# Comparability criteria for intercalibration phase 2

*Outline of the procedure*
*to compare and harmonise the national classifications of ecological status*
*according to the WFD intercalibration exercise*

*March 12th, 2010*

*Nigel Willby – University of Stirling, UK*
*Sebastian Birk – University of Duisburg-Essen, DE*

With contributions from
*Didier Pont, Olivier Delaigue (FR)*
*Mike Best, Geoff Phillips, Martyn Kelly, Ben McFarland, Steve Coates, Clare Scanlan (UK)*
*Jürgen Böhmer, Torsten Berg (DE)*
*Gert van Hoey (BE)*
*Henning Karup (DK)*
*Teresa Alcoverro Pedrola (ES)*

### *Summary of proposals*

- *The comparability of classifications is a dual product of the relative positioning of class boundaries and the variability in EQR values, either between pairs of countries or between countries and a common metric. Both aspects need to be considered when comparing classifications. This document establishes a series of steps for undertaking such comparisons and to achieve harmonisation through the analysis of quantitative datasets.*
- *Certain minimum requirements in terms of the size of datasets and the relatedness of national methods must be satisfied to undertake meaningful quantitative comparisons. Methods that cannot satisfy these requirements either do not share the same assessment concepts and should thus be revised, or may be better suited to comparison via qualitative approaches. In data-intensive methods the use of synthetic data, repackaging of samples or resampling of databases should be considered to increase data availability.*
- *A benchmarking procedure is proposed based on a population of sites identified by each country according to common standards for abiotic variables that are relevant to the pressure(s) being assessed. Application of national methods to benchmark sites allows typological or biogeographical differences between countries to be detected and then removed by normalisation. Without normalisation it is assumed that typological or biogeographical differences play no part in explaining differences in classification between countries.*
- *A mechanism for the benchmark normalisation of national EQRs within Option 3 is developed which enables methods to be applied appropriately to biological data of different nationalities held within a common dataset. The lack of normalisation within previous Option 3 comparisons is likely to have exaggerated differences between national methods when compared to comparisons undertaken using Option 2 in which benchmark normalisation has been routine. Biological data from national benchmark sites should be analysed via standard ordination approaches or with respect to biological common metrics and interpreted in the light of supporting environmental data to ensure that differences between MS can be adequately accounted for.*
- *A 'pseudo-common metric' is developed for use in Option 3 based on the average EQR of each combination of 'independent' countries. The pseudo-common metric behaves very similarly to the global mean EQR but has the advantage of being statistically independent of the country in question. This step opens the way for a regression approach to harmonisation in Option 3 and the establishment of a harmonisation band that will optimise the placement of class boundaries.*
- *The harmonisation band is a useful concept because it readily conveys the magnitude and direction of deviation of national methods from the global average view. However, improved statistical justification of the band is needed; the ± 0.05 threshold is arbitrary and has often been misinterpreted as equating to ± one quarter of a class. It is suggested that where >3 countries are participating in an exercise the harmonisation band is delimited by the 95% confidence limits of the mean boundary values of the different methods expressed on a common scale. With fewer countries ± 0.05 should form a default threshold. The centre of the band forms a harmonisation 'guideline' which would maximise the agreement between classifications.*
- *Direct comparison of commonly assessed sites using categorical classifications that respond to adjustments in class boundary values is recommended for measuring agreement amongst classifications. Variability in classifications has been neglected in the previous Option 2.*
- *Synthetic common datasets created using simulated EQR values with realistic error distributions based on normalised common metric versus national EQR relationships*

*allows the comparison of categorical classifications to be extended to Option 2 if no suitable common data sets can be assembled.*

