

**ASSESSING WATER QUALITY
OBJECTIVES:
DISCUSSION PAPER**

Rob Goudey

December 1999

ASSESSING WATER QUALITY OBJECTIVES: DISCUSSION PAPER

Rob Goudey

Environment Protection Authority
40 City Road
Southbank Victoria 3006
Australia

Printed on recycled paper

© State of Victoria, December 1999

ISBN: 0 7306 7569 6

FOREWORD

The Victorian Catchment Management Council has a legislative requirement to report on the condition and management of the State's land and water resources. An effective and efficient Statewide water quality monitoring and assessment program is an important component of overall integrated natural resources monitoring and assessment.

In 1996 the Council, in partnership with the Department of Natural Resources and Environment and the Environment Protection Authority, completed a review of the State's water quality monitoring network. Since then the Environmental Water Quality Monitoring Committee has been working towards implementing the recommendations to deliver an integrated and resource efficient approach to water quality monitoring and information management in Victoria.

One recommendation was to introduce new processes to analyse the data being collected so that information more useful to management agencies is produced. EPA undertook an investigative study to determine design requirements for a program intended to statistically test attainment of environmental water quality objectives at a Statewide level. This report presents the findings of the study, including a discussion of the implications of various options for the statistical assessment of compliance with water quality objectives. The Committee will consider these implications in developing further network improvements.

Professor Barry T Hart
Chair, Environmental Water Quality Monitoring Committee
Catchment Management Council Member

TABLE OF CONTENTS

1. BACKGROUND.....	1
2. WHAT WE HAVE BEEN DOING TO ASSESS COMPLIANCE.....	1
3. WHAT’S WRONG WITH WHAT WE’VE BEEN DOING TO ASSESS COMPLIANCE?.....	2
Objectives based on absolute limits for individual values.....	2
Reporting the percentage of samples exceeding a given limit.....	3
Estimation of extreme percentiles	4
4. OPTIONS FOR COMPLIANCE ASSESSMENT.....	6
Approaches which don’t use a statistical hypothesis testing framework.....	6
Advantages of Option 1	6
Disadvantages of Option 1	6
Advantages of Option 2	7
Disadvantages of Option 2.....	7
“Pass/Fail” compliance monitoring (Option 3)	7
Example of “Pass/Fail” compliance testing	8
Advantages of Option 3	8
Disadvantages of Option 3.....	9
Approach based on ability to detect breaches of particular magnitudes with a specified power (Option 4)	10
Example including design and testing of a time-relative water quality objective	11
Advantages of Option 4	12
Disadvantages of Option 4.....	12
5. RECOMMENDATIONS.....	13
6. ACKNOWLEDGEMENTS.....	13
7. REFERENCES	13

1. BACKGROUND

The Environment Protection Authority has operated an inland fixed sites water quality monitoring network in Victoria since 1984. Water samples collected monthly at fixed locations on streams and lakes are analysed for physico-chemical indicators of pollution. The data from these analyses are used to support the primary objective of the monitoring network, which is to test for temporal trends in water quality. A secondary use of the data has been to determine whether those water bodies comply with State Environment Protection Policy (SEPP) objectives specified for each indicator (EPA 1995). Although this additional purpose is considered secondary, it is also considered to be an important use. This is reflected in the EPA Corporate Plan (1997-2000), which states the strategic goal for water quality as:

“The long term goal for the water quality program is to have no breaches of the environmental quality outcomes agreed by the community and expressed in State environment protection policies”.

There have been concerns about whether the network data can be validly used for the purpose of assessing compliance with SEPP objectives as they are currently written. The Victorian Water Quality Monitoring Network (VWQMN), of which the EPA network is now a part, was reviewed in 1996 (DNRE, CALPC and EPA 1997). A study has been undertaken into the use of statistical methods to assess compliance with water quality objectives to address one of the recommendations of the review.

This report contains a discussion of the implications of various options for the statistical assessment of compliance, with the intention of providing timely input to the impending review of the Waters of Victoria policy.

Although this study has been conducted in the context of water quality of inland waters, there is the potential that the recommended approach may be beneficial to other areas of EPA monitoring and compliance assessment.

2. WHAT WE HAVE BEEN DOING TO ASSESS COMPLIANCE

The water quality objectives for indicators monitored in various stream segments are listed in SEPPs. Objectives are typically stated in terms of limits on individual, median and percentile values of pollution indicators over each year.

- If any individual value of an indicator measured at a site during a particular year exceeds its SEPP limit for individual values, then the site has failed to comply with the limit for that year.
- If the annual median of values of an indicator measured at a site during a particular year exceeds its SEPP limit for the median, then the site has failed to comply with the median limit for that year.
- If the annual 90th percentile of values of an indicator measured at a site during a particular year exceeds its SEPP limit for the 90th percentile, then the site has failed to comply with the 90th percentile limit for that year.

