



Index Stability Score:

An adjunct tool for

Water Quality Index

Reporting

Doug McCarty

Associate Hydrogeologist

November 2019

OFR-20-01

Contents

1.0 Introduction and Background	3
2.0 Summary of the Method	3
3.0 Problem Statement	6
3.1 Demonstration Discussion	6
3.2 Set Size as an Index Determinant	8
3.3 Natural Variability.....	9
3.4 Criteria for Data-Adequate Indices.....	9
4.0 Proposal: Index Stability Score	10
4.1 First term – Mathematical/statistical sufficiency.....	10
4.2 Second term – Natural variability.....	12
4.3 Third term – Data representativeness.....	12
4.4 General Index ISS – Core parameter coefficient	13
4.5 Lorenz coefficient	14
5.0 Index Stability Score Formula.....	15
6.0 Conclusion.....	15
7.0 References	16

1.0 Introduction and Background

The Arizona Water Quality Index (AWQI), a tool developed to communicate water quality information in a concise and understandable way to the public, water quality professionals, and decision-makers, was rolled out in 2015. The WQI operates on a set of disparate water quality data with water quality standards that differ by constituent from locations having differing designated uses. It generates a single standardized number reported on a scale from 0 to 100, with 100 representing excellent water quality. The criterion for a top score is the uniform attainment of water quality standards applying to the reach or site identified. The index can be applied (with alterations) to both general sets of water quality data, and to sets based on specific analytes of interest. Unfortunately, the intervening years since the index was introduced have seen the index pressed into the service of short-term quarterly or monthly time frame reporting it was not designed for, with little apparent recognition that the numbers reported on these time frames are far from reliable and subject to wide fluctuations from one event to the next. The purpose of this white paper is to lay out a case and an approach for the development of an accompanying Index Stability Score on the same index scale of 0-100 to serve as an adjunct reporting tool and to be reported with the score or index value. The Index Stability Score would give users of the index or score reporting a numerical snapshot of how reliable and stable the index or score is considered to be.

2.0 Summary of the Method

The Canadian Council of Ministers of the Environment (CCME) published a technical report in 2001 (CCME, 2001a) that outlines the components and calculations of the CCME index. As this index was the index largely adopted by Arizona, with adaptive changes to Arizona's environment and the specifics of its water quality programs, the original CCME discussion of the index framework is reproduced here as the most succinct explanation of the index composition¹ :

CCME Water Quality Index Formulation

The index consists of three factors:

Factor 1: Scope

F1 (Scope) represents the extent of water quality guideline non-compliance over the time period of interest. It has been adopted directly from the British Columbia Index:

$$F_1 = \left(\frac{\text{Number of failed variables}}{\text{Total number of variables}} \right) \times 100$$

¹ Equations in CCME excerpt not included in document equation numbering.

Where **variables** indicates those water quality variables with objectives which were tested during the time period for the index calculation².

Factor 2: Frequency

F2 (Frequency) represents the percentage of individual tests³ that do not meet objectives (“failed tests”):

$$F_2 = \left(\frac{\text{Number of failed tests}}{\text{Total number of tests}} \right) \times 100$$

The formulation of this factor is drawn directly from the British Columbia Water Quality Index.

Factor 3: Amplitude

F3 (Amplitude) represents the amount by which failed test values do not meet their objectives. F3 is calculated in three steps. The formulation of the third factor is drawn from work done under the auspices of the Alberta Agriculture, Food and Rural Development.

(i) The number of times by which an individual concentration is greater than (or less than, when the objective is a minimum) the objective is termed an “excursion” and is expressed as follows.

When the test value must not exceed the objective:

$$\text{excursion}_i = \left(\frac{\text{Failed Test Value}_i}{\text{Objective}_j} \right) - 1$$

For the cases in which the test value must not fall below the objective:

$$\text{excursion}_i = \left(\frac{\text{Objective}_j}{\text{Failed Test Value}_i} \right) - 1$$

ii) The collective amount by which individual tests are out of compliance is calculated by summing the excursions of individual tests from their objectives and dividing by the total number

² The index definition of a **variable** is a set of results for the same measured chemical or physical property. Examples of variables include dissolved oxygen, total arsenic, pH, lead, and dissolved copper.