- *The direct comparison of classifications (whether of real or synthetic cases) supports the generation of a standard set of statistically robust metrics, such as the multi-rater kappa coefficient, that are transferable and thus comparable across all Options.*
- *The proposed approaches consider only the upper class boundaries (i.e. the High-Good and Good-Moderate), in order to measure changes in the level of agreement at these boundaries pre- and post- harmonisation. To achieve a focus on a specific boundary classifications are aggregated into classes above or below that boundary (e.g. <good or >good at the Good-Moderate boundary). Agreement is then assessed by comparing kappa values based on (i) original class boundaries and (ii) revision of class boundaries to the harmonisation guideline or (iii) to within the range of the harmonisation band. This process avoids changes to class boundaries by individual countries that would produce only trivial gains in agreement of classifications.*
- *Generic thresholds for judging the acceptability of levels of agreement amongst methods post-harmonisation are not supported, although a realistic absolute lower limit is proposed based on the kappa coefficient. Differences between water body types, quality elements, IC types or GIGs in the maturity of methods, levels of standardisation of sampling, understanding of pressure-response relationships, approaches used in identification of reference sites, extent of taxonomic agreement, range of pressures assessed, and constraints on species distribution are factors that all demand a more flexible approach to judging acceptability.*
- *The presentation of harmonised upper class boundaries must be supported by an ecological characterisation that effectively maps the boundaries onto an ecological gradient that is defined using changes in relevant structural and functional attributes of the quality element.*

## 1 Definitions and concepts

Harmonisation = a state of agreement between national classifications that cannot be *significantly* improved upon by further changes to the position of their upper class boundaries (i.e. High-Good and Good-Moderate). At this point the degree of bias amongst classifications can be considered to be minimised with the result that all methods display a similar level of ambition. All countries can then be considered to have a similar interpretation of the thresholds of deviation from reference conditions that are contained in the normative definitions.

Extent of harmonisation = a numerical measure of the level of agreement between classifications that results following harmonisation of upper class boundaries. A multi-rater kappa coefficient is proposed to quantify this agreement. Indices used in the previous IC phase include the absolute average class difference between countries, although this is also sensitive to the level of agreement with respect to lower class boundaries. The extent of harmonisation can only be derived from direct comparisons of classifications on a common dataset using statistics appropriate for the analysis of categorical data. The distribution of predicted class boundary values on a common metric scale with respect to the global mean, as used in Option 2, is not a comparable statistic. This reflects the average level of bias in classifications, not the level of class agreement, and thus takes no account of inherent variability in the predicted values.

The extent of harmonisation is an emergent property of the level of variability in the relationship between pairs of values (Figure 1) and hence of the uncertainty in the positioning of a class boundary. Thus, it is emphasised that, even after harmonisation of class boundaries, the *extent* of harmonisation that is achieved might be unacceptably low due to persistently large scatter in the relationships between pairs of countries, or between a national method and the common metric. This issue does not apply in Option 1 since all countries share the same method, thus ensuring that differences in classification can only be due to the positioning of class boundaries.