- If an indicator monitored at a site does not fail to comply with any of these limits during a particular year, then the site is said to be in compliance for that year.

The sampling frequency for the monitoring network is currently monthly. Given that compliance is evaluated annually, the assessment is made based on no more than 12 observations. Individual concentrations and statistics based on these 12 observations are assumed to be representative of the water body segment and time period over which the assessment is made. No estimates of uncertainty in assessing compliance are derived.

A summary of numbers of breaches of SEPP objectives for water quality at EPA sites is published in EPA's Annual Report. Breaches of various water quality objectives are now also published in the VWQMN Annual Report for all sites in the Statewide network.

3. WHAT'S WRONG WITH WHAT WE'VE BEEN DOING TO ASSESS COMPLIANCE?

Statistical hypothesis tests aim to infer something about a statistical *population* based on a statistical *sample* from that population. In general, difficulties occur with the current compliance assessment procedures because they do not make allowance for the fact that water quality measurements are actually a random sample from a statistical population (ie the waterbody), and so are subject to random variation.

Without allowance for statistical uncertainty in sampling, a statistical inference is not possible, and so the assessment of compliance is limited to the sample rather than applying to the waterbody. Since water quality policy objectives clearly relate to the waterbody, there is a logical inconsistency between the way the objective is written and the way it is assessed. That inconsistency may result in a declaration of noncompliance while the water body actually complies with the objective, or a declaration of compliance while the water body actually does not comply with the objective.

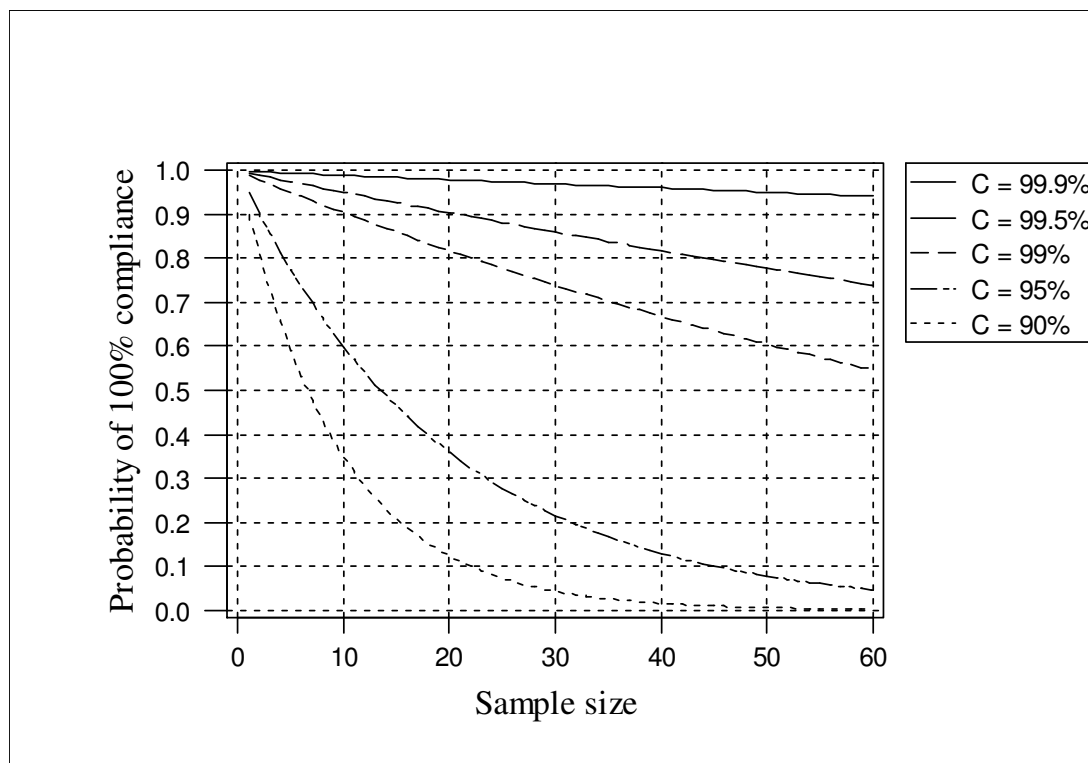
More specifically, problems arise when attempting to (i) assess compliance relative to absolute limits on individual values, (ii) report the percentage of samples exceeding a given limit, or (iii) estimate extreme percentiles. These three problems are discussed below.

