4.

$$index_i = \begin{cases} 100 & \text{if } x_i \le benchmark_i \\ 0 & \text{if } x_i \ \ge \ WCS_i \\ \left[1.0 - \left| \dfrac{x_i - benchmark_i}{WCS_i - benchmark_i} \right| \right] \cdot 100 & \text{otherwise} \end{cases} \tag{4}$$

where x_i is as above; *benchmark*$_i$ is an ecosystem health guideline value; *WCS*$_i$ is a value of x_i at which ecosystem health would be compromised ('worst case scenario').

According to Jones et al. (2013) benchmark (or guideline) values for the EHMP are based on either the 20th percentile (if objective is to be *above* a target – such as dissolved oxygen) or 80th percentile (if objective is to be *below* a target – as is usually the case with phys/chem concentrations) computed for minimally disturbed reference sites. Worst case scenario values are derived from either the 10th or 90th percentile from *all* sites.

The Fitzroy EHI uses equation 4 as well although the determination of *benchmark/guideline* and *WCS* values appears to be more subjective with the former being one of {water quality objective; ecosystem health guideline; trigger value; expert opinion} and the latter simply "the value of x_i at which ecosystem health may be compromised" (Jones et al. 2013).

Finally, we are aware of other local variants to the above scheme (such as the Reef Rescue Monitoring Program) but have not included them in this report due to insufficient computational detail and/or 'interim' status. For example, Schaffelke et al. (2011) note that the inshore water quality index developed by the Australian Institute of Marine Science (AIMS) is an interim metric "as further research and data analysis need to be undertaken". We also note that the Fitzroy EHI proposed by Central Queensland University (Jones et al. 2013) is pending final review and endorsement by the Science Panel.

2.5.2 Benchmarks and Guideline Values

The locally popular baseline comparative methods of the previous section require the specification of a 'benchmark' or 'guideline' value for every index. Given this quantity plays a pivotal role in the numerical value of the computed index and subsequent report card classification scheme, its quantification should be the subject of a separate study. As noted by Hijuelos and Reed (2013):

> "comparison to the baseline can only be made meaningful if the desired direction of change is well understood. Setting of targets may consider the expected effects of restoration or protection projects that have been or will be implemented such that the target represents an expected post-construction system state. Targets should be specific to the reporting region and be scientifically justified. Validation procedures to determine the robustness of the performance measures and the scoring thresholds should be employed, particularly when odelling is involved. These typically require separate validation datasets that are often unavailable".

It is important that changes relative to a benchmark are meaningful *and* can be measured with a degree of precision that is commensurate with the 'distance' represented by the term $\left(WCS_i - benchmark_i\right)$ appearing in equation 4. For example, if the measurement error in individual x_i values is of the same order as the difference $\left(WCS_i - benchmark_i\right)$ then the comparison is rendered ineffectual.

In developing the Fitzroy EHI, Jones et al. (2013) mentioned the issue of benchmark (also referred to as a *reference threshold*) selection but provided no details as to how this was to be implemented although examination of spreadsheets from the Fitzroy Partnership web site reveals that in many instances the 80th and 90th percentiles of empirical data have been used for the benchmark and WCS respectively.

Another aspect of guideline selection that requires careful consideration is what we have termed *'protection harmonisation'*. Guidelines for different indicators will invariably reflect different levels of protection and beneficial use and the simple aggregation of the resulting indices may not result in a meaningful assessment of overall ecosystem health. Such considerations may provide support for the adoption of an unequal weighting system (see next section).

In their recent review, Connolly et al. (2013) strongly supported the 'distance from a guideline' approach although advocated the need to adopt locally-relevant guidelines in preference to regional or nationally-derived values. We support this view, although suggest additional investigations be carried out to investigate (a) potential advantages in using the Canadian concept of an 'exceedence space' rather than a one-dimensional comparison; and (b) the impact on computed indices resulting from the interplay between: choice of guideline value; choice of worst case scenario value; and choice of computational method (including different weighting schemes – see next section). These investigations should countenance both 'undisturbed' and 'impacted' systems to better characterise and understand the performance characteristics of constructed indices.

Before moving on to the issue of indicator aggregation and weighting, we digress momentarily to reflect on current 'benchmarking' practices.

Assessing condition relative to a benchmark – are we doing it right?

This is a rhetorical question – the short answer is that no one really knows. However, the following 'quasi-real' example provides cause to reassess conventional wisdom. We call it quasi-real because the numbers used have been artificially generated by a process that is informed by and reflects the real environment using published results for the Fitzroy Basin.

Implicit in the application of equation 4 is that both the benchmark and worst case scenario figures are static 'lines in the sand' that are free of uncertainty. The only time this is true is when *aspirational* benchmarks (discussed in section 3.4.1) are used. For *empirical* and *modelled* benchmarks, there is an associated *distribution* from which the numerical values for the benchmark and WCS are derived. This distinction is illustrated in Figure 5.

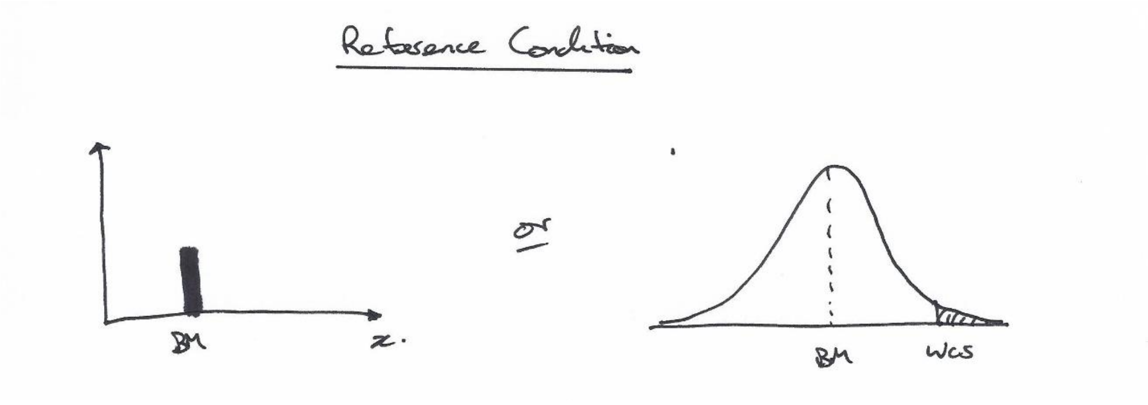


Figure 5. Two views of a 'benchmark' – as a static number that has no uncertainty (left) or a *statistic* derived from an empirical distribution associated with a reference condition (right).

If we accept that both *reference* and *test-site* conditions are variable and hence there is always a *distribution* of results and not just a single number, then a range of possibilities defining 'change' in condition is apparent. The current method of assessing 'change' using equation 4 essentially only contemplates gross changes in 'location' – that is a wholesale shift of the indicator's distribution to the left or right. While this is no doubt important, the method fails to acknowledge other types of distributional change – for example changes is *dispersion* i.e variability of the response. Figure 6 illustrates the effect of changes in location and dispersion for a he effect of changes in location and dispersion for a *symmetrical* distribution while Figure 7 illustrates a change in an *asymmetrical* distribution. Given that nearly all water quality and many other ecological indicators and metrics have asymmetrical distributions, it is of interest to explore Figure 7 in more detail since this lack of symmetry makes it more difficult to assess the 'significance' of change.