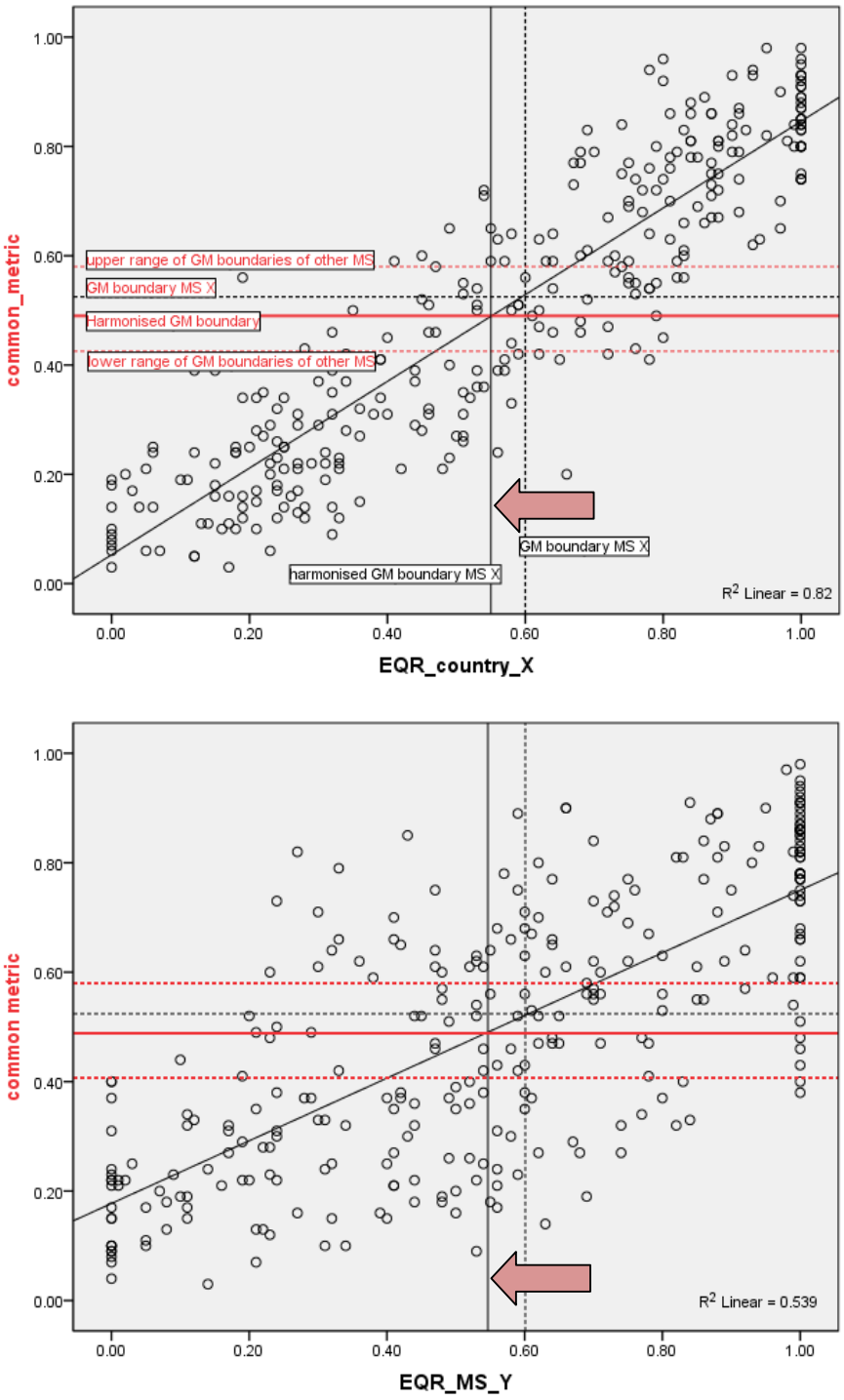



Figure 1: Harmonisation and levels of agreement. In both cases the country in question needs to adjust its Good-Moderate boundary. After this adjustment the class boundaries of both countries are equally harmonised. However, the level of variability is much lower in the case of country X (prediction standard error = 0.11) than Y (prediction standard error = 0.19) and X will consequently show a greater level of agreement with other classifications after adjustment of its boundaries. Option 2 has largely focused on boundary harmonisation while Option 3 has mainly considered levels of agreement. Both are equally pertinent to intercalibration.

## 2 Data considerations

Comparing the position of class boundaries on a common scale, and measuring the extent of harmonisation associated with these boundaries requires numerical datasets at either a national level (Option 2) or at a GIG level (Option 1 and 3). These datasets need to be as large as possible and to cover the widest possible gradient in ecological quality. All countries should contribute to common datasets drawing samples from the population of sites in a given IC type that cover the full range of ecological quality that is represented nationally. In Options 1 and 3 it will generally be difficult to proceed with a quantitative comparison using common datasets composed of <20-25 discrete cases that all countries are subsequently able to classify.

If countries employ relatively data-intensive methods to classify water bodies (e.g. based on the mean or some other property of the distribution of a set of spatially or temporally discrete samples, rather than classifying samples individually) this will limit the size of available datasets at a national or GIG level. In such cases, presenting data at a water-body level is also likely to compress the range of ecological quality that is represented since data aggregation is liable to concentrate values into the moderate-good range. In order to create datasets of sufficient size and that adequately cover the ecological gradient countries should not be constrained by presenting their classifications at a national water body level but should rather consider opportunities for repackaging data into smaller units (e.g. sub-water body level where shorelines or basins represent discrete units) that are still adequate for the application of all national methods. Other possibilities include the creation of synthetic water bodies based on randomised sampling of datasets composed of individual samples.

In Option 2, the size of national type-specific datasets should not be a major constraint provided that these datasets are large enough to capture the full range of ecological quality that exists within that country. If the ecological gradient covered by a national dataset is highly constrained, that country should consider the inclusion of data collected in other neighbouring countries using compatible methods in order to extend the available gradient.

Intercalibration exercises involving bilateral comparisons are preferably avoided. The merging (rather than splitting) of IC types is actively encouraged to increase the number of countries participating in a given exercise. This will also serve to increase the data available at a national or GIG level. Ordination analyses should be undertaken to demonstrate the existence of discrete biological types and to identify

those countries lying at the transition between types and who might then participate in multiple exercises. Where appropriate these analyses should extend to the cross GIG level to maximise the number of countries involved.