³ The index definition of a **test** is an individual sample result on a specific day and time for a specific variable. An example of a hypothetical test for a site might be the concentration of dissolved copper on January 1, 2000. An example of a **failed test** would be a dissolved copper exceedance of the water quality standard for a designated use at the site on Jan. 1, 2000.

of tests (both those meeting objectives and those not meeting objectives). This variable, referred to as the normalized sum of excursions, or *nse*, is calculated as:

$$nse = \frac{\sum_{i=1}^n excursion_i}{\text{Total number of tests}}$$

iii) *F3* is then calculated by an asymptotic function that scales the normalized sum of the excursions from objectives (*nse*) to yield a range between 0 and 100.

$$F_3 = \left(\frac{nse}{0.01 * nse + 0.01} \right)$$

The CCME WQI is then calculated as:

$$CCMEWQI = 100 - \left(\frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \right)$$

The factor of 1.732 arises because each of the three individual index factors can range as high as 100. This means that the vector length can reach

$$\sqrt{100^2 + 100^2 + 100^2} = \sqrt{30,000} = 173.2$$

as a maximum. Division by 1.732 brings the vector length down to 100 as a maximum.

Discussion

...

Applying the Index

Applying this index to water quality data sets must be done with due regard to how the index is formulated. Experience with the British Columbia index has shown that misapplication, or use of the index for purposes for which it was not designed, can lead to erroneous conclusions. There are several rules for application that should be taken into consideration:

a) Index comparisons should only be made when the same sets of objectives are being applied.

...

b) Index comparisons should only be made using the same sets of parameters.

...

c) Care should be taken with older data.

d) The index should be run on parameter sets relevant to the water body being tested.

...

e) Minimal data sets should not be used.

The CCME WQI was not designed to replace proper evaluation of water quality conditions through thorough assessment of water quality chemicals of concern. The CCME WQI should not be run with less than four parameters and four sampling visits per year.

Despite these restrictions on its use, the CCME WQI has been successfully applied in several Canadian jurisdictions and has produced values that contain valuable information with regard to trends through time and spatial discrimination of impacted and non-impacted sites. The committee feels that it has application as a management and communication tool if applied appropriately.

(CCME, 2001a)

For more details on the background and origin of the index, the reader is referred to the Arizona Water Quality Index white paper.

3.0 Problem Statement

3.1 Demonstration Discussion

A selected parameter was considered in an analysis in part to assess the sensitivity of index responses to varying set sizes. *E. coli* was chosen for testing; hypothetical data sets of sizes n=four, n=10, and n=25 were constructed, with single and multiple (2) exceedances at both just over the standard and at ten times the standard. Five different calculation protocols were tested – the default linear multiple calculation, the default formula substituting log 10 values for linear values, the default formula substituting natural log values for linear values, and a simple difference of both log 10 values and natural logarithm values. For the purposes of this objective, the differing calculation protocols are unimportant, save that comparisons should be confined solely within a given calculation protocol. Results are displayed in Figure 1.

1 exceedance at 10x standard						1 exceedance at just over standard					
n =	Linear Calc	Log10 space	LN space	Log10 Diff	LN Diff	n =	Linear Calc	Log10 space	LN space	Log10 Diff	LN Diff
4	40	64	64	62	56	4	64	65	65	65	64
10	66	92	92	90	85	10	93	93	93	93	93
25	80	94	94	94	92	25	97	97	97	97	97
2 exceedances at 10x standard						2 exceedances at just over standard					
n =	Linear Calc	Log10 space	LN space	Log10 Diff	LN Diff	n =	Linear Calc	Log10 space	LN space	Log10 Diff	LN Diff
4	24	48	48	45	37	4	49	50	50	50	49
10	52	85	85	82	74	10	85	86	86	85	85
25	70	94	94	92	88	25	94	94	94	94	94

Figure 1. Exceedance simulations on index responses

3.2 Set Size as an Index Determinant

The demonstration highlights the importance of set sizes as a major determinant of index responses. It is clear to see in the yellow-shaded columns of Figure 1 that the same raw number and magnitude of exceedances can result in vastly different index readings. Numbers range from index values of [64 – 97] and [49 – 94] for one and two exceedances just over the standard respectively, and from [40 – 80] and [24 - 70] for one and two exceedances respectively at values of 10 times the standard for linear factor calculations. These variations are entirely attributable to the number of samples under consideration in each set; a fixed number of exceedances in a set can constitute a widely varying percentage of a set size without minimum size requirements (floors). Logarithm calculation protocols did not vary as much as linear ones, but still varied considerably based on set size. Large sets are not a concern; the index response as set size grows larger decreases and stabilizes as it approaches its asymptotic limit, and the confidence interval around a large-set index narrows as the set grows, assuring a more reliable number. However, small sets are a great concern, as index values are highly unstable and subject to great changes simply due to the expansion of set size. This variability is independent of the addition of any other exceedances or the magnitude of any additional exceedances.

