Collaborative Project (large-scale integrating project)
Grant Agreement 226273
Theme 6: Environment (including Climate Change)
Duration: March 1st, 2009 – February 29th, 2012

WISER

Water bodies in Europe: Integrative Systems to assess Ecological status and Recovery

DELIVERABLE

Deliverable D6.1-2: Manuscript reviewing components of uncertainty and their assessment, including guidelines for estimation and quality assurance

Lead contractor: **Bournemouth University (BourneU)**
Contributors: **Ralph Clarke (BourneU)**

Due date of deliverable: **Month 30**
Actual submission date: **Month 34**

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Content

Overview

This deliverable D6.1.2 is the following draft manuscript to be submitted as a WISER paper to a refereed journal:

Assessing uncertainty of indicators of Water Framework Directive ecological status class

Ralph T. Clarke

Centre for Ecology and Environmental Sciences, School of Applied Sciences, Bournemouth University, Talbot Campus, Poole, Dorset, BH12 5BB, UK

Abstract

Since the introduction of the European Union Water Framework Directive in 2000, considerable effort has been made the Member States to develop biological assessment and monitoring systems for the ecological status class of all of their water bodies (river stretches, lakes, transitional and coastal waters) based on one or more biological quality elements (fish, macroinvertebrates, diatoms-phytoplankton, macrophytes and physical habitats). In accordance with the WFD, these assessment systems have usually been derived by the use of one or more biological indices (often termed metrics) derived from the sampled biological taxonomic composition and diversity, which are converted to Ecological Quality Ratios (EQRs) through standardised by reference condition values of the metric(s) for each water body type and then classified into one of five ecological status classes. All of these steps and every sampling and other methodological decisions you make can affect the waterbody assessment and are potential sources error or uncertainty.

In this paper, the various sources of uncertainty are considered in more detail. The best-available datasets for assessing uncertainty in WFD status class of UK rivers based on macroinvertebrate sampling are used to demonstrate how spatial and temporal variance in metric values can be estimated. New free-available software WISERBUGS (WISER Bioassessment Uncertainty Guidance Software) is described which can help to quantify the effect of this estimated sampling variability on the confidence of assigning water bodies to status classes.

Keywords

Water Framework Directive, WFD, uncertainty, confidence, ecological status class, sampling variation, metric, multi-metric indices

1. Introduction

Any ecological index is of little use without some understanding of the sources and sizes of the sampling error and other uncertainties in its estimation (Clarke et al., 1996).

The European Water Framework Directive (WFD) (European Union, 2000) requires Member States to assess, monitor and, where necessary, improve the ecological quality of its water bodies (river stretches, lakes, transitional/estuarine and coastal waters). The WFD prescribes that such bioassessments should be based on the values for one or more Ecological Quality Ratios (EQRs), each classified into one of five ecological status classes (high, good, moderate, poor, bad), where the EQRs represents the extent of discrepancy between the values of biological parameters observed for a water body (WB) and the values of the same parameters expected for that type of WB if it was in reference condition. The biological parameters usual summarise some aspect of taxonomic diversity or composition as quantitative indices (often referred to as metrics) The overall status class for a WB is based on the use of EQRs and estimated status classes for one or more sampled (or surveyed) biological quality elements (BQEs), namely fish, macroinvertebrates, diatoms-phytoplankton, macrophytes and physical habitats. The WFD ecological status class of European rivers, lakes, transitional and coastal waters has been one of the most high profile “ecological indicators”. Any such classification measures of aquatic ecological quality are of little value without some knowledge and quantitative estimates of their susceptibility to sampling/surveying error and other uncertainties and of the confidence in assigning individual water bodies to ecological status classes. In recognition of this, the WFD states that ‘estimates of the confidence and precision attained by the monitoring system used shall be stated in the river basin monitoring plan’ (European Union, 2000, Annex V, section 1.3.4).

Understanding the causes of change, and especially decline, in WFD ecological quality and providing advice on measures to improve quality, including for water body management plans, requires some quantitative knowledge of the relationship between potential stressor variables and the biotic response measures (Johnson et al. 2006). This relationship is often assessed by developing statistical or maybe more mechanistic models, calibrated by field observations and estimates of all variables and model parameters. If the modelled relationship between observed values of the biotic metric and the estimated values of the stressor variable(s) is very good, then not only must the underlying relationship be strong (although necessarily causal), but the sampling errors in the observed biotic metric values for each waterbody (or site) must be low relative to the total variance in metric values between all WB in the relationship (Fig. 1(a)). (In addition, the estimation errors for the stressor variable values (e.g. lake mean annual total phosphorus concentration) for a each waterbody must be small relative to the total variance in stressor variable values amongst all waterbodies of the same type (i.e. same estimated/predicted reference condition values). However if, as is common in field-based ecological modelling, the observed data-based biotic-stressor relationship is not very strong, then it is important and very useful to know whether this is because the underlying true relationship is weak (cases (b) and (d) in Fig. 1) or because the true relationship is strong but spoilt by our high sampling errors in estimating the biotic metric values (case (c) in Fig.1). Therefore in any such modelling, for WFD or other objectives, it is useful to determine the sampling variance (i.e. precision) in the estimates of site and waterbody biotic index values, especially relative to between-waterbody and total variability amongst all waterbodies, sites and samples of similar types (e.g. with similar reference condition values).