By way of example, we have examined published results for the Fitzroy EIH Program[1] . Based on the summary statistics provided, we have constructed appropriate distributions that have similar properties. Figure 8 shows our inferred distributions at Fitzroy Catchment Site 1751. It is not at all clear from an inspection of Figure 8 whether or not the 2011 results are *significantly* worse than reference condition.

[1] http://riverhealth.org.au/report_card/ehi/Fitzroy/Fitzroy%20PhysChem%20Data.xls

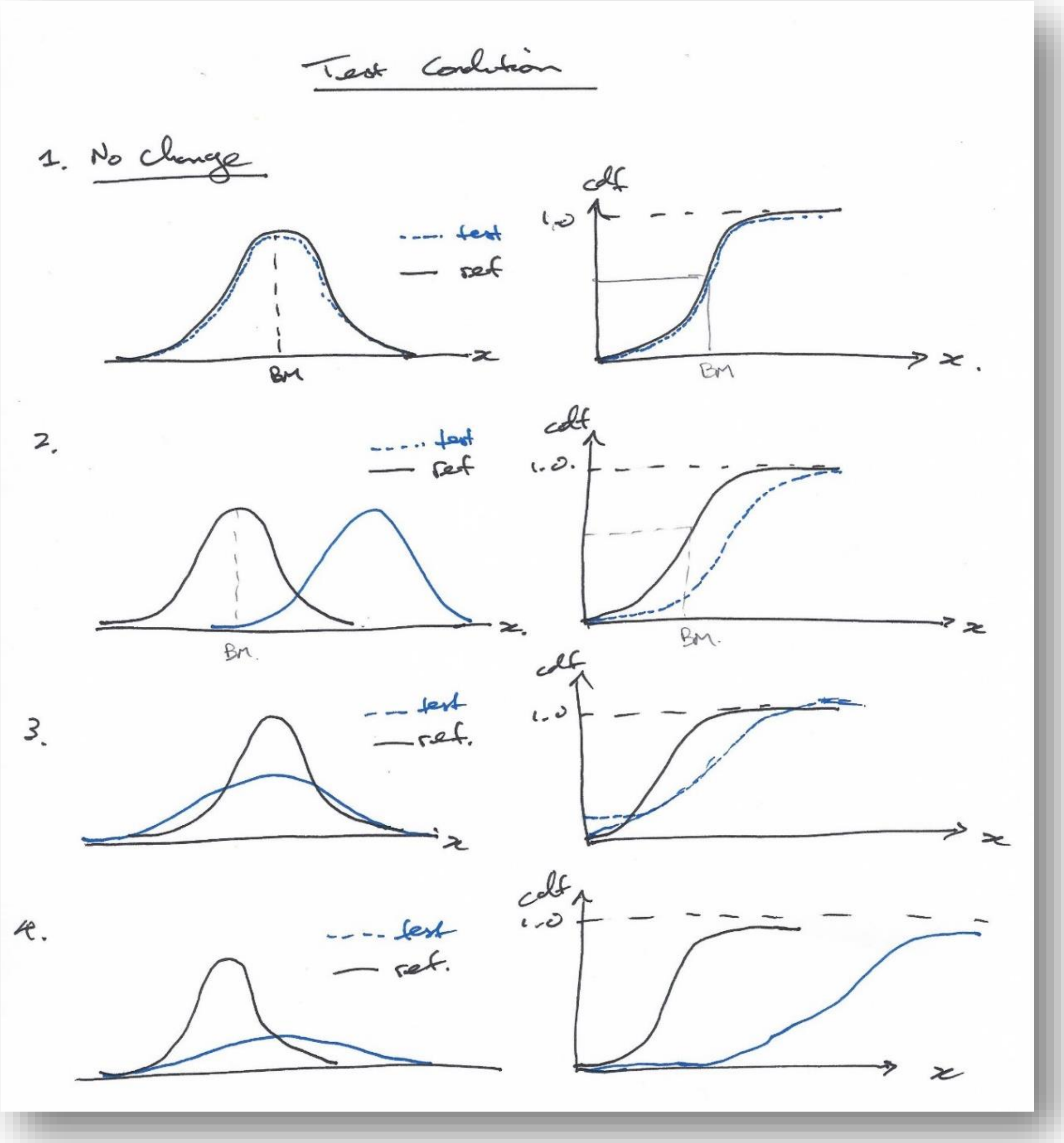


Figure 6. Examples of distributional changes in location and scale. Probability density function (*pdf*) on left; cumulative distribution function (*cdf*) on right.

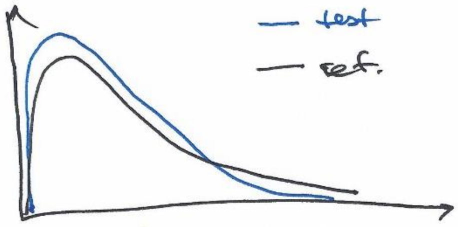


Figure 7. Hypothetical distributions of an indicator / parameter at test and reference locations.

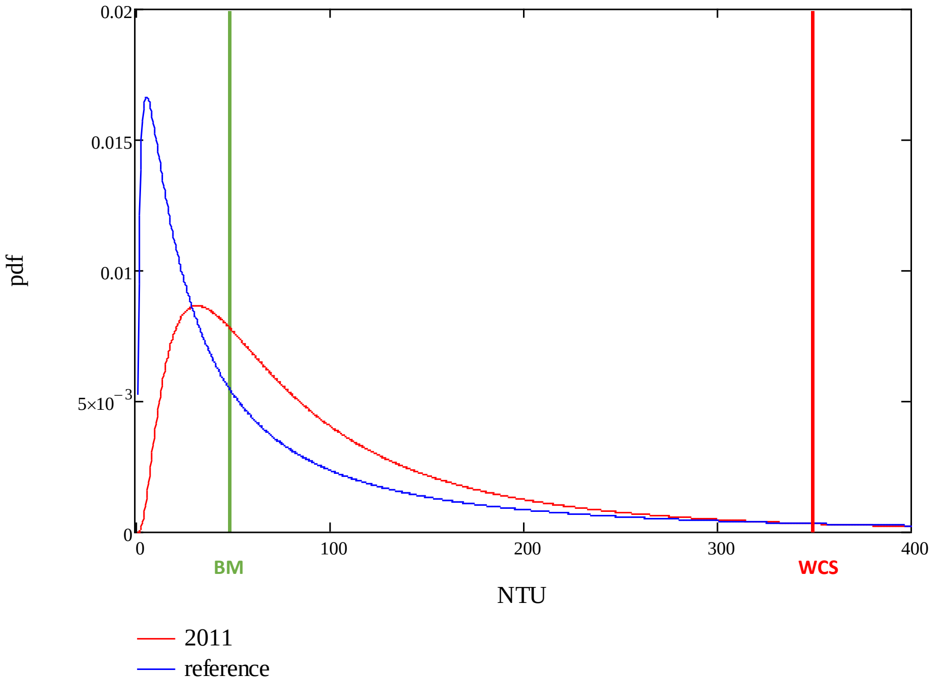


Figure 8. Inferred theoretical distributions for reference and 2011 turbidity results at Fitzroy Catchment Site 1751. Benchmark turbidity (green line) is 50 NTU and Worst Case Scenario value (red line) is 350 NTU.