Consequently, use of the index on small sample sets is not recommended. Water quality sampling is an exercise in statistical characterization, with random variations possible for any given sample. The first sample of a set of visits may turn up an exceedance, or the 25th sample might. A larger set of values collected leads to more certainty about measures of central tendency, values and ranges of outlier events, and the percentage of visits that can be expected to result in exceedances of water quality standards. But little can be said with confidence about a single value, or a low number of values, since variation in a population is distributed randomly and an exceedance of standards may occur in any given sampling event, whether the first or the tenth. Index numbers generated from only a few samples have a much higher degree of uncertainty associated with them, and a much lower level of reliability and confidence that can be ascribed to the accurate characterization of the water body. As a provisional rule-of-thumb, set sizes of less than three visits should not be evaluated.

Recommendations for minimal dataset sizes necessarily imply recommendations on time frames for which the index is employed. ADEQ considers data collected at least seven days apart as the minimum time to establish temporal independence of samples. This rule is used to aggregate samples for consideration in the water quality assessment. Using this rule, one month is the minimum to achieve a four-sample set, but logistical considerations for sampling routinely prevent this from occurring. Practically speaking, achieving a minimum data set of four for even a qualified index value requires a minimum of two months of collection time, since field visits are only rarely scheduled on a persistent bi-weekly basis. More typically, visits are made on a monthly or quarterly basis: monthly visits on average might be expected in monsoon season when conditions can change rapidly, while quarterly visits are usual to characterize water quality when conditions are stable and not expected to change much. Seasonal or semi-annual time frames for index determinations are the minimum that make practical

sense for generation of indices, with much longer time frames (annual, 5 year, ten year, historical) more suitable yet. The state of Oregon, for example, employs a ten-year time frame in the use of their water quality index. The water quality index is not a tool that is designed for, nor is it suitable for, snapshot evaluations and cursory judgments about water quality conditions.

3.3 Natural Variability

In addition to the problems posed by consideration of sets that are mathematically too small for the generation of reliable indices or scores, another factor that should be considered is that of natural variability through time and season. One or two samples do not constitute an adequate analysis design to determine if representative conditions have been characterized sufficiently and accurately. Water quality in the Southwest varies widely by season – base flow conditions, summer monsoon storms, winter storms, and spring snowmelt in certain locations all demonstrate dramatically different and distinct water quality characteristics. Furthermore, where storms are concerned, differing storm durations, intensities, and frequencies all result in different water quality responses at a given location. These responses may be rapid and successive in response to a series of storms, or they may be distributed widely and erratically throughout the seasons, with base flow conditions interrupted without warning and for short bursts by quick-moving storms. Hydrological responses in nature occur randomly, and cannot be predicted with any degree of certainty far in advance; thus, a survey design that incorporates these uncertainties and allows for them yields much more reliable trend or summary data over time. For this reason, it is recommended that a minimum of one hydrologic years' worth of data be used in any index calculation.

3.4 Criteria for Data-Adequate Indices

The following criteria, in addition to mathematical terms addressing the concerns of the preceding two sub-sections, are included in an Index Stability Score to provide the minimum assurance that data sets with a general water quality index score can be considered data-adequate and thus reliable to a quantifiable degree. Analyte WQ indices will reflect characteristics in this list consistent with application of an index determination to a single analyte.

- The percentage of all parameters required for a core parameter evaluation represented in the data set
- Each considered parameter (or analyte, if single) has a minimum of ten values
- Data coverage for each considered parameter (or single analyte) represents at least three of the four seasons (seasonal distribution)
- The percentage of analyte records exhibiting values with a "GT" (Greater than) designation (typically, *E. coli* data) is considered as adding uncertainty to the data set
- Both base flow and storm flow conditions are represented
- The duration of the period of record of the set is at least one year

4.0 Proposal: Index Stability Score

The Index Stability Score is presented on the same 0-100 scale as WQ scores/indices. To the extent possible, the mathematical basis of the WQI calculation is followed in the same general format.

4.1 First term – Mathematical/statistical sufficiency

The first term of the stability score addresses the data set's *statistical sufficiency* in evaluating whether the data set is large enough so that the central limit theorem begins to take effect. When the central limit theorem begins to be operative, data sets behave as normally distributed data sets, even if the dataset is a distribution other than normal. This lends added reliability to measures such as the mean and standard deviation and thus provides an indication that sufficient data is present to reduce index variability when considered in sample-by-sample calculations. The term is a pro-ratio based on an ideal minimum set size. The n used as a threshold (denominator) for the term is 30, and proration for the term is normalized on a 0-100 scale. Table 1 summarizes values applied.