Objectives based on absolute limits for individual values

An *absolute limit* on a water quality indicator is one which no individual value of the indicator is permitted to exceed, eg "No individual mercury concentration shall exceed 0.00005 µg/l". The main difficulty with statistical assessments of compliance of individual measurements is that a variance estimate is required for a statistical test to be performed. The question is then one of what data to use to calculate that variance. Two possibilities are to collect replicate data during the assessment period or to derive a variance from historical data. However, the current compliance assessment method does not make use of any variance, and so does not provide for a statistical test of the hypothesis of compliance.

Another difficulty relates to the dependence between *the probability of a breach* and the *sampling rate*. Consider a waterbody which, in reality, meets an objective for a percentage (C) of the time. Under the assumptions of independent random sampling and no temporal trend, the probability of detecting *no* failed measurements depends on the sample size. This

is represented diagrammatically in the following figure, which shows the chance of declaring 100% compliance in samples of various sizes from populations with true percentage compliance as specified (ie for C=99.5%, 99%, 95% and 90%).



Clearly, the probability of obtaining 100% compliance is strongly dependent on the sampling rate. The figure also shows that it is possible to fabricate an illusion of improved performance by taking fewer samples. In addition, this dependence on sampling rate may lead to errors in comparing compliance rates of different sites and years if sample sizes have varied.

For a sample of size 12 (eg monthly sampling over one year), this means that a site which fails to comply for 5% or 10% of the time (ie C = 95% or 90% respectively) has a significantly large chance of being found to be compliant.

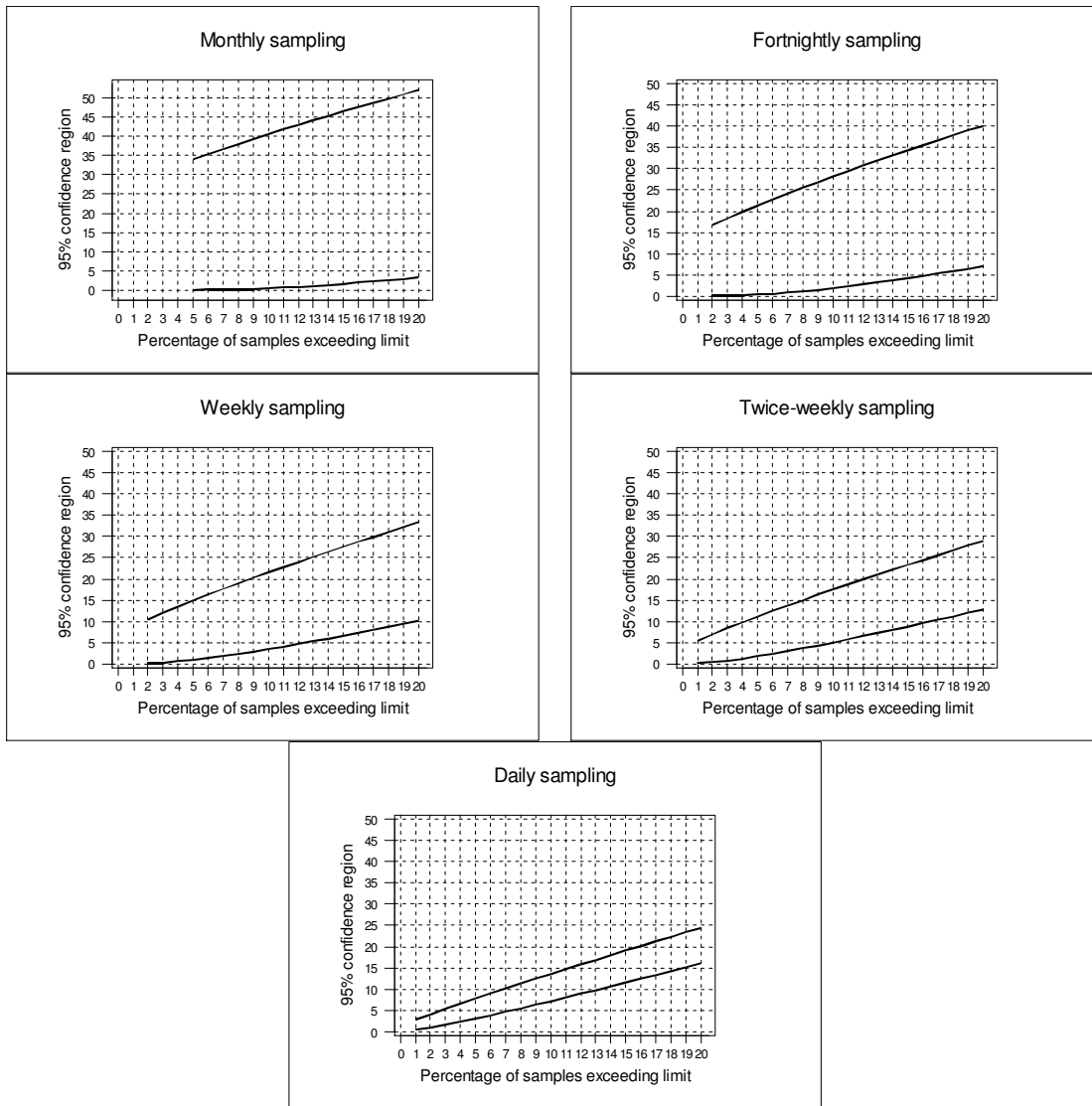
Reporting the percentage of samples exceeding a given limit