If insufficient quantitative data remains between countries attempting to intercalibrate their classifications the only remaining option is for some form of qualitative comparison based on direct ecological characterisation of class boundaries.

## 3 Steps of the procedure

All necessary steps to conduct the comparison and harmonisation of national status classifications are summarised in Table 1 and described in the following. Flow charts depicting these steps are given in Figures 6 and 7.

Table 1: Overview of process steps

| Step | Description |
|---|---|
| 1 | *Benchmarking* |
| 2 | *Benchmark Normalisation* |
| 3 | *Boundary translation using ordinary least squares regression* |
| 4 | *Progression requirements* |
| 5 | *Harmonisation guideline* |
| 6 | *Translation to national EQR* |
| 7 | *Assessing levels of agreement pre and post harmonisation* |
| 8 | *Optimising placement of class boundaries* |
| 9 | *Translation of normalised EQRs* |
| 10 | *Judging acceptability* |
| 11 | *Ecological characterisation of class boundaries* |

*Step 1 Benchmarking*

For all IC options at the start of each exercise it is necessary to benchmark the national biology in a given IC type in order to remove any sub-typological differences between countries. Without this step classifications are only directly comparable if there are no differences between countries in the biology of their reference (or other benchmark) sites. This may be the case for countries in close geographic proximity, but in most exercises, due to the size of each GIG, the breadth of conditions within an IC type, and the range of sampling methods used, it is likely that a gradient of ecological status will exist between countries in terms of the national view of reference condition (or an alternatively defined population of benchmark sites). Each country must nominate a set of national reference sites belonging to the relevant IC type that have been screened against agreed abiotic criteria that are relevant to the pressure(s) being assessed. Thus, for assessment of riverine eutrophication, for example, national benchmark populations could be

drawn from sites with annual mean orthophosphate concentrations of 20-50 µg/L. If a country employed a geographical analogue approach in establishing reference sites, and therefore used unimpacted sites from a different country as the basis for its method, these sites should be submitted for benchmarking. If no reference sites exist within a type, or for some participating countries, the benchmarking dataset must be reset to a lower threshold for which all countries can provide data (alternative benchmarking). The benchmarking process must be independent of national classifications (i.e. countries cannot simply nominate the sites they classify as high status as being their benchmark sites). Benchmark sites may, however, form a subset of the common dataset that is used in intercalibration.

Annex III of the IC Process Guidance provides further specification on how to derive reference and alternative benchmarks for intercalibration.

*Step 2 Benchmark Normalisation*

*Options 1 and 3*

Each country must apply its method to the benchmark dataset of every other country. This will establish where on the national scale the benchmark population of sites from each country lies. In applying national methods it is advantageous if the national EQR is allowed to extend over its natural range, rather than being artificially truncated at an upper value of 1. The median value of each benchmark population on the national scale must then be used to normalise the EQR of each country when its method is applied to sites from a different country (Figure 2). After this step the focus of intercalibration is strictly on the relative positioning of class boundaries since any effect of typological and biogeographical differences is minimised. If there is a shortage of benchmark data, geographically adjacent countries may elect to aggregate data to increase the confidence in the median value if there is no reason to expect significant typological differences to exist. Similarly, if there is no evidence that the median value of a national EQR differs between several countries, all of which can provide adequate data, a generic benchmark can be applied for this subset of countries. Biological differences between national benchmark populations should subsequently be explored with reference to biological common metrics or via ordination of biological data and such differences should be clearly accountable using supporting environmental data.

Construct a common dataset of examples of an IC type from the relevant GIG. All countries comparing methods should contribute data and should provide sites covering the full extent of the quality range in their country. Countries should apply

their methods to all data in the common dataset, normalising their national EQRs for each country according to the values established by the benchmarking exercise. This establishes the normalised national EQR (nEQR) to minimise typological or biogeographical differences between countries. The benchmark normalisation step also has the attraction of correcting for potential inequalities in the volume of data contributed to the common dataset by different countries.