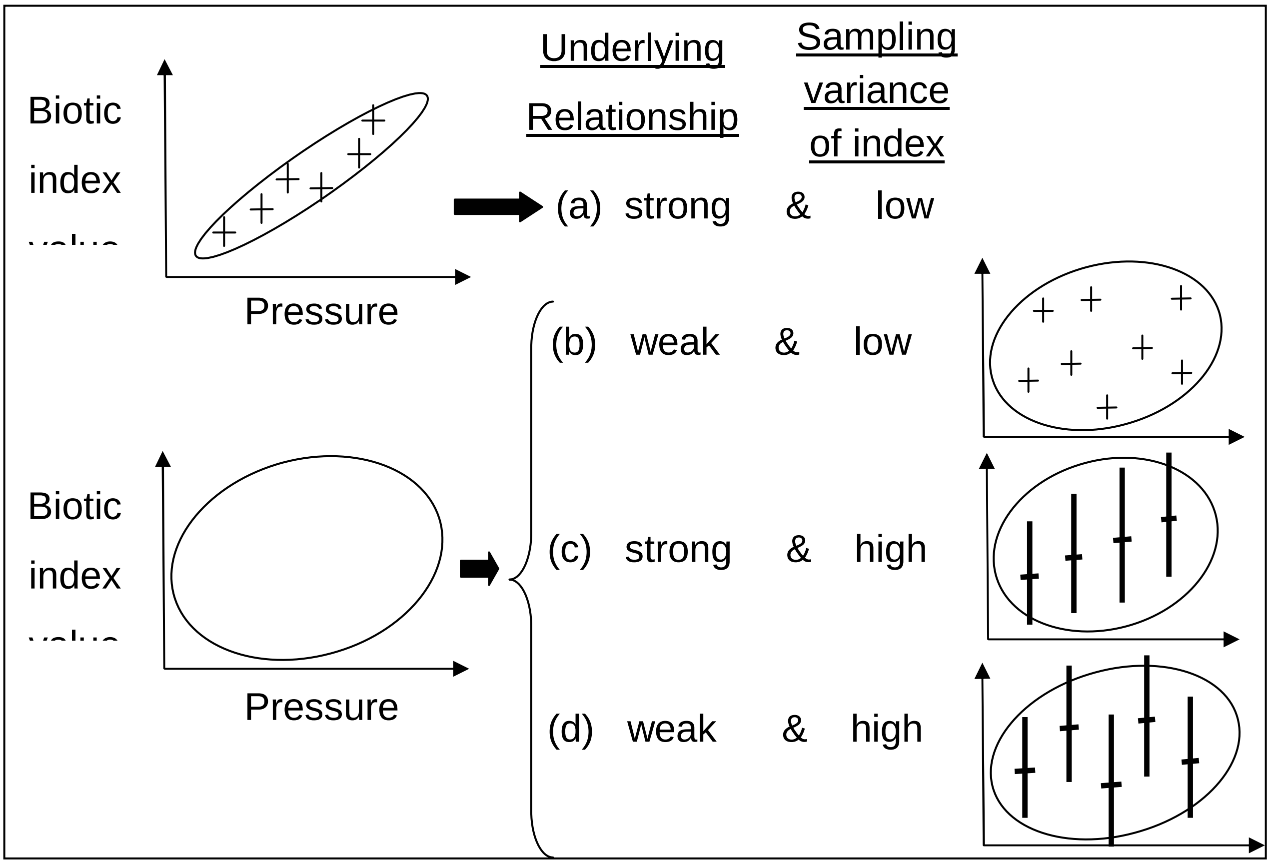


Fig.1 Relationships amongst sites between observed values of biotic index and pressure variable Differentiating three possible causes (b-d) of a poor relationship requires estimating sampling precision of biotic index (horizontal bar and vertical lines denote water body index true mean and sampling error respectively)

In their recent review of the achievements made in the first 10 years of the WFD, Hering et al. (2010) concluded that “Future challenges still remain, including the estimation of uncertainty in assessment results and a revision of rules in combining the results obtained with different Biological Quality Elements”. This current paper attempts to contribute towards improving understanding and assessment of sampling and other uncertainties in estimates of metrics, EQRs and WFD ecological status class and the implications for confidence and uncertainty of multi-metric and multi-BQE water body assessments.

2. Bioassessment uncertainty

It is useful to remember that an estimate of a water body status class is made, and assumed to apply, for both a specific area in space (i.e. the water body, lake, estuary or stretch of river) and for a period in time (the assessment period) which could be one day, one month, one season, one year or maybe, three or five years (in the case of national long-term surveillance monitoring).

2.1 Sources of bioassessment uncertainty

The total uncertainty and potential error in estimating the true status class for a WB for a period is due to the combined effects of:

(i) spatial biological variability within the WB

(ii) temporal biological variation within the assessment period

(iii) the choice of sampling/surveying methods and sampling personnel

(iv) the sub-sampling and sample processing protocols, including taxonomic identification and variation in expertise of personnel used

(v) errors in setting appropriate reference condition values due to limitations in the available reference sites' data and/or uncertainty in the predictive modelling of their biota-environment relationships

(vi) the choice of biological indices and the method of their conversion to EQRs

(vii) the choice of status class limits

(viii) the choice of multi-metric indices and class rules and/or multi-BQE rules.

It is useful to be remain aware that every methodological decision you make can affect the WB assessment and its true uncertainty. However, the practical way to progress is to acknowledge that there is no absolute true or correct WB classification protocol. We should therefore aim, at least initially, to assess the uncertainty in our bioassessments due to the spatial and temporal sampling variation and sample processing errors, conditional on the chosen overall WB classification protocol, namely conditional on the choice of BQEs, sampling/surveying methods, metrics, EQRs and status class rules. This is estimating the sampling precision of our chosen method. In one sense, the actual accuracy (variation about the "true" WB value) is unknowable as the true quality depends on which subjective aspects of the biota and which methods we use to define WB quality. We can only assess a form of accuracy by the strength of some form of correlation between our biotic measures and independent WB condition measures based on the extent of anthropogenic modifications and stresses operating at the water body (Johnson et al. 2006). However, the emphasis of the WFD approach is to base assessments on the biological rather than the chemical conditions. In the longer term we should try to compare different WB classification protocols (based on different methods, metrics, reference sites, BQEs and status class rules), to learn from their discrepancies and improve our bioassessment protocols.