Percentage of samples exceeding a particular limit (e.g. a policy limit) can be derived as

$$\frac{\text{total number of samples exceeding limit}}{\text{total number of samples collected}} \times 100\%$$

The intention in calculating this value is usually to provide an estimate of the proportion of time that the limit has been exceeded at a particular site during the assessment period.

The following figure shows 95% confidence intervals for estimates of the percentage of observations exceeding a limit. A noticeable feature is the degree of uncertainty conferred by monthly sampling. If the true percentage of values exceeding a limit is 5%, then a sample of size 12 will provide a very uncertain estimate of that percentage (ie 0% - 34%).

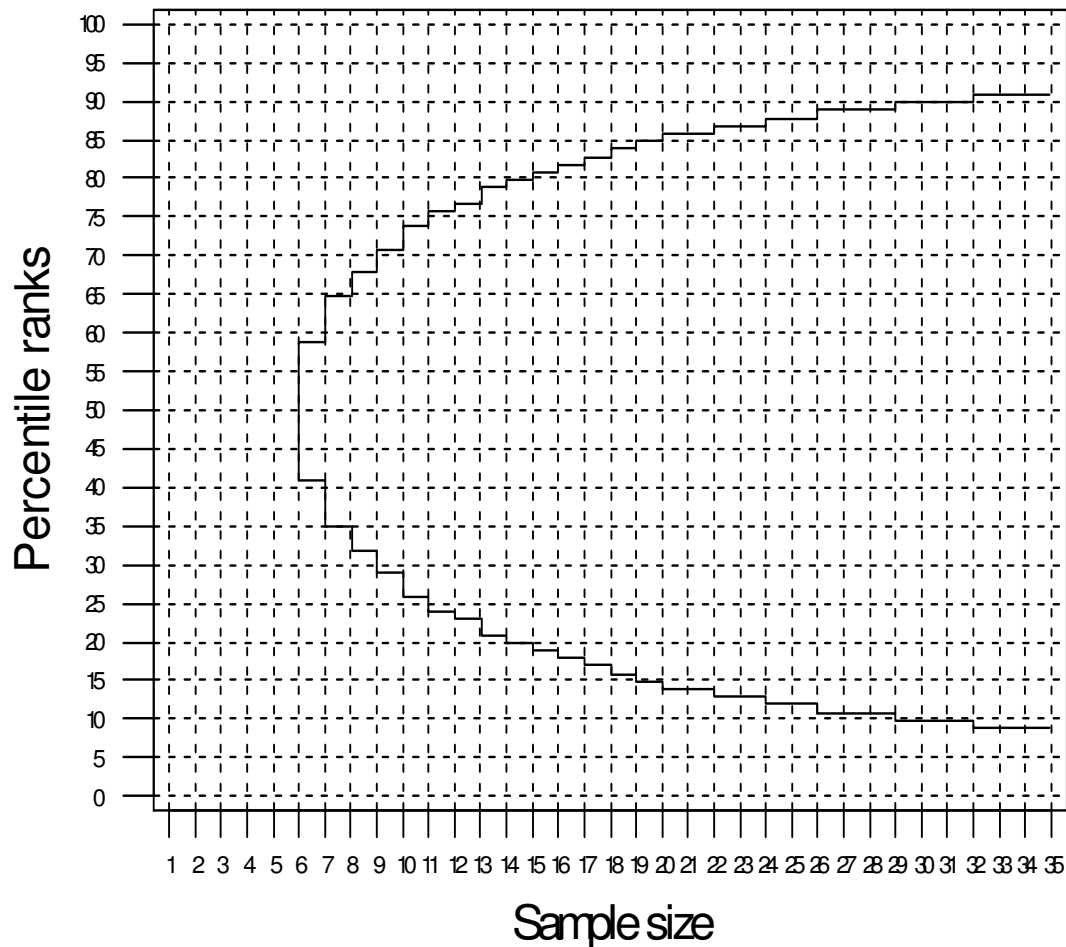


The practice of comparing individual samples to a set limit and then using these comparisons to estimate the percentage of values above limit for the whole assessment period will result in considerable uncertainty, even for more intense sampling frequencies (such as fortnightly). The use of these values to make comparisons between sites and between years would therefore be unwise. A possible solution may be to lengthen the assessment period to include more data (eg 3 years), however this will also lengthen the return time for assessment of compliance.