Failure to carry out benchmark normalisation increases the risk that countries will make larger adjustments to their methods than are justified**.**

| Site | Data source | Benchmark | National EQR of Country A | National EQR of Country B | *Calculation: EQR of Country A* | *Calculation: EQR of Country B* | Normalised National EQR Country A | Normalised National EQR Country B |
|---|---|---|---|---|---|---|---|---|
| A01 | Country A | Yes | 0.97 | 0.80 | | | 1.11 | 1.00 |
| A02 | Country A | Yes | 0.82 | 0.78 | National EQR of A | National EQR of B | 0.94 | 0.98 |
| A03 | Country A | Yes | 0.87 | 0.99 | Median EQR of A at Benchmark Sites of A | Median EQR of B at Benchmark Sites of A | 1.00 | 1.24 |
| A04 | Country A | No | 0.54 | 0.65 | | | 0.62 | 0.81 |
| A05 | Country A | No | 0.73 | 0.34 | | | 0.84 | 0.43 |
| A06 | Country A | No | 0.34 | 0.21 | | | 0.39 | 0.26 |
| B01 | Country B | Yes | 0.89 | 0.99 | | | 1.00 | 1.27 |
| B02 | Country B | Yes | 0.76 | 0.76 | | | 0.85 | 0.97 |
| B03 | Country B | Yes | 0.99 | 0.78 | National EQR of A | National EQR of B | 1.11 | 1.00 |
| B04 | Country B | No | 0.43 | 0.54 | Median EQR of A at Benchmark Sites of B | Median EQR of B at Benchmark Sites of B | 0.48 | 0.69 * |
| B05 | Country B | No | 0.32 | 0.43 | | | 0.36 | 0.55 |
| B06 | Country B | No | 0.12 | 0.22 | | | 0.13 | 0.28 |

Normalise the national EQR
for each national dataset in the common database individually
by dividing the actual EQR value of the site by the median EQR value
at benchmark sites of the country providing the data.

* Example: 0.54 / 0.78 = 0.69

Figure 2: Example of the country-wise normalisation of national EQRs in IC Options 1 and 3

*Option 2*

Apply the common metric separately to each population of benchmarking sites submitted by each country. Take the median common metric value of each population of sites and use this to normalise the common metric values for each country. Thus if the median common metric value for the benchmarking sites of country A is 1.1 and for country B is 0.9 the common metric values for country A and B should be divided by 1.1 and 0.9 respectively. This establishes the normalised common metric (nCM). The opportunities for data aggregation between countries, as noted for Option 1 and 3, also apply here.

*Step 3 Boundary translation using ordinary least squares regression*

*Options 1 and 3*

All countries should apply their methods to the biological data and submit a benchmark normalised EQR (nEQR). The common dataset may also contain data from other countries situated within the GIG that do not currently have a compliant

method, providing that all the participating countries can apply their own methods satisfactorily to this data.

For each country in turn, using ordinary least squares regression (OLS), relate the nEQR to a 'pseudo common metric' (PCM) constructed by averaging the EQR values of the remaining countries. Thus, in an exercise involving five countries, for country A relate the nEQR of country A (the 'test' country) to the average nEQR of countries B, C, D and E (the 'assessors'); repeat for country B versus the average of countries A, C, D and E etc. If any one country is not significantly correlated with the average view of the countries ($p>0.001$) it must be excluded from the process. Recalculation of PCM values following exclusion of any one country is probably only necessary where there are small numbers of intercalibrating countries (<4), since any one country has a small influence on the PCM and the influence of the country that is removed is common to all combinations of PCMs in a given exercise. Prior to exclusion any such country must be given the opportunity to improve its method (e.g. by removal or re-weighting of component metrics) to achieve a satisfactory correlation with the average view of the other countries. The continued failure of a country to correlate significantly with the majority view indicates a major conceptual difference in assessment making it incomparable with other countries.

Take the model formula for each regression and determine the PCM value (i.e. the average EQR of the remaining countries) that equates to the upper class boundaries for each national method. Therefore, for country A, if $y = mx + c$ where $y$ = the PCM value, $m$ = the regression slope, $x$ = the EQR value of country A and $c$ = the regression intercept, derive the value on the PCM scale for values of $x$ corresponding to the High-Good and Good-Moderate class boundaries. Repeat this for each country's regression.