Ideally, each potential source of uncertainty in our WB assessments should be scientifically assessed and quantified, either from suitable existing datasets, or else from new scientific studies specifically designed with replication at the appropriate spatial and temporal field scale, replication of any sub-sampling, and/or use of multiple personnel to assess the extent of inter-operator effects metric and status class variability for a water body.

This can help us revise our monitoring sampling design for each water body and our overall bio-assessment methodology to improve the precision and/or cost-effectiveness of our monitoring scheme. If we understand which sources of variation make substantial contributions to the uncertainty associated with an assessment, monitoring strategies can be designed to reduce this uncertainty and hopefully give an acceptable level of confidence.

3. Estimating spatial and temporal sampling variances

The WFD status class for a waterbody is usually based on the one or more metrics from one or more BQEs. For any one metric, the metric value (and derived EQR) used for a waterbody is usually based on some form of average of the metric values from the samples obtained from that waterbody over the assessment and monitoring reporting period.

In order to derive estimates of the sampling uncertainty for the estimated metric (and EQR) value for a water body, it is necessary to have estimates of the various sources of spatial, temporal and sampling processing errors relevant to that WB and period of time.

3.1 Example using macroinvertebrate datasets and metrics for UK rivers

As a real example illustrating the effects of spatial and temporal variability on the uncertainty of water body, I use results from an analysis of a combination of UK government environment agencies' UK datasets for river macroinvertebrate samples based on the RIVPACS sampling and sampling processing procedures (Murray-Bligh, 1997) and the RIVPACS bioassessment system(Clarke et al. 2003) (Table 1).

For WFD reporting purposes, the UK agencies propose reporting the river quality for each river water body (i.e. cohesive river stretch) as the average quality over a three year period. The RIVPACS approach uses a predictive statistical model of the macroinvertebrate-environment relationship between UK-wide reference sites to set site-specific expected (E) values for each macroinvertebrate metric which are then compared with the observed (O) values as EQR (O/E) ratios. The expected values are also season specific to allow for natural variation in macroinvertebrates between (RIVPACS) sampling seasons (Feb-May, June-Aug, Sept-Nov) (Clarke et al. 2003). Therefore, the sources of variance in the observed (O) values of metrics which affect the sampling variance of river water body average quality over a three-year period are replicate sampling variability at the same site on the same day (σ_R^2), spatial variability between sampling sites within the water body (σ_S^2), within-season (σ_W^2) and between-year-within-period (σ_Y^2) temporal variability (Table 1). There is also a potential spatio-temporal interaction variance.

Although the datasets are the best available for the UK, they are not ideal as no single dataset enables us to estimate all of the above variance components across a wide range of water bodies (Table 1). However, it was possible to fit a statistical mixed model to the combined datasets involving (assumed constant) average values for each variance component while allowing for (fixed effect) differences between combinations of water bodies, seasons and periods. The mixed models were fitted using the REML (Residual Maximum Likelihood) procedure within the Genstat statistics package, but can also be fitted using the lme and lmer mixed model functions in the R programming language.

Table 1. Spatial, temporal and replicate sampling data structure of UK river sites datasets, together with the sampling uncertainty variance components which can be estimated from each dataset.

Dataset	Sampling structure			
BAMS	16 study sites (4 qualities x 4 physical types) x 3 seasons x 3 replicates (1^{st} & 3^{rd} operator A, 2^{nd} operator B) – one year only			
TAY	28 (mostly good/high quality) sites in Tay region of Scotland x 4 replicates x 2 seasons per year for most years over period 1988-1997			
SEPA	418 Scottish SEPA sites (from high to bad quality) sampled in each of 2-3 seasons per year over period 1990-2004 (181 cases of 2 samples on different days in same season)			
DOVE	Dove Catchment in Central England : 5 WFD waterbodies with 3, 2, 3, 1 and 1 monitoring sites per water body (moderate to high variable quality), mostly one sample in spring and autumn for most years 1983-2007			
	Variance components (Y denotes dataset contributes data to estimation)			
	Replicate	Within-season temporal	Inter-year within 3-yr period	Between site within WB Spatial
BAMS	Y			
TAY	Y		Y	
SEPA		Y	Y	
DOVE			Y	Y

Further details of the datasets and the approaches used to estimate variance components are given in Clarke (2009). The variance component models were fitted to each of the two macroinvertebrates indices which are currently used for national assessments and monitoring of UK rivers, namely the number of BMWP (Biological Monitoring Working Party) families present (NTAXA) and the BMWP Average Score Per Taxon of the families present (ASPT). The richness metric NTAXA was analysed on the square root scale as Clarke et al. (2002) showed that this transformation removed the tendency for replicate sampling variance of NTAXA to increase with the replicate mean NTAXA value for a river site and thus made sampling variance independent of site type and quality, supporting the use of single variance component estimates for all river sites (when based on the RIVPACS sampling protocol).