An examination of common measures of location for the two distributions in Figure 8 does not help either since the *median* turbidity of 75 NTU in 2011 represents an increase while the *mean* has decreased from 158 NTU (reference) to 119 NTU (2011 result). An examination of the two cumulative distribution functions reveals the nature of the dilemma (Figure 9). The two *cdfs* cross over at the 77th. percentile. Percentiles *below* P_{77} are numerically higher for the 2011 data while percentiles *above* P_{77} are numerically smaller for the 2011 data. Thus our assessment of whether 2011 is “better” or “worse” than the reference distribution is ambiguous – it depends on how we wish to interpret the results and by what measure. For example, if we are concerned about *chronic* effects and sediment load to the system, then a comparison of means is probably most appropriate. However, if we are more concerned about acute impacts associated with elevated TSS concentrations, then it could be argued that 2011 represents a better outcome since the extreme turbidities are not as high as under the reference condition.

To compound this indeterminacy, there is the related issue of *grading* this site’s turbidity results. If we adopt the current methodology used by the Fitzroy Partnership (which is the same as the EHMP and many other agencies) a grade is assigned on the basis of a comparison of the average result with the relevant benchmark (equation 4). The (theoretical) average for site 1751 in 2011 is 119 NTU which results in a score of 23 which then converts to a “D” (i.e “Poor”). Alternatively, if we score each observation separately and then average the scores we obtain a result of 75 which then converts to a “B” (i.e “Good”). This simple example has served to highlight a fundamental, and unresolved issue with the current scoring system – namely, that it is a blunt instrument that is incapable of resolving anything other than gross changes in condition. We very much suspect that this ‘inertia’ is compounded as a result of aggregation and averaging over sites, times and sub-regions.

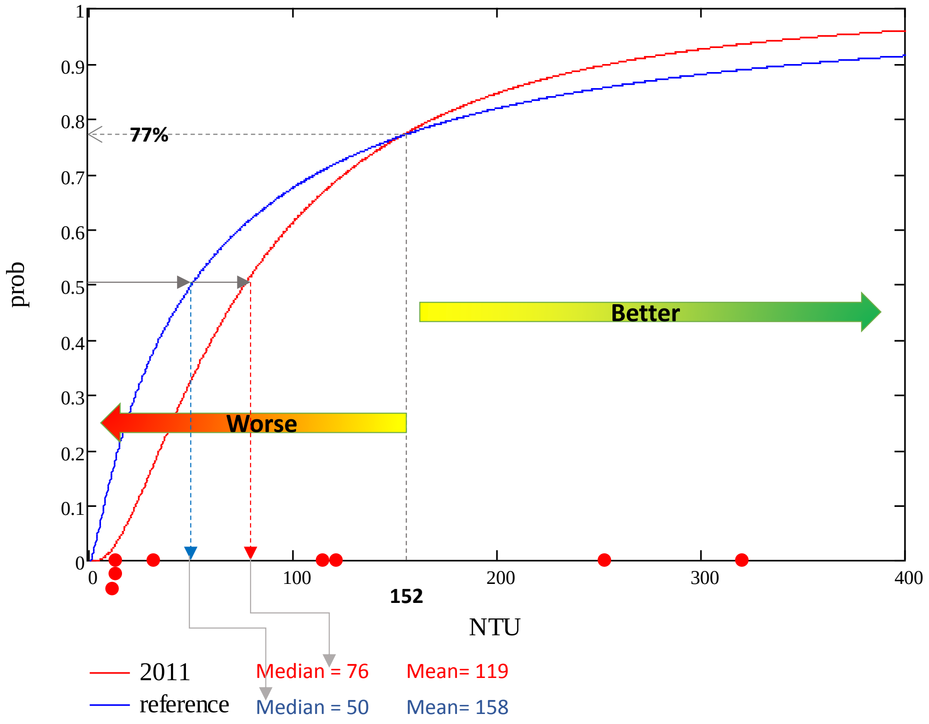


Figure 9. Cumulative distribution functions for the distributions in Figure 8. The percentiles below the 77th. are numerically *greater* (i.e. worse) for the 2011 data while the percentiles above the 77th. are numerically *smaller* (i.e. better) for the 2011 data. Individual data points indicated by solid red circles.

2.5.3 Combining multiple indices and weighting

The method by which a number of indices are combined into a single metric is a key aspect of the overall report card methodology with the choice "likely to have a strong impact on the final scores and their sensitivity to changes" (Connolly et al. 2013). The possibility that different weighting schemes may lead to different assessments of ecosystem health for the Fitzroy Basin was also noted by Jones et al (2013b) and they concluded that further research into this area was required. Accordingly, the Fitzroy EHI utilises an equal weighting of all indices.

The weighting issue is pervasive and, to our knowledge, no clear advice has emerged despite numerous studies in which alternative approaches have been evaluated (Emerson et al 2012). Borja and Dauer (2008) claimed that the "most difficult challenge in index development is selecting and combining metrics in a manner that is complex enough to capture the dynamics of essential ecological processes but not so complex that its meaning is obscured".

A number of candidate weighting strategies are available and some of these have been listed in Table 1 together with advantages and disadvantages.

Table 1. Advantages and disadvantages of a select number of weighting strategies (*adapted from Williams et al. 2010)*

Approach	Advantages	Disadvantages
Equal weighting (index score is average of all indicators)	simple to understand and communicate; do not have to justify weighting rationale	Assumes all indicators are of equal importance
Geometric mean (weight towards lowest score)	Penalises more imbalanced scores; the more imbalance, the lower the score	more difficult to interpret and communicate
weight according to importance to overall health and/or objectives	if done correctly, should provide a more accurate assessment than other weighting schemes	element of subjectivity in weighting scheme introduces unquantifiable bias
weight proportional to precision of component scores	index scores that have low confidence are down-weighted	precision can be difficult or impossible to quantify
use only the worst score	simple to understand and communicate	wasteful of all remaining information – other scores merely serve as 'place holders; uncertainty in extreme values will be larger than an aggregated measure.

2.6 Validation procedures

An independent validation of the index methodology is crucial if it is to gain wide acceptance as a useful and meaningful tool to measure, monitor and manage overall ecosystem health. The fundamental questions to be answered by a validation study are:

- Do the results make sense when applied to a wide variety of situations, places, and circumstances?
- Does the aggregation process produce a classification that accords with expert opinion?
- Does the index have good signal-to-noise properties?
- Do statistically significant changes in the index correspond to ecologically significant changes?
- Is the direction and magnitude of trends in the index over space and time consistent with the direction and magnitude of spatial-temporal trends in observed ecosystem condition?

Clearly, the use of 'expert' opinion is a key component of the validation process, although expert opinion on the qualitative description of water quality can be variable (Neary 2012) which would necessitate the use of a relatively large panel of water quality experts and a properly designed validation experiment in order to partition and test components of variation in validation scores.