If more than 30 samples exist, the sample sub-score is maximized at 100 for all n 's over 30.

Mathematically, the term shall be represented as the percentage of samples relative to a set considered complete at 30 samples:

C_1

$$\text{Where } C_1 = 100 * \frac{n}{30}$$

And n = number of samples for consideration

Table 1. Statistical sufficiency sub-score

n	Term proration	Subscale score	
1	0.03	0.0	*
2	0.07	0.0	*
3	0.10	10.0	
4	0.13	13.3	
5	0.17	16.7	
6	0.20	20.0	
7	0.23	23.3	
8	0.27	26.7	
9	0.30	30.0	
10	0.33	33.3	
11	0.37	36.7	
12	0.40	40.0	
13	0.43	43.3	
14	0.47	46.7	
15	0.50	50.0	
16	0.53	53.3	
17	0.57	56.7	
18	0.60	60.0	
19	0.63	63.3	
20	0.67	66.7	
21	0.70	70.0	
22	0.73	73.3	
23	0.77	76.7	
24	0.80	80.0	
25	0.83	83.3	
26	0.87	86.7	
27	0.90	90.0	
28	0.93	93.3	
29	0.97	96.7	
30	1.00	100.0	
>30	1.00	100.0	

*Insufficient samples to calculate standard deviation; 0 by default

4.2 Second term – Natural variability

The term for **natural variability** takes advantage of a *mathematical property* of the set: the maximum coefficient of variation (CoV) of the data set equals the square root of the number of values in the set. Consequently, the ratio between the CoV and \sqrt{n} will always be a value between 0 and 1. Multiplication of this value by 100 puts the sub-score on a 0-100 scale, and subtraction of the value from 100 re-orientes the score so that a high score represents a dataset with less variability (more reliability), while a low score represents widely-fluctuating values in the reported range.

Mathematically, the term shall be presented as:

C_2

$$\text{Where } C_2 = 100\left(1 - \frac{CoV}{\sqrt{n}}\right)$$

$$\text{And } CoV = s / \bar{X}$$

$$\text{And } s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

$$\text{And } \bar{X} = \frac{\sum X}{N}$$

4.3 Third term – Data representativeness

The third term, **data representativeness**, represents a summation of various conditions intended to characterize whether the set *covers enough different natural conditions* to be truly representative in nature. These conditions include the following:

- Presence of events sampled in storm flow conditions
- Presence of events sampled in base flow conditions
- Seasonal distribution – at least one sample in each of three seasons (counting as one criterion)
- Percentage of results not exhibiting “greater than” values
- Duration of sampling record – percentage of year elapsed between first and last sample. At least one year is required to maximize this criterion sub-score. Durations beyond one year receive full credit.

The number of factors that are not percentage values outlined shall be summed as either being met (1 pt.) or not (0 pts) for a set, then divided by the total number of criteria applicable. The proportion of pass/fail scores to the entire list is then applied as a weight. The percentage criteria are averaged and

weighted by the proportion of total criteria reporting as percentages and summed with the other weighted score. The final weighted value is multiplied by 100 for the third sub-score.

Mathematically, the term shall be presented as:

C_3

$$\text{Where } C_3 = 100 * \left(\sum_{i=1}^n \frac{i}{n} \right) + \sum_{i=1}^2 w_i * x_i$$

w_i is the weight applied to i th term of the percentage sub-criteria

x_i is the percentage of the i th term expressed in decimal form

4.4 General Index ISS – Core parameter coefficient

The General Water Quality Index (GWQI) ISS has a coefficient equal to the percentage of core parameter analytes represented in the data set applied to the overall calculation to ensure incomplete data sets meeting all other criteria do not erroneously generate a high GWQI. Testing in early iterations of ISS development revealed that a single pass/fail item for core scope coverage originally included in the data representativeness term of the index was not sufficiently sensitive enough to penalize the stability score adequately when core parameter coverage was incomplete.⁴ The embedded criterion for core scope coverage was removed from the score and replaced by this coefficient. Since core parameter coverage considerations do not apply to analyte WQIs, the coefficient is not used in that context.

Mathematically, the term shall be presented as:

p_G

Where p_G is the percentage of core parameter analytes represented for a member of the same designated use grouping

⁴ In one test case, a high GWQI was generated with an ISS of apparent acceptability based on only one parameter being present out of a possible estimated 11-12 total in the data set. Application of the core parameter coefficient corrected the erroneous impression.