Table 2. Estimates of replicate, temporal and spatial sampling variance components of macroinvertebrate metrics (ASPT and the square root of NTAXA) based on a combination of UK river sites datasets

	Variance estimate (% of total in brackets)			
	Replicate	Within-season temporal	Inter-year within 3-yr period	Between site within WB (Spatial)
Metric	σ^2_R	σ^2_W	σ^2_Y	σ^2_S
√NTAXA	0.0576 (37%)	0.0350 (22%)	0.0365 (23%)	0.0266 (17%)
ASPT	0.0654 (28%)	0.0596 (26%)	0.0209 (9%)	0.0873 (37%)

For these two indices, on average, roughly one-third (28% ASPT, 37% NTAXA) of the total variance in values which occurs within a water body over a three year assessment period is due to simple variability in values between replicate samples taken at any single sample site on a single day (Table 2). The RIVPACS macroinvertebrate sampling protocol used is a multi-habitat fixed time sampling method with no sub-sampling for the identification of taxa present; using a less reliable method might lead to greater inter-replicate variability. Based on this limited analysis, spatial variability between sites within a WB is a greater source of the total within-WB sampling variance for ASPT (37%) than for NTAXA (17%), suggesting that the type of taxa (i.e. their nutrient stress tolerances and BMWP scores) varies relatively more between possible sampling sites within a water body than the macroinvertebrate taxonomic richness.

3.2 Sampling precision of water body biological metric values

Having estimates of the various variance components for indices enables us to assess which metrics (and sampling methods) are most susceptible to sampling variability. For example, within the Europeam FP5 STAR project (Furse et al. 2006), Clarke et al. (2006,a,b) estimated the replicate sampling (and sub-sampling) variance as a percentage (P_{samp}) of the total variance in metric values across all samples and sites of varying quality within a WFD stream type for each of a wide range of European macroinvertebrate sampling methods and stream types; metrics and methods with relatively low P_{samp} have higher sampling precision and greater potential to provide reliable measures of river status class.

The variance component estimates for the selected metrics can be used to estimate the typical sampling precision obtained with each of a range of sampling regimes for a WB monitoring scheme. This can help design the most cost-effective sampling scheme for assessing and monitoring ecological quality using these metrics.

Consider the previous UK rivers macroinvertebrates example, where the WFD assessment is based on the average quality and thus average index values for a water body over a three-year period. If a sampling scheme involves taking r replicate samples at each of s sampling sites on each of w dates within the sampling season in each of y years (1, 2 or 3) within the 3-year assessment period, then the sampling variance (σ_M^2) of the mean metric value is :

$$\sigma_M^2 = \sigma_R^2 / rswy + \sigma_W^2 / wy + \sigma_Y^2 (1 - y/3)/y + \sigma_S^2 / s \qquad \text{(Eq. 1)}$$

There are only three years in any one WFD reporting period and therefore if samples have been taken in all three years, the years are effectively temporal statistical strata and the variance between years does not influence the sampling precision of the WB period mean. When sampling has not occurred in some years, the estimate of average quality for the whole period is potentially biased as some 'strata' have not been sampled. However, within the framework of using variance components, this extra uncertainty has be included as a term involving typical inter-year within-period variance but with a finite population correction (1-y/3) to allow for the fraction of all (i.e. 3) years sampled (Cochran, 1977). A similar logic applies to spatial stratification of a WB into zones or area-defined habitat types. For example, if a lake has been sub-divided into sections (such as near, mid and far-shore, or shallow, mid-water and deep sampling zones) and one or more samples taken from each section, then the sections are effectively statistical strata and the sampling variance of the sample mean metric value for the whole lake does not depend on the variance in metric values between sections if they have all been sampled, but it does depend on the variance between sites within each section. However, it is important to have some understanding of how much variability in the biota and metric values occur between spatial sections and (potential) strata relative to other sources of sampling uncertainty, as this will guide whether the water body sampling scheme can benefit from using such sections as statistical strata.

3.3 Confidence of status class depends on precision of sampling scheme

It is misleading to say "this is (definitely) the class of this water body". It is more realistic to say "we estimate these are the probabilities of this water body being of each status class based on our sampling/survey design for this assessment period and this assessment method". Given the WFD goal for Member States to achieve good or better ecological status for all water bodies (ideally by 2015 but with possible extension to 2027, (Hering et al., 2010)), then it is especially useful to have estimates, based on our monitoring and assessment scheme, of the confidence (i.e. probability P_{pass}) that each WB is of good or better status or conversely the confidence ($P_{fail} = 1 - P_{pass}$) that the WB failed to achieve good status. With limited resources for remediation measures, it is important to concentrate efforts on improving those water bodies for which we are most confident the ecological status is inadequate, within the practicalities of any river basin management plan. The confidence with which we can assign a water body to a WFD status class and the likelihood of failing to achieve good or better status are dependent on the accuracy with which we can estimate the WB mean values of the metrics (and EQRs) involved in the chosen bioassessment method. This depends heavily on the precision of our sampling scheme for the water body over the assessment period (Table 3). Taking more replicate samples from the same site on the same day, although the cheapest form of replication, only reduces uncertainty due to small-scale spatial heterogeneity between samples from the same site on the same day. At the other extreme, if the sole aim was to estimate average WB quality over the three year period, a statistically efficient strategy might be to take a sample from one (or more) different site(s) in each year, as this provides some spatial and temporal coverage replication (scheme 4 in Table 2) even though with such a scheme we cannot identify the relative importance of spatial and temporal variability in the observed metric values.