The PCM behaves in a very similar manner to the global mean nEQR of all countries, but has the advantage of being statistically independent of the country that is being tested (i.e. the y axis is truly independent of the x axis which is not the case when the common metric is composed of the global mean). The bias in the PCM between different combinations of countries reflects the relative precautionarity of individual countries. In the case of three countries the mean of any one country may have a more significant influence on the PCM value, especially if there are marked differences in bias between the three countries. In this instance the 'common' metric should be formed from the country that exhibits the largest correlation with the global mean of the three countries. Thus, if the

nEQR of country B has the largest correlation with the mean nEQR of countries A, B and C, the nEQR of country B will form the common metric. Country A and country C should then calculate where, on the scale of country B, their upper class boundaries would lay. A PCM is not required in exercises involving only two countries.

The purpose of this step is to obtain an unbiased estimate on the PCM scale of the boundaries of a national method. Thus, regressions should not be constrained to a simple monotonic form provided it can be shown that other linear models (e.g. polynomial) would lead to a statistically improved fit. However, most curvilinearity is expected to fall in the lower half of the quality gradient.

*Option 2*

Within each independent national dataset conduct an OLS regression between the national EQR and the nCM. In performing this regression it will be advantageous if the national EQR is permitted to cover its full natural range rather than being artificially truncated at a value of 1.0. The model should then be used to obtain the predicted values for each national EQR at the class boundaries on the nCM scale. Since the prediction error will be utilised at Step 7 to simulate sets of EQRs for use in cross-comparison of classifications it is desirable to constrain the prediction error by excluding major outliers from the national dataset. This step should only be performed after establishing the models used to predict class boundary values. Removal of outliers should not be used as a means of changing the basic national EQR versus nCM relationship. Regard should also be given to the section on data requirements.

National regressions with the common metric should not be constrained to a simple monotonic form provided that additional variables (e.g. quadratic transformations of the national EQR) can be shown to lead to a statistically improved goodness of fit. In general it is expected that most relationships will follow a simple linear form over the middle-upper part of the quality gradient.

*Step 4 Progression requirements*

For quantitative comparisons to progress beyond this point several checks are required on the performance of regressions established in Step 3.

*Option 1*

No tests are required since all countries apply the same approach; EQRs are therefore perfectly correlated. Regression is still required in order to translate the normalised national EQRs onto a pseudo common metric scale that can then be employed in harmonisation. We do not consider here the case in which all countries apply the same approach including the use of identical class boundaries.

*Option 2*

In an ordinary least squares (OLS) regression of the common biological metric(s) against each national EQR the relationship must be highly significant ($p<=0.001$) and meet the assumptions of normally distributed error and constant variance (homoscedasticity) of model residuals. The common metric must adequately represent all methods (preferably $r^2>0.5$ but absolutely $r^2>0.3$) which may therefore require the use of several different metrics in parallel. The slope of the regression should lie between 0.5 and 1.5.

*Option 3*

The regression between each method and the pseudo common metric must be highly significant ($p<=0.001$) and should conform to the standard requirements for normal distribution of error and homoscedasticity of model residuals. The slope of the regression should lie between 0.5 and 1.5. Under Option 2 and Option 3 trials with synthetic data indicate that countries with a regression $r^2$ of <0.3 between the national EQR and the common metric are liable to exhibit a level of agreement with other classifications that is so low as to force their removal at a later stage of the IC exercise. While such countries may technically have a highly significant correlation with the common metric (depending on sample size) when the $r^2<0.3$ it is strongly recommended that the national method is subject to further improvement (e.g. by removal or reweighting of selected metrics). It is similarly undesirable for the strength of the correlation between national methods and the common metric to display pronounced variation between participating countries. Thus, as a guideline the minimum observed $r^2$ value across a group of countries should be at least half the observed maximum $r^2$.

Fulfilment of these requirements means that harmonisation of national classifications is *potentially* achievable, it is not assured.

*Step 5 Harmonisation guideline*

By fitting national class boundaries to each of the nEQR versus PCM relationships (Options 1 and 3) or the EQR versus nCM relationships (Option 2) it is possible to

establish the predicted values on a PCM or nCM scale for each upper class boundary. This will yield the familiar spread of values seen in Option 2 and a midpoint represented by the global average of the predicted values (Figure 3). This mid point is considered to represent the harmonisation 'guideline'. The more closely national class boundaries approach this guideline the greater the resulting level of harmonisation of their classifications and the lower the level of bias between methods. This principal applies irrespective of the error associated with the projection of each national class boundary onto a common scale.