Another difficulty flagged by Whittaker et al. (2012) is the lack of guidance on how to assess the degree to which a water quality index is representative of the underlying data – the difficulty being that there is no 'observed' index against which to evaluate and compare different constructs for index calculation. In the absence of a 'true' index value, a common validation technique is to examine the correlation between the index and constituent variables. Simulation studies have also been used in this context (Feio et al. 2009).

Notwithstanding issues of expert bias and between-expert variation, we outline below a possible strategy for validating an index and/or assessing competing methods of index construction.

The idea is relatively straightforward: present 'raw' data or low-level summaries of the data used to construct the index to a group of *N* 'experts' and ask them to classify the results using the same scheme as that used to classify index values. The results are then summarised in a standard two-way contingency table (Figure 5).

	Expert classification				
Index score	Excellent	Good	Fair	Marginal	Poor
95-100	F_{11}	F_{12}	F_{13}	F_{14}	F_{14}
80-94	F_{21}	F_{22}	F_{23}	F_{24}	F_{25}
65-79	F_{31}	F_{32}	F_{33}	F_{34}	F_{35}
45-64	F_{41}	F_{42}	F_{43}	F_{44}	F_{45}
0-44	F_{51}	F_{52}	F_{53}	F_{54}	F_{55}

Figure 10. Two-way contingency table for presenting results of a validation experiment for a single index. Cell entries are frequencies (counts).

Conventional contingency table analysis tools will enable inference to be drawn about the degree of association between the index score and the experts' classification. Note that in the table of Figure 5 $\sum_{i=1}^{5}\sum_{j=1}^{5}F_{ij}=N$. As a rule of thumb, for this design, *N* would have to be in excess of 100. Finding and engaging more than 100 experts may be problematic in which case the number of categories would have to be collapsed. This design can be extended to undertake more comprehensive assessments. For example, an analysis of a four-way contingency table using a *multinomial* ANOVA model would allow an assessment of multiple indices in a number of sub-regions. In this case (and using the same experts throughout) we have $\sum_{i=1}^{5}\sum_{j=1}^{5}\sum_{k=1}^{s}\sum_{l=1}^{r}F_{ijkl}=rsN$ where F_{ijkl} is the frequency observed for the i^{th} index category and j^{th} expert category for sub-region *k* and index *l*. This arrangement allows for quite sophisticated hypotheses to be tested, for example that the 2-way association between index score and expert score is consistent for different indices and/or across sub-regions. This type of analysis is non-standard and further advice and assistance would be required to construct a validation experiment using this approach.

2.7 Recommendations

With respect to indicator / index development we recommend the following:

5. A staged approach such as that used for the development of the SEQ WQMS be adopted;

6. Undertake targeted investigations using *existing data* to investigate and assess:
   - the merits of various computational methods such as the CCME WQI method; and
   - the implications of equal and unequal weighting schemes.
7. Undertake validation study using group of experts and methodology provided here (or suitable alternative);
8. Using *validated* indices, apply to existing data to quantify spatial correlation structure and temporal variation.

3. Report Cards

3.1 Rationale

The use of simple, visual and descriptive tools to summarise and assess numerical observations on numerous ecosystem indicators is commonplace in Australia and worldwide. While there has been considerable development associated with the computational processes and aggregation procedures, relatively less effort has been devoted to validation and performance assessment (Borja et al. 2009). A review of the Chesapeake Bay monitoring program noted the potential for drawing contradictory conclusions when assessments were based on different indicators (U.S. Government Accountability Office 2011). The underlying assumption of the evidence-based policy agenda that research, statistics and indicators can lead to policies that will work better, on the assumption that 'scientific' information could guide social affairs has also been brought into question (Herzi and Dovers 2009).

Thus one of the most significant challenges in report card development is to eliminate ambiguity through aggregation but not to over-smooth whereby the resulting score has low signal-to-noise properties and high inertia to change.

Whatever approach is adopted by the GHHP we believe that the validation component should be given high priority to ensure the report card scores are meaningful and accord with expert (subjective) assessment. This is a view supported by Connolly et al. (2013).

3.2 Objectives

As suggested by Dennison et al. (2013), the purpose of a report card is to **integrate** (monitoring data); **engage** (stakeholders) and **catalyse** (actions). To be *effective* the report card should:

- Use carefully constructed indices based on a select list of indicators to evaluate the status of the system;
- Have the ability to assess long- and short-term trends in ecosystem condition based on a validated aggregation and scoring methods;
- Provide transparency in the methodologies used to produce the grades;
- Communicate the results in a way that is both meaningful and understandable to multiple audiences.

(Adapted from Hijuelos and Reed 2013).

3.3 Criteria for report card development

The development of a reporting framework requires the derivation of highly aggregated 'scores' that reflect essential attributes or dynamics of the system that can be used to track changes over time and support decision-making. These scores should be based on the following:

- Relevant to ecologically important functions or processes;
- Sources of spatial-temporal variation quantified, correctly utilised and interpreted;
- Sufficient statistical power to detect trends in both time and space on scales that are relevant and meaningful;
- Human, financial and capital resources available to implement performance assessments in cost-effective and timely manner;
- Useful for management decisions and program refinement.

(Adapted from Hijuelos and Reed 2013).

3.4 Methodology

A simplified schematic of the steps involved in the construction of a report card is shown in Figure 6.

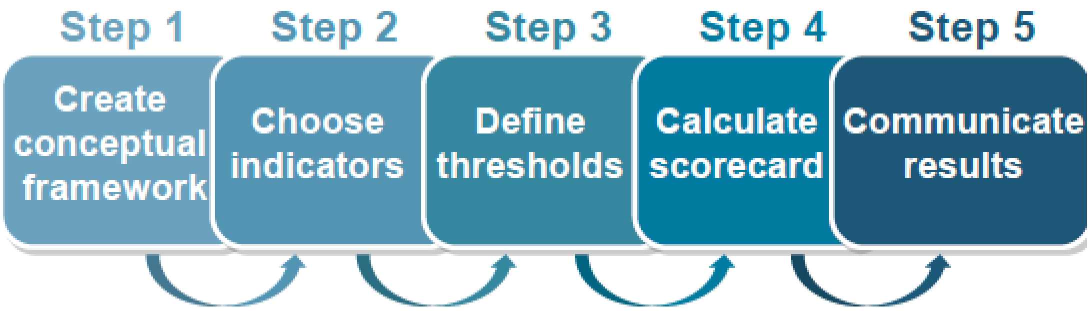


Figure 11. Steps in report card development. ***(Source: Dennison et al. 2013).***

We understand that the GHHP has completed step 1 and is in the process of finalising step 2. Issues associated with defining thresholds / benchmarks / guidelines have been discussed in section 2.5.2. Step 4 remains as a significant challenge for which we believe there are two broad strategies. These are outlined in Table 2 together with an assessment of the advantages and disadvantages of each. Irrespective of which approach is adopted by the GHHP, careful consideration needs to be given to: (i) the computation of indices; (ii) the aggregation of indices into report card scores; and (iii) the translation of a numeric score to an ecosystem classification – including a taxonomy for this classification (Figure 7). These are discussed further in the following section.