Table 3. Illustrative example of how confidence of water body (WB) status class depends on sampling scheme (1-4) used to estimate WB metric mean values. Observed mean ASPT = 6.4, good/moderate class boundary ASPT value for this WB type = 6.0, variance component estimates as in Table 2; observed class is 'good', P_{fail} = probability true class is moderate or worse

	Replicate variance	Within-season temporal	Inter-year within 3-yr period	Between site within WB Spatial	Variance of mean	P_{fail}
	σ^2_R	σ^2_W	σ^2_Y	σ^2_S	σ^2_M	
Sampling	0.0654	0.0596	0.0209	0.0873		
Scheme	*r*	*w*	*y*	*s*		
1	1	1	1	1	0.2262	20%
2	3	1	1	1	0.1826	17%
3	1	1	3	1	0.1290	13%
4	1	1	1 different site each year		0.0562	5%

3.4 Uncertainty of class depends on the spatial and temporal scale of extrapolation

In many monitoring schemes for a river stretch or lake, it is often only possible to take a biological sample or survey at one site on one occasion on which to estimate ecological quality. The WFD (European Union, 2000, Annex V, section 1.3.4) recommends that for operational monitoring using macroinvertebrates, fish or macrophytes, sampling/surveying should be at least once every three years. The resulting estimates of ecological status are often implicitly intended to represent average quality over the three year period It is useful to realise that the confidence we can have that this is the true ecological status class diminishes with the area in space and period in time over which this assessment is used to represent quality, or more specifically average quality; as illustrated in Table 4.

Table 4. Confidence of estimated water body status class depends on spatial and temporal extent of application/extrapolation. Example of one sample taken at one site on one day with observed ASPT = 6.4, good/moderate class boundary ASPT = 6.0, variance component estimates as in Table 2; P_{Good} = Confidence of observed (Good) class or better

Spatial-temporal scale for assessment	Uncertainty variance	Estimated uncertainty Variance	P_{Good}
same site -same day	σ^2_R	0.0654	94%
same site – season average	$\sigma^2_R + \sigma^2_W$	0.1250	87%
same site – 3 year average	$\sigma^2_R + \sigma^2_W + \sigma^2_Y$	0.1459	85%
whole water body- 3 year average	$\sigma^2_R + \sigma^2_W + \sigma^2_Y + \sigma^2_S$	0.2332	80%

4. Uncertainty of Reference Conditions and conversion of metrics to EQRs

The WFD requires that the observed (*O*) values of metrics are standardised to EQRs, ideally on a scale 0-1. A general approach to achieving for this is by the standardisation :

$$EQR = \frac{O - E_0}{E_1 - E_0} \qquad \text{(equation 2)}$$

where E_1 = Reference Condition value (= value of metric for which $EQR = 1$)

and E_0 = value of metric for which $EQR = 0$

Any EQR values calculated from equation (2) which are negative are always reset to zero. The EQR could be a RIVPACS-type O/E ratio where E_1 is set a RIVPACS model-based site-specific expected value and E_0 is set to zero. When several EQRs are used to create a multi-metric index (MMI) by averaging their values, each EQR is forced into the range 0-1 by setting any EQR values greater than 1 to 1.

4.1 Potential sources of error in setting Reference Condition values of metrics

A wide range of factors can influence the errors in estimates or model-based predictions of the Reference Condition values (upper "anchor" value E_1) of each metric for the group of sites or water bodies to be assessed. These include:

(i) Inadequate information & knowledge

- Inadequate set of RC sites for all or some WB types
- Not involving all "relevant" environmental variables
  (e.g. WFD System A or B Types or predictive model variables)
- Not making optimum predictive model
  (e.g. RIVPACS type model v Neural Networks (e.g. from the EU PAEQANN project); mechanistic model functions/parameters)

(ii) Sampling variation in RC sites' sample data (SE of mean)

(iii) Inconsistent data

- Existing Data from different sampling methods/standards combined to set RC
- Test site's observed sample value and RC data values based on (partially) different sampling methods

Similarly, errors in estimating the lower "anchor" value (E_0) will also have implications for EQR values; this is especially important for multi-metric indcies, where individual metric EQR values are directly averaged prior to sub-division to status classes.

4.2 Method for setting Reference Condition values for a metric

Various methods of setting the Reference Condition (E_1) value of a particular metric for a particular site/waterbody or environmental-similar group of sites/waterbodies can be used depending the data available. Obviously the reference condition or high quality sites used to determine the E_1 values should be sampled in the same way as the samples for the sites being assessed. The following are several possible options in roughly decreasing order of preference.

i) If a suitable RIVPACS-type predictive model involving an adequate number of environmental similar reference condition sites is available, then the E_1 values are best based on RIVPACS-type site- and season- specific predictions of the expected fauna and metric values.

ii) In the absence of a RIVPACS model, if a suitable number of reference condition sites of an environmentally-similar type are available, the E_1 values can be based on the mean or median metric value for these sites.

iii) If a suitable number of an environmentally-similar type of 'high' quality sites (of uncertain reference condition) are available, the E_1 values can be based on the mean, median or perhaps an upper percentile (75% or 90%) value of the metric for these sites.

iv) If only a very small number of an environmentally-similar type of 'high' quality sites (of uncertain reference condition) are available, then the E_1 values cannot be reliable estimated and might be based using the maximum of the few values available. However, the maximum value is not a stable measure and increases with the number of sites on which it is based.

v) If no 'reference condition' or 'high' status sites are available then, some form of hind-casting or extrapolation to reference conditions will be necessary to provide appropriate values of E_1.

5. Ecological assessments based on multiple metrics and/or multiple BQEs

The WFD requires that the status class for any WB, whether river, lakes, transitional or coastal water, should be based on one or more metrics and EQRs derived from each BQE and then the overall WB assessment should be based on combining the individual BQE assessments. In particular, the WFD (European Union, 2000, Annex V, section 1.4.2 (i)) requires that the overall class for a WB "shall be represented by the lower of its values (classes) for the biological and physico-chemical monitoring results for the relevant quality elements". The choice of BQEs to involve depends on their perceived reliability in measuring and responding to changes in pressures on this type of water body. Part of this reliability is determined by the susceptibility of each BQE and metric to sampling uncertainty.