Estimation of extreme percentiles

It is possible to calculate confidence intervals for percentiles using sample data. When reporting a 95% confidence interval for a given percentile, it is desirable that the entire confidence interval lie within the range of the sample data (ie. somewhere between the minimum and the maximum). In this way there can be at least 95% certainty that the sample data at least contain the 'true' percentile which is being estimated.

The figure below shows the percentile ranks corresponding to smallest and largest percentiles estimable with (nonparametric) 95% confidence for a given sample size.



It can be seen that

- At least six data values are needed to derive a 95% confidence interval for the population median.
- Monthly sampling over one year (12 data values) allow the estimation of 25th, 50th (the median) and 75th percentiles (i.e. the *quartiles*) and their 95% confidence intervals, but is insufficient for calculating 80th and 90th percentiles.
- 95% confidence intervals for the 80th and 90th percentiles would require sample sizes of 14 and 29, respectively.

If indicators exhibit seasonal variation (eg total Phosphorus), then fortnightly sampling over one year would be needed to allow estimation of seasonal medians. Fortnightly sampling of non-seasonal indicators (eg Mercury) would allow annual estimation of the 80th percentile but not the 90th percentile. For 80th and 90th percentiles estimated from monthly data over one year, there can only be 93% certainty and 72% certainty, respectively, that the ‘true’ population percentiles are within the range of the sample data. Such percentiles can not be reliably estimated at the 95% level of confidence on an annual basis using monthly data collection.

4. OPTIONS FOR COMPLIANCE ASSESSMENT

Given the preceding discussion and the present resource limitations for monitoring, there appear to be four main options:

1. Allow the present methodology to continue. Do not modify either the water quality objectives or the monitoring program.
2. Leave the monitoring program as it is, but modify water quality objectives so that they relate to assessment of the samples only.
3. Retain the current expression of water quality objectives in policy, but adjust the monitoring program to allow sufficiently accurate estimates of percent compliance, medians and upper percentiles to be calculated.
4. Adopt a compliance assessment procedure which recognises both the limitations on resources and the requirement to detect breaches of particular magnitudes with the specified power.

Approaches which don't use a statistical hypothesis testing framework

Options 1 and 2 are straight forward in their application because they involve no change from the methodology which is already in place (see page 1). A detailed discussion of difficulties with the current approach was given on pages 2 to 5 of this report.

Advantages of Option 1

- No modification of water quality objectives
- No changes to sampling frequencies
- Greater flexibility to change things at a later time (eg reallocate resources) because there are no statistical sample size requirements to satisfy.