4.5 Lorenz coefficient

For analyte WQI data sets with fewer than ten records or GWQI data sets with an average of fewer than 10 records per analyte, a Lorenz coefficient based on a cubic curve between 0 and 1 is applied to penalize calculations made on small data sets. For more than ten records (or an average of more than ten), the coefficient is 1. This coefficient is intended to discourage use of the WQI where stability is lacking. For users who insist on reporting with data sets smaller than 10 records, QA disclaimers can be applied and the coefficient removed. In such a case, the reported value is no longer considered an *index* value with full QA assurance, but a *score* with no accompanying assurance of QA integrity.

Water quality scores, where Lorenz coefficients are dispensed with, are considered to have stability scores of 0.

Mathematically, the term shall be presented as:

L_c

Lorenz coefficients are presented in Table 2.

Lorenz curve coefficients	
0	0.000
1/10	0.001
2/10	0.008
3/10	0.027
4/10	0.064
5/10	0.125
6/10	0.216
7/10	0.343
8/10	0.512
9/10	0.729
10/10	1.000

Table 2. Lorenz coefficients for small data sets

5.0 Index Stability Score Formula

The final Index Stability Score formula follows the general model of the GWQI formula:

$$\text{Index Stability Score} = L_c * p_G * \frac{\sqrt{C_1^2 + C_2^2 + C_3^2}}{1.732}$$

$$\text{Where } C_1 = 100 * \frac{n}{30} \quad (\text{Statistical sufficiency sub-score})$$

$$\text{And } C_2 = 100 * \left(1 - \left(\frac{CoV}{\sqrt{n}}\right)\right) \quad (\text{Natural variability sub-score})$$

$$\text{And } C_3 = 100 * \left(\sum_{i=1}^n \frac{i}{n}\right) + \sum_{i=1}^2 w_i * x_i \quad (\text{Data representativeness sub-score})$$

And coefficients are as described previously.

Each sub-score term can range from 0 to 100 before further operations. Squaring each term and summing all under the radical has a theoretical maximum of 30,000. Taking the root of the theoretical maximum yields a value of 173.2. A divisor of 1.732 then normalizes the vector length to a result no greater than 100, with 100 representing high confidence of a stable and reliable score, and 0 representing no confidence in the score stability.

6.0 Conclusion

The Index Stability Score formulation proposed here is intended to provide an adjunct to Water Quality Index reporting. A noted weakness of WQI reporting in its first three years of existence has been the inability of users to assess how much stock to put into the number reported; a 100 reported for a site with two data points is given the same credibility as a 100 reported with a long historical record of water quality data. Many historical sites in TMDL reporting, for example, have calculated WQIs of 90 or better based on one, two or three data points, which gives a misleading sense of assurance of the water quality in areas of known impairment. The former method of qualifying data by attaching a “Q” or a “P” (for *qualified* and *provisional* qualifiers, respectively) to the value has been easily disregarded; numbers can readily be stripped of their qualifiers and used independently of consideration of them. Additionally, the

suffixes posed problems with data recording and storage, since they created functional issues with sorting and calculations. The Index Stability Score is a separate number reported with the index value that circumvents some of these limitations. However, no reporting conventions can ultimately replace the considered and careful judgment of the person evaluating the data. It remains incumbent upon users of the index/score values to use the information responsibly and with due care.

A second noted distinction to come about because of the development of this Index Stability Score is actual formal set size criteria (GWQI: average of the grouped by parameter sample counts ≥ 10 ; AWQI: >10) marking the difference between water quality indices and water quality scores. A clear and necessary distinction is now formally established between **index reporting** and **score reporting**. In this way, where internal requirements mandate short-term reporting on less-than-ideal set sizes, the numbers can still be generated and supplied, with the caveat that they are now formally called **scores** (with a default ISS of 0 and no assurance of quality/certainty in reporting). Index calculations with data set sizes consistent with recommendations are now formally called **index values**, and the methodology, more rigorous and quantifiable, is consistent with using the index in scientific studies, such as in long-term trend analysis and statewide studies by basin or watershed.

7.0 References

- Canadian Council of Ministers of the Environment. 2001(a). *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.
- Canadian Council of Ministers of the Environment. 2001(b). *Canadian water quality guidelines for the protection of aquatic life: CCME Water Quality Index 1.0, User's Manual*. In: Canadian environmental quality guidelines, 1999, Canadian Council of Ministers of the Environment, Winnipeg.
- McCarty, Doug, 2015. *The Arizona Water Quality Index: A Communication Tool for Water Quality Summaries*, 2015. Arizona Department of Environmental Quality, Phoenix, AZ.
https://static.azdeq.gov/wqd/reports/wqi_18.pdf