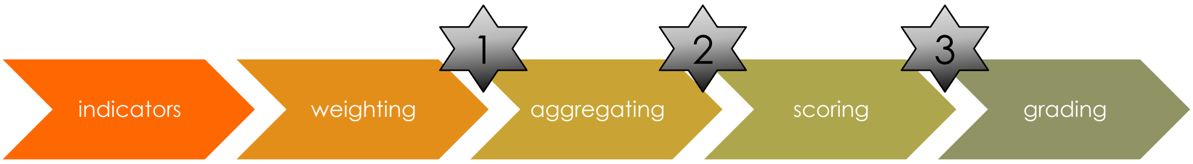


Figure 12. Scorecard grading process. Numbered stars indicate stages where computational/statistical procedures need to be developed.

Table 2. Two broad strategies for report card development.

Strategy	Advantages	Disadvantages
'cherry-pick' - from other projects around Australia and around the world based on what you think will work best for Gladstone.	- cost effective; - short(er) lead time; - low risk	- known problems are inherited; - combination of 'best bits' of others may be sub-optimal for Gladstone
'roll your own' - develop program based on current best practice, recognising that some of the information-gaps need to be plugged by targeted R&D	- tailor-made therefore fit-for-purpose; - more tightly coupled and better integration; - moves you up the innovation scale -> increased recognition	- R&D component may 'weigh you down' - both financially and with implementation; - no guarantee of superior outcome; - high(er) risk

3.4.1 Classification schemes and taxonomies

The identification of a grading taxonomy for a report card is rather arbitrary and is a function for the ISP. There are no hard and fast rules and experience elsewhere suggests this is essentially a deliberative process to reach consensus on the identification of breakpoints in the aggregated index that generate meaningful descriptors of ecosystem health. Consideration should also be given as to what constitutes reasonable / unreasonable progress even if a target is not achieved (Hijuelos and Reed 2013). The validation procedure outlined in section 2.6 could be useful in assessing the performance of candidate report card scoring methodologies.

To assist the ISP in its deliberations on report card scoring options, we believe it is important to make some distinctions: firstly – is the resultant classification intended to be *relative* or *absolute*? In other words, does the report card label simply rank sub-regions relative ideal conditions for Port Curtis or do the report card labels have relevancy in other contexts / jurisdictions / ecosystems (Figure 8)?

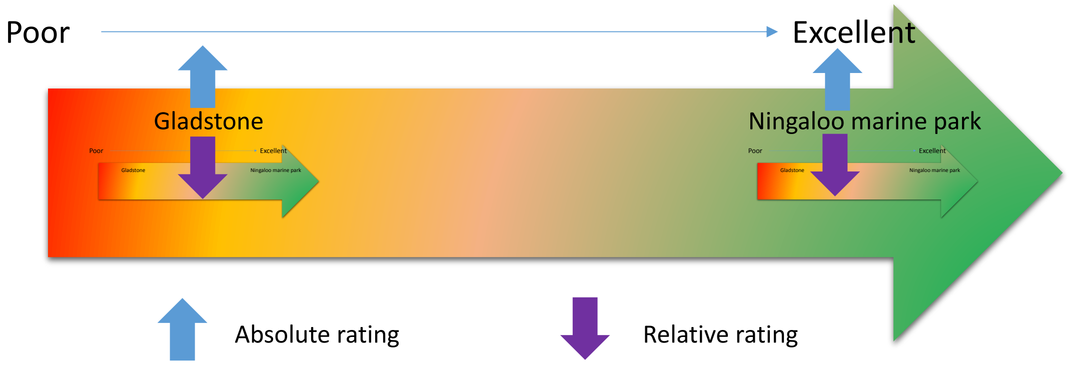


Figure 13. Illustration of absolute and relative ratings. Absolute ratings are meaningful in other settings and can be compared across different environments whereas relative ratings rate sub-regions relative to each other.

Secondly, if the indices underpinning the score card evaluation use a baseline comparative method (see section 2.5.1) then the frequency with which the threshold or benchmark is attained or exceeded will depend on whether the threshold/benchmark is *aspirational*, *empirical*, or *modelled*. An aspirational thresholds is set as a target to achieve for which a classification of 'excellent' for example simply means *progress* towards the target has been 'excellent' – the overall ecosystem status may be less than 'excellent'. Empirical targets for water quality are the most common in Australia since the National water Quality Guidelines (2000) recommend comparison of test site data to percentiles of reference site data. Using this type of threshold allows to make statements about the *frequency* with which we would expect a threshold to be exceeded in the absence of any impacts. For example, if the benchmark for a water quality indicator is set as the 80th percentile of background data, then we expect this indicator to fail 20% of the time even when the test site is no different to the reference site. Finally, in some cases we might have good models having good predictive capability for some parameters. In this case it would be possible to have a target which is based on modelled conditions. An example of this is turbidity in the Western Basin for which sophisticated statistical models have been developed that can predict *background* turbidity under actual or assumed wind, rain, and tidal conditions. There is merit in computing an index of turbidity that takes into account the prevailing exogenous factors rather than one based on a static threshold that has been derived from data over a fixed period. An example of model-based assessments is the aquatic fauna performance measure used in the Everglades Report Card (Hijuelos and Reed 2013).

Options for scoring

Step 3 in Figure 7 requires the specification of a formula, process, or table for assigning grade to an aggregated score. The first decision therefore concerns the choice grades. A number of options exist:

- **Binary** label: "pass/fail"; "improved/declined";
- **Ordinal** label: letters ("A" to "F"); words ("poor" to "excellent")
- **Ratio** (number): e.g. percentage of indices above benchmark;
- **Interval** (number): e.g scale of 0-10.

The difference between ratio and interval numeric grades is that for *ratio* type grades statements of the kind “site A is twice as good/bad as site B” whereas for *interval* type grades only *differences* between grades (and not ratios of grades) are meaningful.

A difficulty with labels is their ‘granularity’ – that is, individual labels span a large range of indicator values (Connolly et al. 2013). This granularity can introduce ‘inertia’ since it may take a large shift in many indicators to alter the grade.

We are unaware of any suggested methods for: (i) deciding which grading option is most suitable; and (ii) ‘optimally’ converting a score to a grade. We suggest that this aspect of the report card development form part of the validation process.

3.5 Recommendations

With respect to report card development we recommend the following:

3. As a matter of priority, undertake a project to re-evaluate index aggregation and scoring methodologies in current use (eg. Fitzroy Basin and EHMP)

    - Investigate alternatives / modifications that better deal with distributional changes in indicators other than gross shift in *location* (eg. mean);

4. Undertake validation study using group of experts to:

    (c) Assist in the development of a ‘formula’ to convert the aggregated and (possibly) weighted indices to a suitable report card grade;

    (d) ‘Road-test’ this formula by applying to existing data to establish that the resulting grades: accord well with expert assessment; reflect meaningful changes in ecosystem status; and adequately reflect differences between sub-regions.

This page intentionally blank

4. Monitoring

The development of an on-going, integrated, cost-effective monitoring program to support the data and information requirements of ecosystem reporting is clearly a priority for the GHHP. We understand that advice will be needed at the *design* stage of monitoring program development – particularly as it relates to more complex issues of:

- Where / when / what to sample and sample sizes;
- Statistical power;
- Identification of sub-regions;
- Allocation of resources including striking a balance between replication and increased spatial-temporal coverage;
- Avoiding sampling redundancy through an understanding of space-time correlation structures;

At this stage however, the provision of detailed advice and specific recommendations associated with the list above is not possible since these matters can only be resolved: (a) via an iterative process involving a multi-disciplinary team and the integration of *non-scientific* issues such as capability, logistics, and cost; and (b) after existing data holdings have been analysed to *quantify* aspects such as the spatial correlation structure in measured indicators and the trialling of indicators over candidate sub-regions.