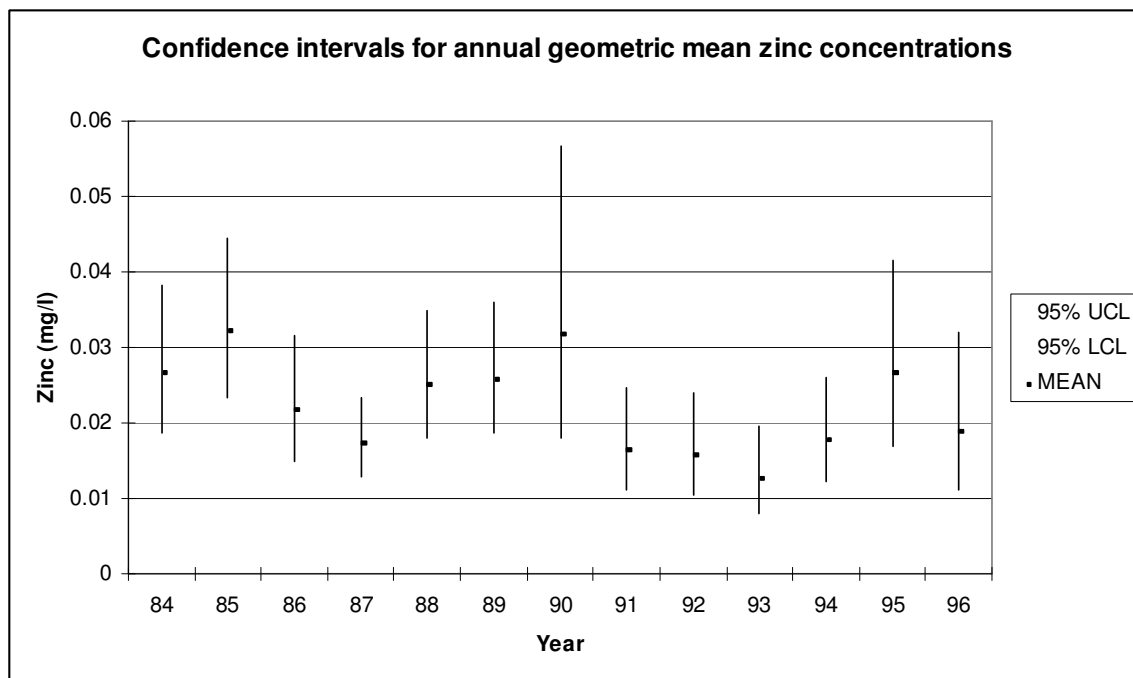
Disadvantages of Option 1

- It is not possible to statistically assess compliance of individual measurements with policy objectives, unless a variance can be estimated using either (i) additional data collected during the assessment period or (ii) historical data. The collection of sufficient additional assessment period data would involve significant additional cost, and is not a realistic option. An estimate of variance from historical data would require prior adjustment of the data for any temporal trends, seasonal variations and serial correlations. Also, the validity of this method rests on the assumption that the assessment period variance has not changed since the historical period.
- It is not possible to derive sufficiently accurate estimates of percent compliance or upper percentiles (ie 80th and 90th percentiles) based on only 12 samples per year.
- It is possible to derive estimates of the median provided that the indicator is not seasonally affected.
- No statistically valid reporting of percent compliance for all objectives is possible.

In case A the entire 95% confidence interval lies below the objective, and in case D the entire 95% confidence interval lies above the objective. The decision in these two cases is clear cut. However, in cases B and C the result of compliance assessment depends on where the benefit of doubt is given. Warn (1989) considers that compliance is unresolved in cases B and C, and points out that the likelihood of an ‘unresolved’ case is inversely related to the sample size.

Example of “Pass/Fail” compliance testing

The following graph was derived using historical zinc data from a VWQMN fixed site. For this example, suppose there is a policy objective of 0.02 mg/l for the annual geometric mean zinc at this site.



If the benefit of doubt is given to the discharger(s), then zinc only failed to comply with the objective in 1985, but complied during all other years. If the result is taken on face value, then zinc complied with the limit for six out of the 13 years. If the benefit of doubt is given to the environment, then zinc only complied in 1993.

If Warn’s approach is adopted, then zinc complied in 1993, failed to comply in 1985, and compliance is unresolved for all other years. For this particular example, to reduce the number of unresolved cases to just fewer than half would require the samples to have been collected approximately every ten days.

Advantages of Option 3

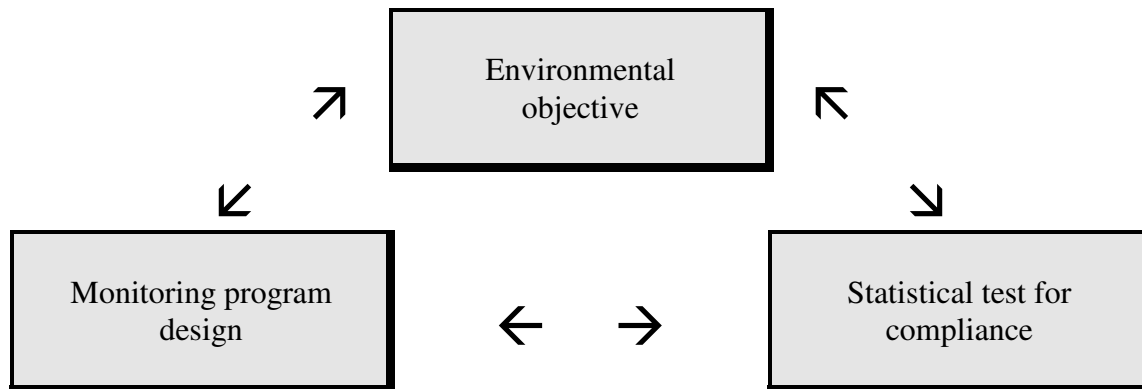
- No changes to the way water quality objectives have been written.
- Conclusions about compliance with percentile limits could be drawn from the current monitoring network with a greater level of confidence.
- This approach has been quite well documented in the water quality literature since the late 1980’s, and represents a considerable improvement on the previous two options.