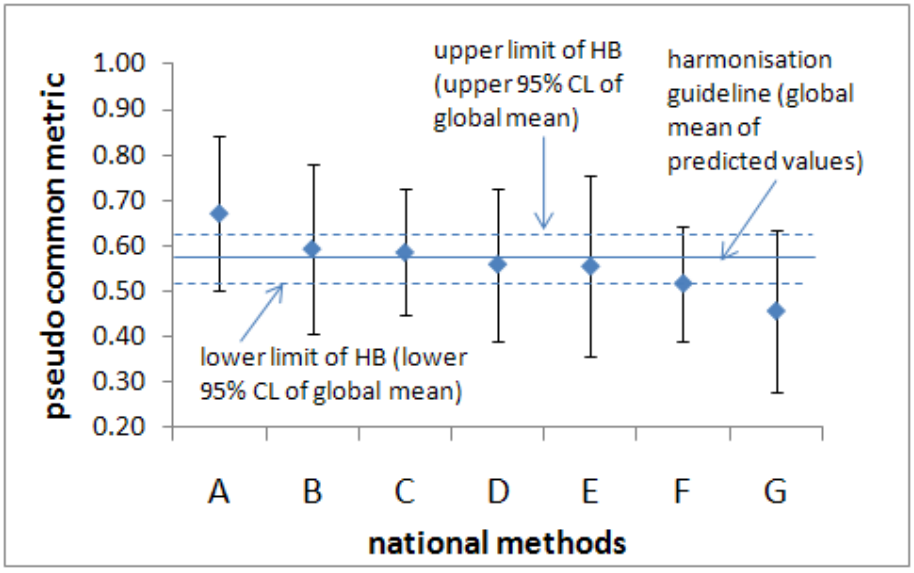



Figure 3: Anatomy of the harmonisation band. The harmonisation guideline, based on the global mean of all countries is 0.57. The 95% confidence limits are 0.05 meaning that the upper and lower limits of the band fall at 0.62 and 0.52 respectively. Two countries, A and G fall outwith the harmonisation band and are thus required to lower and raise their boundary accordingly. The values shown for each country are the predicted value (± standard error of prediction) on the pseudo common metric scale of the normalised national EQR that represents the Good-Moderate boundary.

Having established a harmonisation guideline value for each upper class boundary the concept of a harmonisation band could, in principle, be dispensed with. However, the harmonisation band was a prominent feature of intercalibration in IC phase 1, using Option 2. Moreover, a fixed guideline value will necessitate changes to class boundaries by all countries (unless they happen to fall on the global mean), yet in many cases the changes required are likely to be so small as to have a trivial effect on the level of agreement between classifications.

In IC phase 1 a harmonisation band based on the mean boundary ± 0.05 was applied in Option 2, the 0.05 being an arbitrary value designed to reflect some of

the uncertainties inherent within the IC process. It is important to note that this band width is an absolute value; it cannot be equated to a class width since 0.05 only equates to one quarter of a class if the mean width of the classes above and below that class boundary are both 0.2. In previous exercises this has not been the case; class widths have generally been less than this (and occasionally more) and the two classes either side of the boundary have not been of equal width. Establishing a Good-Moderate class boundary harmonisation band using the criterion of quarter of a class would also imply knowing the position of the Moderate-Poor boundary in order for the lower part of the Good-Moderate boundary harmonisation band to be correctly positioned.

The harmonisation band should preferably be a statistical property of the distribution of predicted class boundary values on the common metric (PCM or nCM) scale. An effective approach would be to use the upper and lower 95% confidence limits of the global mean to set the outer edges of the band. Countries would be expected to fall inside this band. In practice this is then only likely to require adjustments to class boundaries by a relatively small number of countries. The positioning of class boundaries by these outlying countries will account for the majority of disagreements in classification. In tests with synthetic data the 95% confidence limits have ranged between 0.02 and 0.08 but have generally been close to the original Phase 1 value of 0.05. The more countries that are involved in an exercise the greater the reliability of the mean and its confidence limits. When only two countries are involved the confidence limits of the mean will have little value for setting the band width. Consequently, when <4 methods are being compared it will be preferable to default to a harmonisation band width of ± 0.05.

It is important to note that the harmonisation band does not consider the uncertainty associated with predicted class boundaries on the common metric scale. The level of this uncertainty will influence the extent of harmonisation that is possible after adjustment of boundaries but a large uncertainty cannot be used to inflate the width of the harmonisation band.