Much has already been written about the guiding principles, methodologies, and processes associated with monitoring program design (eg. Hedge et al. *undated*) and will not be repeated here.

What we can do however is highlight some of the pitfalls and recommend a broad strategy. With respect to the first of these it is interesting to note that a review of the United States Environmental Monitoring and Assessment Program (EMPAP) cited the following reasons for the failure of aquatic monitoring programs:

- The Objectives for monitoring are not clearly, precisely stated and understood;
- Monitoring measurement protocols, survey design, and statistical analysis become scientifically out-of-date;
- Monitoring results are not directly tied to management decision making;
- Results are not timely nor communicated to key audiences in terms they can understand.

Partnerships such as the GHHP are precisely that – a collection of individuals, groups, and organisations sharing a common interest but not necessarily identical goals and priorities. As stated on the GHHP website *"Gladstone Healthy Harbour Partnership (GHHP) is a forum to bring together parties (including community, industry, science, government, statutory bodies and management) to maintain, and where necessary, improve the health of Gladstone Harbour"*. Given the diversity of backgrounds and perspectives represented by the constituent members, it is critical that Partnership member have a shared view about the role of monitoring and understand their position in the 'data-information space'. To this end, Fox and Mann (2010) classified monitoring activities, the drivers for the activities, and an organisation's positioning on the data-knowledge continuum (Figure 14).

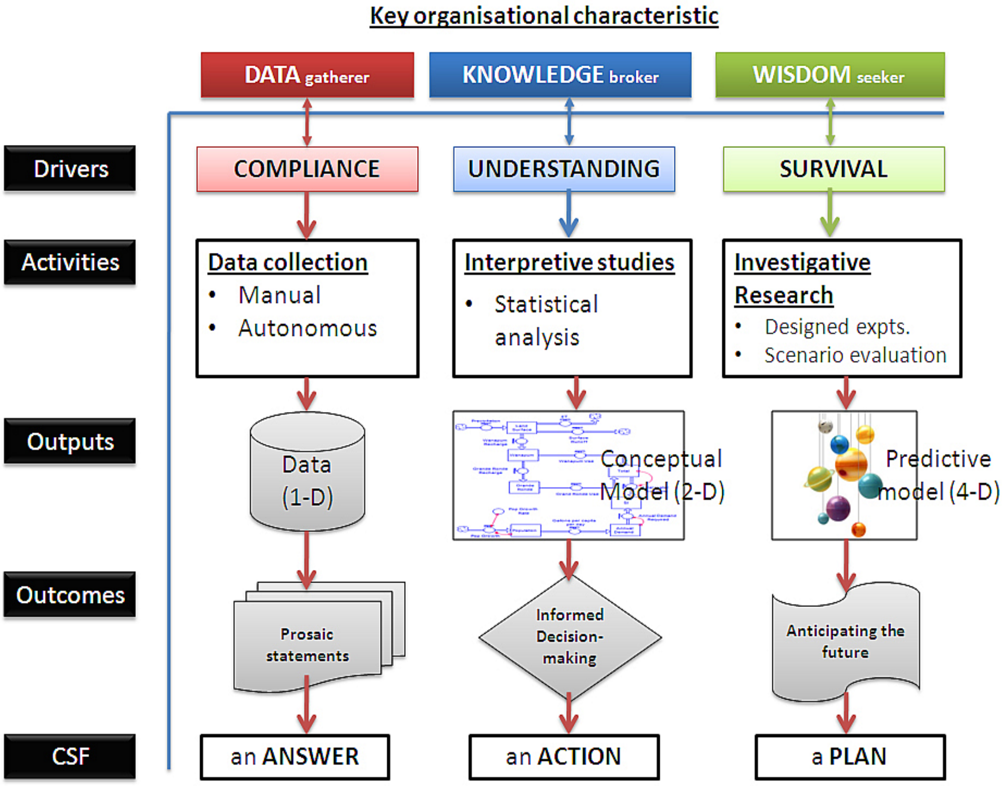


Figure 14. Taxonomy of organisational data – information gathering (CSF = critical success factor).

4.1 Design and sampling criteria

Again, this is an area where there is no shortage of existing advice, recommendations, and strategies. While individual programs differ in their sequencing of tasks, there is a high degree of similarity in terms of 'organisational structure'. For example, Figures 15, 16, and 17 show, respectively, the monitoring frameworks as recommended by the Australian Government (ANZECC/ARMCANZ 2000), for GBRMPA (Hedge et al. *undated*), and for the Adelaide Coastal Waters Study (Henderson et al. 2006).

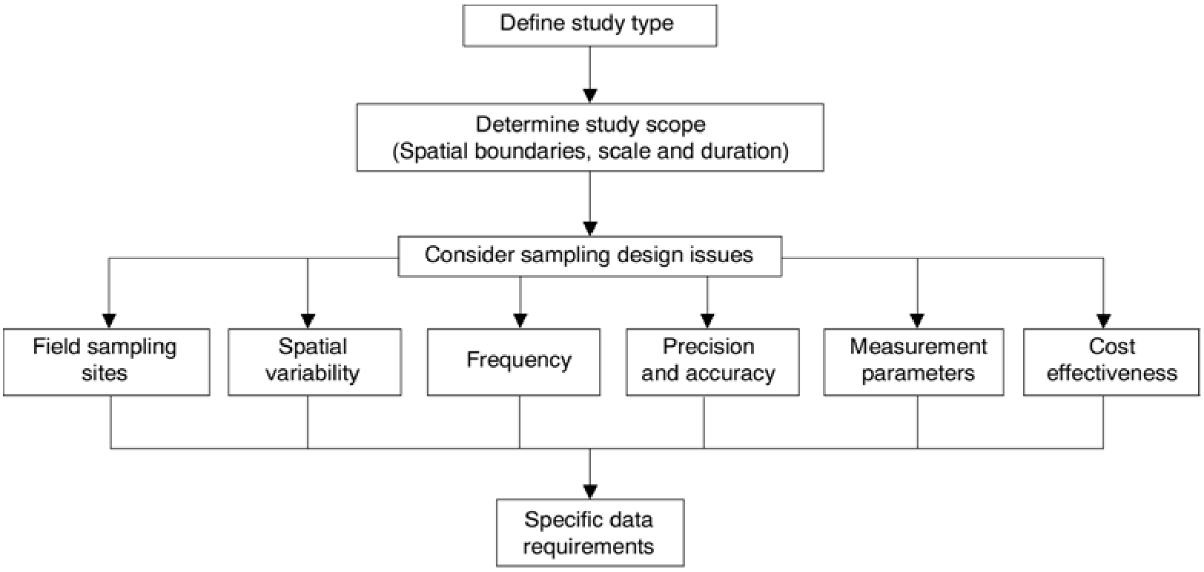


Figure 15. Framework for designing a monitoring program. (*Source*: Figure 3.1 ANZECC/ARMCANZ 2000).