5.1 Consequences of uncertainty on use of 'Worst case' rules

The WFD prescribes use of the worst-case or "one-out all-out" (OOAO) rule, whereby the overall class for a WB is the worst of the classes based on each individual BQE. Although this may be logical as a precautionary rule in an ideal world where the status based on each BQE can be measured without error; in practice the inevitable uncertainty associated with the sample-based estimated class for each metric and BQE leads to problems of probable under-estimation of the true overall class. As a simple illustration, if the true mean value for a WB is just above the Good/Moderate (G/M) boundary when based on each of M indices, then for each index, there is roughly a 50:50 change that the sample mean value will be below the G/M boundary. In the worst case rule, the probability than all M sample mean index values will be above the G/M boundary is 0.5^M, so with say M=3 indices, the probability that the WB will be classified as moderate of worse is very high 0.875 ($1 - 0.5^3$) even though the true mean value on each individual index would classify the WB as Good or better (Table 4).

Table 4. Illustration of implications of sample variability on use of worst case (one-out-all-out) rule on multiple BQEs or indices (assuming sampling uncertainty of BQEs or indices is uncorrelated)

BQE (or index)	Probability sample estimated EQR is 'moderate or worse	Examples	
		(a)	(b)
B1	P_1	0.5	0.3
B2	P_2	0.5	0.3
B3	P_3	0.5	0.2
Worst case	$1 - (1-P_1)*(1-P_2)*(1-P_3)$	0.875	0.608

More importantly, even when the individual indices or BQEs true mean vales are well above the G/M boundary, using the worst case rule can still lead tot he overall WB class being more likely than not estimated as moderate or worse.

The precision of using a worst case rule or a multi-metric index can be reduced by adding an extra metric with relatively high sampling variance and low precision (for details see Clarke et al. 2006b)

Borja (2010) compared the use of the OOAO principle with an alternative existing integrated assessment, based on the same multiple BQE data (chemical, phytoplankton, macroalgae, macroinvertebrates and fish) for 14 transitional and 4 coastal water bodies in the Basque region of Northern Spain over the period 2002-08. He found that the OOAO method indicated a lower status class than the integrative method for 18% of 125 (WB by year) cases for coastal waters and 58% of 224 cases for transitional waters. Borja (2010) found that the majority of disagreements for transitional waters were due to the observed sample status class for macroalgae being lower than for the other elements noted that macroalgae were considered to have the lowest reliability. Re-assessments excluding macroalgae reduced the disagreements between OOAO and the Borja's integrated approach from 58% to 32%, with the OOAO method now showing greater agreement of general improvement in WB quality with time (Fig.2 in Borja 2010).

One solution to problem of implementing the OOAO rule with large-scale sampling variability might be to adjust individual index EQR class limits downwards; but then individual metrics will have less power to detect moderate or worse quality. This is a complex issue as the ideal adjustment might depend on the number of other indices or BQEs involved.

I suggest that a better approach might be to take the median of the classes based on the individual indices and/or individual BQEs to be used in the overall WB status classification. Such an approach is an option in the new WISERBUGS software (Clarke 2011) discussed below.

6 Estimation of SD of waterbody mean value using the R software package

As an example, suppose we have taken two (N_R = 2) replicates macroinvertebrate samples from each of three (N_S = 3) sites around a lake and our best estimates of the metric variance due to between replicate variability and between site spatial variability for this WB are V_R (say 0.48) and V_S (say 0.36) respectively, then the estimate of the uncertainty SD associated with the lake mean metric value across the six samples is the square root of:

$V_S / N_S + V_R /(N_R \text{ x } N_S) = 0.36 /3 + 0.48 /(2 \text{ x } 3) = 0.12 + 0.08 = 0.20$; thus SD = 0.447

and this would the estimate of the uncertainty SD of this metric for this lake required for input into the WISERBUGS software (see section 7) to assess confidence of status class.

If sampling scheme had involved taking all six sample from the same site (avoiding inter-site travel and equipment transport costs), then the uncertainty SD would be much higher:

$V_S / N_S + V_R /(N_R \text{ x } N_S) = 0.36 /1 + 0.48 /(6 \text{ x } 1) = 0.36 + 0.08 = 0.44$; thus SD = 0.663.

Sampling at a single site around the lake can never reduce the uncertainty SD below 0.6 (i.e. below the square root of the between-site variance of 0.36)

If the degree of sampling and spatial replication varies between sites and water bodies, then the formula for the estimated variance and thus SD of the WB mean metric value is more complex, but here we give a brief illustrative example of how these estimates can be obtained using the R software package

The software package R is freely available from www.r-project.org. This package has several routines which can be used to fit mixed models (that is those involving both 'fixed' level factors and 'random' level factors

If the estimates of variance components for a set of waterbodies are obtained by analysing their replicate, spatial and (maybe) temporal variability all together using a mixed model approach in the R software package using the routine *lme* or *lmer*, treating WB as a 'fixed' effect factor, then the estimates of both the mean and its SD for each WB (even with unequal replication) are automatically available in the model 'summary' as Fixed effects 'Values' and 'Std.Error' respectively.