Disadvantages of Option 3

- It is not possible to statistically assess compliance of individual measurements with policy objectives, unless a variance can be estimated using data collected during either the assessment period or an historical period. The same difficulties associated with the derivation of that variance apply.
- 14 samples per year would be needed to estimate an 80th percentile with 95% assurance that the 'true' percentile is within the range of the data. 29 samples per year would be needed to estimate a 90th percentile with 95% assurance that the 'true' percentile is within the range of the data. If indicators vary seasonally, then fortnightly sampling would be required in order to estimate seasonal medians, and upper percentile estimates may not be feasible at all.
- Sampling frequencies required to estimate percent compliance to within desired accuracies require increased resources and may prove prohibitive in terms of cost. There may be significant potential for criticism of diversion of more resource money to more monitoring rather than actions.
- Although this approach has been well documented in the literature, it has also been criticised in the environmental monitoring literature (Green 1989). The basis of the criticism was that the approach ignores both
 - (i) the magnitude of difference between *acceptable* and *unacceptable* levels which is considered important to test for, and
 - (ii) the power to detect such a difference given that one has occurred.
- Because the size of the difference between *acceptable* and *unacceptable* levels has not been specified, there is the potential for wasting money by either:
 - (i) sampling with inadequate intensity, and so providing inadequate power to detect noncompliance at magnitudes of concern, or
 - (ii) sampling with excess intensity, and so providing redundant effort and expense, and the tendency to detect 'effects' smaller than those of environmental concern.In such instances, the level of statistical significance has not been adequately matched to the environmental significance.

Approach based on ability to detect breaches of particular magnitudes with a specified power (Option 4)

This approach considers questions such as “How should the objective be written?”, “What monitoring design is required?” and “Which statistical test should be used?” to be logically interconnected, since the answer to any one of those questions depends on the answers given to the other two.



An objective which is meaningful in an environmental sense needs to be testable using the monitoring data. The monitoring program must allow a sufficient quality and quantity of data for testing the objective. The statistical test should be one which is the most appropriate for testing the objective using the monitoring data. Inconsistencies can easily arise where one or two of these components have been specified without proper consideration of the remaining component(s).

Water quality objectives may be written in a form which is either *absolute* or *relative*, according to the following table.

Time periods

		Assessment period only	Assessment period and historical period
<u>Sites</u>	Assessment site only	<i>Absolute objective</i>	<i>Time-relative objective</i>
	Assessment site and comparison site(s)	<i>Site-relative objective</i>	<i>Site-and-time-relative objective</i>

An absolute objective contains limits in terms of the original scale of measurement, whereas the limits in a relative objective are expressed relative to either an historical period, a comparison site, or some aspects of both.

The objective first specifies the size of the deviation from acceptability which will be tested for. A convenient way to do this is to state an *acceptable* level and an *unacceptable* level.

The acceptable level may be set based on some absolute value, an historical period value, a comparison site value, or the historical differences between the assessment site and comparison site values.

The unacceptable level specifies the deviation from acceptability at which noncompliance will be declared.

A statistical test can be used to test whether measurements were more likely to have arisen from an *acceptable* population or an *unacceptable* population.

The objective next specifies the allowable sizes of the following two probabilities:

- the *power* of the test, which is the probability that a deviation of the specified magnitude will be detected if it occurs, and
- the *significance level* of the test, which is the probability that noncompliance is declared erroneously.

The relative sizes of these two probabilities may depend on the relative costs of missing noncompliance versus declaring noncompliance erroneously.

An estimate of the variance is then required, and this could be derived either by way of a pilot study or by analysis of historical data. Once the variance has been obtained, the next step is to determine whether there is a practicable sample size which will allow a statistical test with the desired power and significance level. If not, then some compromise of the power and significance level may be possible. If a practicable sample size can not be achieved, then that particular objective may not be assessable for compliance in the manner attempted. This means that the indicator may be inappropriate in a monitoring framework, and this should have significant influence regarding the choice of objectives for policy and indicators for monitoring.

Example including design and testing of a time-relative water quality objective

This example uses turbidity data from a VWQMN site. Suppose there are *acceptable* and *unacceptable* policy objectives of 12.8 NTU and 30 NTU respectively for geometric means. The acceptable level has been chosen based on historical data from the 1984-88 period, and represents a suitable *target* value for turbidity at that site. The unacceptable level has been chosen based on studies of effects of elevated turbidity on stream biota. Suppose that the policy also states that the probability of correctly concluding noncompliance must be greater than 90%, and the probability of correctly concluding compliance must be greater than 90%.