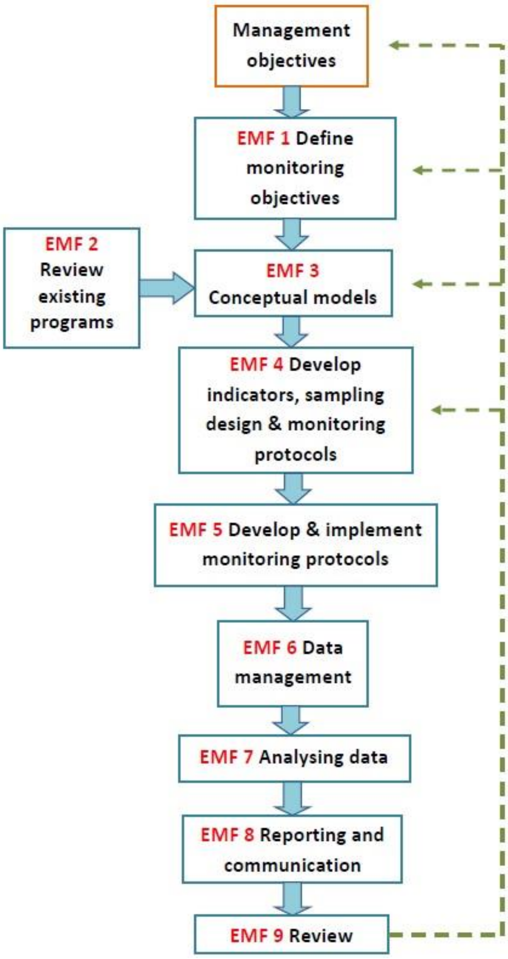

Management objectives
EMF 1 Define monitoring objectives
EMF 2 Review existing programs
EMF 3 Conceptual models
EMF 4 Develop indicators, sampling design & monitoring protocols
EMF 5 Develop & implement monitoring protocols
EMF 6 Data management
EMF 7 Analysing data
EMF 8 Reporting and communication
EMF 9 Review

Figure 16. Steps associated with the development of an integrated monitoring framework for the GBRWHA. (*Source: Hedge et al.* undated*)*.

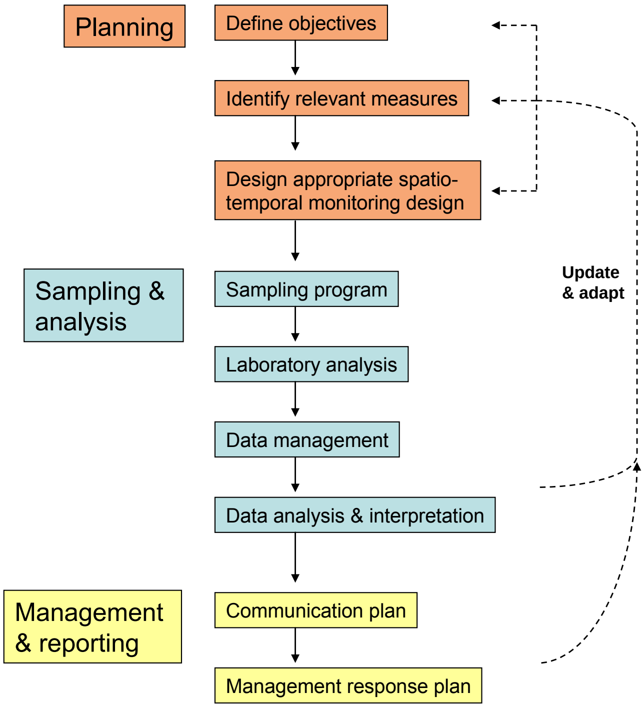


Figure 17. Integrated monitoring program design for the Adelaide Coastal Waters Study. (*Source:* Henderson et al. 2006).

4.2 Suggested strategy

With respect to monitoring program development we suggest the following:

4. Adopt a high-level framework such as that suggested in the National Water Quality Management Strategy (ANZECC/ARMCANZ 2000) to assist in the identification of sample design elements.

5. Use the process outlined here (Figure 18) (or similar) to ensure the logical sequencing of additional investigative and validation studies required to inform the monitoring program design.

6. Develop field sampling and data analysis protocols on the basis of final decisions associated with: index computation; sub-region identification; and report grading 'formula'.