Figure 2 illustrates the approach and mixed model output using R. It is an example involving a single metric '*BioIndex*' for each of 4 samples from each of 9 sites from each of 8 lakes (1-8), except for lake 5 which only had one sample from each of 6 sites. It shows how to specify the correct mixed model in the *R lmer* routine and how the *lmer* output for the lakes treated as a 'fixed' effect gives the estimate of the lake mean metric value (as 'value') and its SD (as 'Std.Error'). Notice the larger SE for lake 5, because of its lack of replication and fewer sampling sites.

The above WB mean values derived by *R* routine *lmer* (or *lme*) can be used directly as inputs into WISERBUGS as observed WB metrics values in the 'Observed metrics values file'.

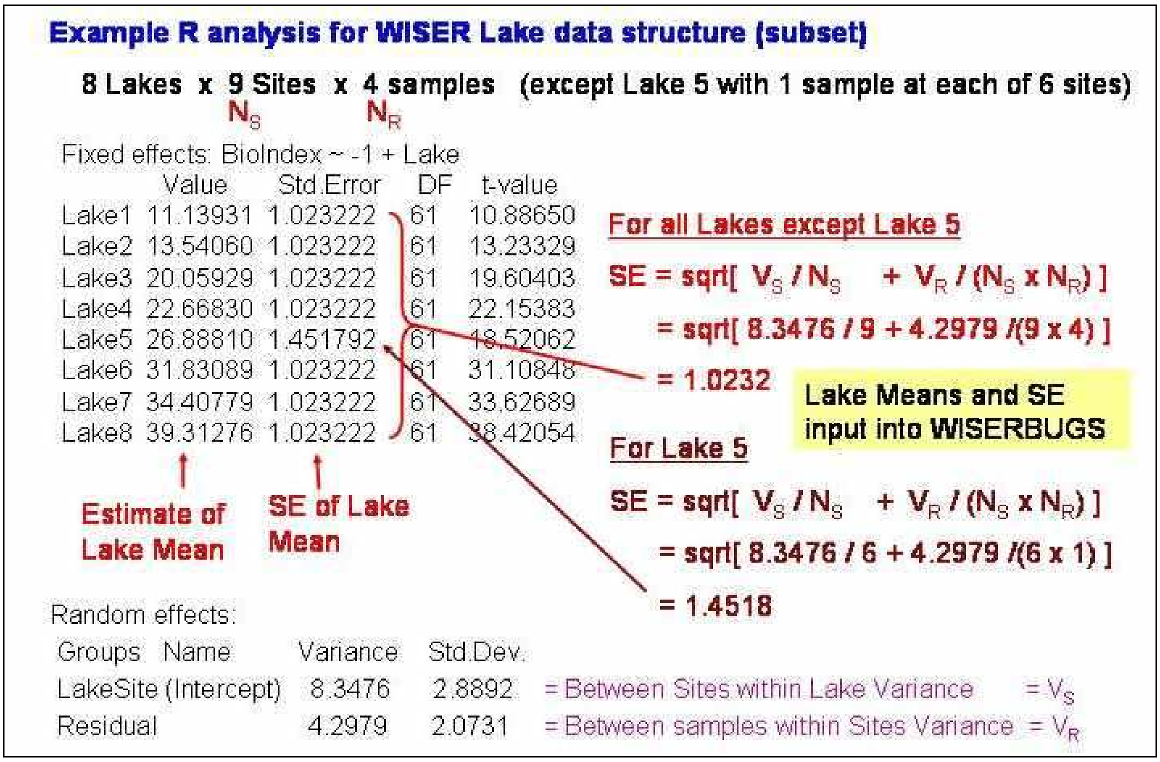
Also, the above estimates of the SD of the WB mean values derived by *R* routine *lmer* (or *lme*) can be used directly as inputs into WISERBUGS as the Uncertainty SD for that metric in column *I* of the 'Metric specification file'

Figure 2: Illustrative output from R mixed model analysis showing how to obtain estimates of the uncertainty standard deviation (SE) for a water body sample mean observed metric value

Example R analysis for WISER Lake data structure (subset)

8 Lakes x 9 Sites x 4 samples (except Lake 5 with 1 sample at each of 6 sites)

```
> #  Now fit Mixed model using 'R' software with Lake as Fixed factor (with value 1-8)

>#   and Site within Lake as Random factor 'LakeSite'

>#  'LakeSite' has unique value for each Site x Lake combination

>model1<-lmer(BioIndex~-1+Lake+(1|LakeSite)
>#              BioIndex ~ FIXED + (RANDOM)

>summary(model1)

Linear mixed model fit by REML
Formula: BioIndex ~ -1 + Lake + (1 | LakeSite)

Random effects:
 Groups   Name         Variance   Std.Dev.
 LakeSite (Intercept)   8.3476    2.8892   = Between Sites within Lake Variance      = VS
 Residual               4.2979    2.0731   = Between samples within Sites Variance  = VR
```


Example R analysis for WISER Lake data structure (subset)
8 Lakes x 9 Sites x 4 samples (except Lake 5 with 1 sample at each of 6 sites)
N_S
N_R
Fixed effects: BioIndex ~ -1 + Lake
Value Std.Error DF t-value
Lake1 11.13931 1.023222 61 10.88650
Lake2 13.54060 1.023222 61 13.23329
Lake3 20.05929 1.023222 61 19.60403
Lake4 22.66830 1.023222 61 22.15383
Lake5 26.88810 1.451792 61 18.52062
Lake6 31.83089 1.023222 61 31.10848
Lake7 34.40779 1.023222 61 33.62689
Lake8 39.31276 1.023222 61 38.42054
Estimate of Lake Mean
SE of Lake Mean
For all Lakes except Lake 5
$SE = sqrt[V_S / N_S + V_R / (N_S \times N_R)]$
$= sqrt[8.3476 / 9 + 4.2979 /(9 \times 4)]$
$= 1.0232$
Lake Means and SE input into WISERBUGS
For Lake 5
$SE = sqrt[V_S / N_S + V_R / (N_S \times N_R)]$
$= sqrt[8.3476 / 6 + 4.2979 /(6 \times 1)]$
$= 1.4518$
Random effects:
Groups Name Variance Std.Dev.
LakeSite (Intercept) 8.3476 2.8892 = Between Sites within Lake Variance $= V_S$
Residual 4.2979 2.0731 = Between samples within Sites Variance $= V_R$