Monitoring design stage

The historical period standard deviation was used to estimate the assessment period sampling frequency required to test the hypotheses:

Null hypothesis: Assessment period mean = Historical period mean

Alternative hypothesis: Assessment period mean > Historical period mean

The estimated sampling frequency was 6 samples per year for this example, ie this is the minimum number of samples per year required to test for variation from the historical geometric mean given that:

1. we wish to be at least 90% certain of detecting an increase at least as large as the difference between the *acceptable* and *unacceptable* values, and
2. we wish to be at least 90% certain of not erroneously declaring noncompliance.

Compliance testing stage

A subsequent test using six samples from the assessment period, and allowing a probability of 90% of correctly concluding compliance, rejected the hypothesis of equality of the historical and assessment period means in favour of the alternative hypothesis. This means that turbidity at the site failed to comply with the objective during the assessment period. This result can be stated with 90% certainty that noncompliance did not occur by chance alone.

If the example had shown the hypothesis of equal means to have been accepted, then the result could have been stated with 90% certainty that an increase at least as big as the difference between the *acceptable* and *unacceptable* values could have been detected had it occurred.

Advantages of Option 4

- The water quality objective would be stated in a manner that allows assessment of compliance at a site.
- This represents a more complete statistical design in that all of the important design elements need to be specified, and is therefore more scientifically defensible.
- The approach ensures that the requirements of the water quality objective, the monitoring design and the statistical test for compliance are either jointly satisfied, or the reasons why they cannot be satisfied are clear.
- The approach can select from a wider range of types of water quality objectives, and so may be more flexible regarding specific monitoring contexts, making more efficient use of comparison site information.
- Because the size of the difference between *acceptable* and *unacceptable* levels and the power of the test both need to be prespecified, any detected noncompliance is of environmental significance as well as statistical significance.
- Objectives expressed as changes relative to historical data share common elements with the aim of the trends analysis program. There is scope for a better interface, or perhaps merging, between the two objectives that the program is intended to meet.

Disadvantages of Option 4

- Water quality objectives would need to be rewritten in a manner that allows the statistical assessment of compliance.
- Sampling rates would need to be determined on the basis of historical data or pilot data, and may vary across the monitoring network for different water quality indicators and beneficial uses.

- Although this approach requires prespecification of the magnitude of difference between *acceptable* and *unacceptable* levels which is considered important to test for, the size of that difference may not actually be known for some indicators and beneficial uses.
- A greater, but not prohibitive, level of statistical conceptual knowledge would be needed on the part of those attempting to interpret the results of compliance assessments. There is potential for initial confusion in interpretation of results due to change in approach and lack of understanding of statistical basis on the part of those doing the interpretation.

5. RECOMMENDATIONS

1. That the current methodology for assessing compliance with water quality objectives be modified to allow logical consistency between the way the objective is written and the way it is assessed.
2. That the feasibility of implementing Option 4 be investigated in the light of
 - (i) the ability to incorporate water quality objectives that are appropriately defined to enable compliance assessment in a policy framework (ie WOV), and
 - (ii) the current network site distribution and limitations on resources.
3. That for whatever approach is adopted, the assumptions and limitations of the approach are made clear and transparent to all stakeholders.

6. ACKNOWLEDGEMENTS

This study was funded and undertaken by EPA for the Environmental Water Quality Monitoring Committee.

The author wishes to thank Lisa Dixon for her guidance and for her extensive comments on drafts of this report. Much of the groundwork for the compilation of this report was completed with the assistance of Bill Lloyd-Smith. Valuable comments and guidance on the statistical contents were also given by Sylvia Esterby, Bruce Mapstone, Tony Warn and Mark Burgman. Thanks are also extended to Dave Robinson, who provided the sample water quality data and advice regarding their interpretation, and to Helen Schilling, who did the necessary followups to ensure publication of this report.

7. REFERENCES

Crabtree, R.W., Crockett, C.P. and Ellis, J.C. (1989) "Continuous effluent consents: modelling and compliance testing" *Environmental Monitoring and Assessment* **12**: 149-164.

DNRE, CALPC and EPA (1997) *Testing the Waters*.

EPA (1995) *State Environment Protection Policy (Waters of Victoria) Draft Schedule F7 - (Waters of the Yarra Catchment)* Publication No. 471 Environment Protection Authority (Melbourne).

Goudey, R. and Lloyd-Smith, B. (1998) *Statistical Assessment of Compliance with Water Quality Objectives* (In press).

Green, R.H. (1989) "Power analysis and practical strategies for environmental monitoring" *Environmental Research* **50**: 195-205.

Warn, A.E. (1989) "Auditing the quality of effluent discharges" *Environmental Monitoring and Assessment* **12**: 165-181.