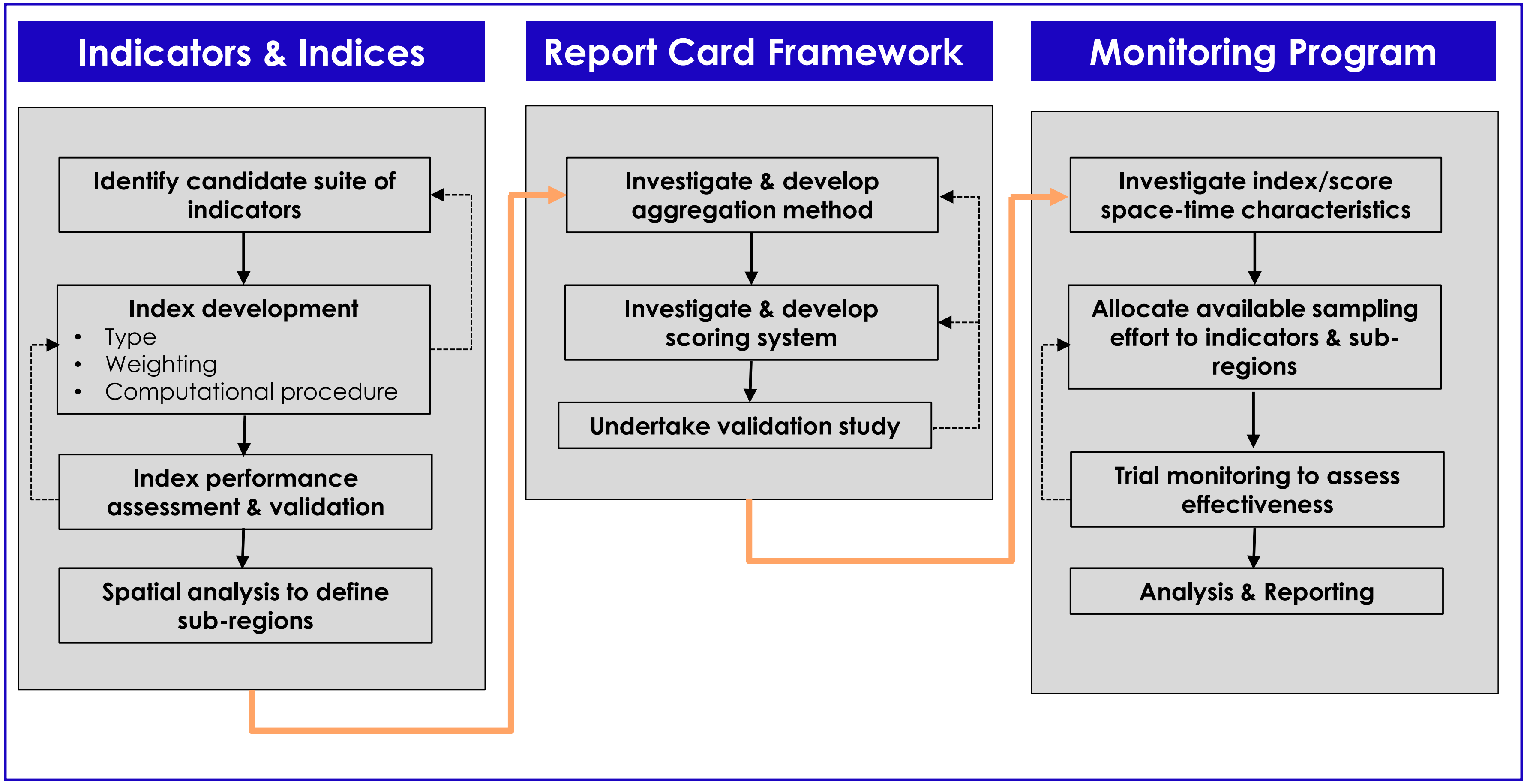

Indicators & Indices
Report Card Framework
Monitoring Program
Identify candidate suite of indicators
Index development
Type
Weighting
Computational procedure
Index performance assessment & validation
Spatial analysis to define sub-regions
Investigate & develop aggregation method
Investigate & develop scoring system
Undertake validation study
Investigate index/score space-time characteristics
Allocate available sampling effort to indicators & sub-regions
Trial monitoring to assess effectiveness
Analysis & Reporting

Figure 18. Suggested steps to develop integrated monitoring and reporting for GHHP.

References

ANZECC and ARMCANZ (2000), National Water Quality Management Strategy, Paper No. 7a Australian Guidelines for Water Quality Monitoring and Reporting, Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand.

Borja, A. and Dauer, D.M. (2008) Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. *Ecological Indicators*, **8**, 331-337.

Borja, A., Ranasinghe, A., Weisberg, S.B. (2009) Assessing ecological integrity in marine waters, using multiple indices and ecosystem components: Challenges for the future. *Marine Pollution Bulletin*, **59**, 1-4.

Bunn, S.E., Abal, E.G., Smith, M.J., Choy, S.C., Fellows, C.S., Harch, B.D., Kennard, M.J., Sheldon, F. (2010) Integration of science and monitoring of river ecosystem health to guide investments in catchment protection and rehabilitation. *Freshwater Biology*, **55**, Supplement 1, 223-240.

Carruthers, T., Carter, S., Florkowski, L., Runde, J., Dennison,W. (2009) Rock Creek Park natural resource condition assessment, National Capital Region Network. Natural Resource Report NPS/NCRN/NRR—2009/109. National Park Service, Fort Collins, Colorado.

Canadian Council of Ministers of the Environment (2001) Canadian water quality guidelines for the protection of aquatic life: CCME Water Quality Index 1.0, Technical Report. In: Canadian environmental quality guidelines, 1999, Canadian Council of Ministers of the Environment, Winnipeg.

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. November 2013. Queensland, Australia.

Dennison, Thomas, Kelsey (2013) Environmental report cards: A tool to integrate monitoring data, engage stakeholders and catalyze actions. Presentation given at International Rivers symposium, Brisbane, 23 September 2013, Australia.

Emerson, J.W., Hsu, A., Levy, M.A., de Sherbinin, A., Mara, V., Esty, D.C., Jaiteh, M. (2012) 2012 Environmental Performance Index and Pilot Trend Environmental Performance Index.Yale Center for Environmental Law and Policy, New Haven.

Feio, M. J., Almeida, S. F. P., Craveiro, S. C., Calado, A. J., (2009) A comparison between biotic indices and predictive models in stream water quality assessment based on benthic diatom communities. *Ecological Indicators* **9**, 497-507.

Henderson, B., Dobbie, M., Harch, B. (2006) An Integrated Environmental Monitoring Program for Adelaide's coastal waters. Final Report for the Adelaide Coastal Waters Study EMP1 Task. CSIRO Mathematical and Information Sciences, Canberra, ACT.

Herzi, A.A., Dovers, S.R. (2009) Australia's indicator-based sustainability assessments and public policy. *Australian Journal of Public Administration*, **68(3)**, 303-318.

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.

Jones, M., Ukkola, L., Eberhard, R. (2103b) The Partnership Program Design for the Development of the Report Card 2010-11 Phase 2, Version 1 May 2013. Fitzroy Partnership for River Health.

Jordan, S.J. and Vaas, P.A. (2000) An index of ecosystem integrity for Northern Chesapeake Bay. *Environmental Science and Policy*, **3**, Supplement 1, 59-88.

de Sherbinin, A., Reuben, A., Levy, M. and Johnson, L. (2013). Indicators in Practice: How Environmental Indicators are Being Used in Policy and Management Contexts. New Haven and New York: Yale and Columbia Universities.

Failing, L., and Gregory, R. (2003). Ten common mistakes in designing biodiversity indicators for forest policy. *Journal of Environmental Management* **68 (2)**, 121–132.

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.

Fox, D.R., Mann, R. (2010) Principles of Statistical Design and Analysis for SCA's Water Monitoring Program with application to the Cox's River Catchment and Wingecarribee Catchment and Reservoir Water Monitoring Framework**.** Environmetrics Australia Report to Sydney Catchment Authority.

Harbans, L. (2011) The introduction to the water quality index. *Water Efficiency*, Sept/Oct 2011, 44-49.

Hedge, P., Molloy, F., Sweatman, H., Hayes, K., Dambacher, J., Chandler, J., Gooch, M., Chinn, A., Bax, N., Walshe, T. (*undated*) An integrated monitoring framework for the Great Barrier Reef World Heritage Area. National Environment Research Program.

Hezri, A.A., and Dovers, S.R. (2006). Sustainability indicators, policy and governance: Issues for ecological economics. *Ecological Economics*, **60**, 86-99.

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 Resources*, **46(11)**, 3544-3552.

Hijuelos, A. and Reed, D. (2013) Methodology for Producing a Coastal Louisiana Report Card, September 13, 2013. The Water Institute of the Gulf.

Kaurish, F.W. and Younos, T. (2007) Developing a standardized water quality index for evaluating surface water quality. *Journal of the American Water Resources Association*, **43(2)**, xx-xx.

Neary, B.P. (2012) A sensitivity analysis of the Canadian Water Quality Index. A report for CCME prepared by Gartner Lee Limited, Ontario, Canada.

United States Government Accountability Office (2011) Chesapeake Bay Restoration Effort Needs common Federal and State Goals and Assessment approach. Report to Congressional Committees.

Whittaker, G.,Lautenbach, S., Volk, M. (2012) What is a good index? Problems with statistically based indicators and the Malmquist index as alternative. 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 R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)

Williams, M. R., Longstaff,B.J. Wicks, E. C. , Carruthers, T. J. B., Florkowski, L. N. (2010) Ecological Report Cards: Integrating Indicators into Report Cards. Pages 79-96 in B. J. Longstaff, T. J. B. Carruthers, W. C. Dennison, T. R. Lookingbill, J. M. Hawkey, J. E. Thomas, E. C. Wicks, and J. Woerner, editors. Integrating and applying science: a practical handbook for effective coastal ecosystem assessment. IAN Press, Cambridge, Maryland.