7. General approach for assessing status class uncertainty using WISERBUGS

New software called WISERBUGS (WISER Bioassessment Uncertainty Guidance Software) has been written within the WISER project to provide a general means of using simulations to assess uncertainty and confidence in any estimates of ecological status class for water bodies based on either single metrics or a combination of metrics, multi-metric indices (MMIs) and multi-metric rules. The User provides prior estimates of the relevant sampling uncertainty for each metric and metric value to be involved in the water body assessments, together with metric status class limits and the rules for combining metrics into an overall water body assessment.

WISERBUGS is designed to be as generic as possible, so that it can be used with a wide range of metrics derived from field site sampling and survey data for any single or combination of biological quality elements (BQEs, namely phytoplankton, aquatic flora, macroinvertebrates and/or fish) and any type of water body (rivers, lakes, transitional or coastal waters).

The program requires the User to provide a 'Metric Specification File' in EXCEL format, in which they specify which metrics are to be used to determine the site or waterbody bioassessments, the individual metric uncertainty estimates and the multi-metric rules for combining information from individual metrics.

The uncertainty in the estimate of the (usually) mean value of a metric for a water body depends on the level of sampling replication on which it was based in terms of replicate sampling, spatial and temporal sampling coverage over the area of the water body to be assessed and the period of time for which the water body assessment is to apply. The estimates of uncertainty in individual metric values can include the sampling standard deviation (SD) due to sampling/sub-sampling variation and (optionally) the SD and bias due to sample sorting and identification.

In practice the uncertainty SD estimates for each metric for each water body or site to be assessed within WISERBUGS must be based on best-available information from replicated sampling studies on this or environmentally-similar water bodies.

The WISERBUGS software and User Manual is freely-downloadable from the WISER web-site (www.wiser.eu).

Acknowledgements

This paper is a result of the project WISER (Water bodies in Europe: Integrative Systems to assess Ecological status and Recovery) funded by the European Union under the 7th Framework Programme, Theme 6 (Environment including Climate Change) (contract No. 226273), www.wiser.eu.

References

Borja, A., 2010. Problems associated with the 'one-out, all-out' principle, when using multiple ecosystem components in assessing the ecological status of marine waters. Marine Pollution Bulletin 60, 1143–1146.

Clarke, R. T., 2009. Uncertainty in WFD assessments for rivers based on macroinvertebrates and RIVPACS. Environment Agency Science Report SC060044/SR4, Bristol. Downloadable from: http://publications.environment-agency.gov.uk/epages/eapublications.storefront/4d8cad0a00f44ada273fc0a8029606a7/Product/View/SCHO0409BPUE-E-E

Clarke, R.T., Furse, M.T., Wright, J.F. And Moss, D., 1996. Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. *Journal of Applied Statistics*, 23, 311–332.

Clarke, R.T., Furse, M.T., Gunn, R.J.M., Winder J.M And Wright, J.F., 2002. Sampling variation in macroinvertebrate data and implications for river quality indices. *Freshwater Biology*, 47, 1735–1751.

Clarke, R.T., Wright, J.F. And Furse, M.T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling*, 160, 219–233.

Clarke, R. T. And Hering, D., 2006. Errors and uncertainty in bioassessment methods – major results and conclusions from the STAR project and their application using STARBUGS. *Hydrobiologia*, 566: 433–439.

Clarke, R.T., Lorenz, A., Sandin, L., Schmidt-Kloiber, A., Strackbein, J., Kneebone, N.T And Haase, P., 2006a. Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics. *Hydrobiologia*, 566, 441–459.

Clarke, R.T., Davy-Bowker, J., Sandin, L., Friberg, N., Johnson, R.K. And Bis, B., 2006b. Estimates and comparisons of the effects of sampling variation using 'national' macroinvertebrate sampling protocols on the precision of metrics used to assess ecological status. *Hydrobiologia*, 566, 477–503.

Cochran, W.G., 1977. Sampling Techniques, 3rd edition, Wiley and Sons, New York.

European Union, 2000. Directive 2000/60/EC. Establishing a Framework for Community Action in the Field of Water Policy. European Commission PE-CONS 3639/1/100 Rev 1, Luxemburg.

Furse, M.T., Clarke, R.T., Winder, J.M., Symes, K.L., Blackburn, J.H., Grieve, N.J. And Gunn, R.J.M., 1995. *Biological assessment methods: controlling the quality of biological data. Package 1: The variability of data used for assessing the biological condition of rivers.* Bristol: National Rivers Authority, (R&D Note 412).

Hering, D., Moog, O., Sandin, L. And Verdonschot, P. F. M., 2004. Overview and application of the AQEM assessment system. *Hydrobiologia*, 516, 1–21.

Johnson, R. K., D. Hering, M. T. Furse & P. F. M. Verdonschot, 2006. Indicators of ecological change: comparison of the early response of four organism groups to stress gradients.Hydrobiologia 566: 139–152.

Murray-Bligh, J., 1997. Procedures for collecting and analysing macro-invertebrate samples. BT001, Environment Agency, Bristol.