*Step 6 Translation to national EQRs*

Having established the midpoint and upper and lower limits of the harmonisation band for each class boundary on the PCM (Options 1 and 3) or the nCM scale (Option 2) these values must be translated back into a national EQR that can then be employed to reset national classifications where necessary. The effect of making

this change on the level of agreement is then subsequently tested through the direct analysis of categorical classifications.

The average value of the boundary on the common metric scale, derived from all the regressions, therefore represents the guideline value, $y_h$, to which all countries must initially harmonise that boundary. The position of each country above or below this mean value indicates its relative precautionarity or bias with respect to the global view of that class boundary. Return to the national regression, invert the formula and determine where the national EQR would need to be positioned in order to achieve the guideline class boundary. Therefore for country A, $x = (y_h\text{-}c)/m$ where $y_h$ = the harmonised value on the PCM or nCM scale. Repeat this step using the upper and lower confidence limits of the harmonisation band in place of the global mean, in order to identify the range of values within which each predicted national class boundary should lie.

*Step 7 Assessing levels of agreement pre and post harmonisation*

*Options 1 and 3*

Populate a prepared spreadsheet with the nEQR values derived from the common dataset. Then apply the original class boundary values expressed on a normalised scale to generate a set of categorical classifications of the data.

Compare the level of agreement between classifications on a boundary-by-boundary basis using the multi-rater kappa coefficient. The kappa coefficient derives from Cohen (1960)[1] and was generalised by Fleiss (1971)[2]. It can be interpreted as a chance-adjusted measure of agreement between three or more raters, each of whom independently classifies a series of cases into one of a set of nominal categories. Kappa has long been recommended for use in social science and medical applications, such as psychological assessment and drug testing where there is a need to test agreement between experts using nominal classifications. Although the kappa coefficient is ideally suited to comparing agreement between WFD classification methods it has so far seen only limited use in intercalibration (e.g. Borja et al. 2007[3], Pont & Delaigue 2010[4]). The multi-rater kappa is not routinely calculated within standard commercial statistical packages, although there are various free online calculators and downloadable macros and

[1] Cohen. J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

[2] Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.

[3] Borja, A., Josefson, A.B., Miles, A., Muxika, I., Olsgard, F., Phillips, G., Rodríguez, J.G. & Rygg, B. (2007) An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. *Marine Pollution Bulletin*, 55, 42-52.

[4] Pont, D & Delaigue, O. (2010) Intercalibration process: kappa coefficient. Version 1.2. Cemagref.

scripts in R. The formula for kappa and the associated asymptotic standard error is also straightforward to set up within a spreadsheet which can then also be used to generate various supporting metrics such as the percentage of exact agreement and the bias of different methods.

Since the focus of intercalibration is only on the upper class boundaries it is appropriate to consider classifications at the simplest categorical level in order to compare boundaries. Pont & Delaigue (2010) describe the use of the multi-rater kappa coefficient in this specific context. Thus, to consider the Good-Moderate boundary the focus is on sites that are either <good (i.e. those classified as moderate, poor or bad) or >good (i.e. good or high). Similarly at the High-Good boundary the focus is on sites that are <high or >high (i.e. good-bad). When comparing classifications at this level of aggregation it is, however, important to note that datasets that are overpopulated with poor or bad sites (that will generally be classified as <good by all countries) may artificially inflate the level of agreement that exists between countries. Similarly an excessive proportion of high status sites may distort the apparent agreement at the Good-Moderate boundary. To avoid this cases that are compared should cover the full range of ecological quality.

Firstly obtain the multi-rater kappa coefficient and its 95% confidence limits for the High-Good and Good-Moderate boundaries using the unadjusted national class boundaries. Note the values of supporting statistics, such as the proportion of exact class agreement. Then substitute the original class boundaries with the nEQR boundaries that correspond to the harmonisation guideline value. This will generate an improved level of agreement between classifications although the increase in agreement that is seen will be constrained by the error in the relationships between countries or with the common metric (except Option 1 where there should be no error). The extent to which further change in boundaries is required will be determined by the value of the kappa value and its associated 95% confidence limit (Figure 4). If the confidence limits attached to the kappa value derived using the original boundaries overlap the confidence limits associated with harmonisation guideline boundaries then no change in boundary position is required; any change would only generate a trivial increase in the extent of agreement. If the kappa values associated with the original and guideline boundaries do not overlap then adjustment to the boundary positions is required (Step 8).

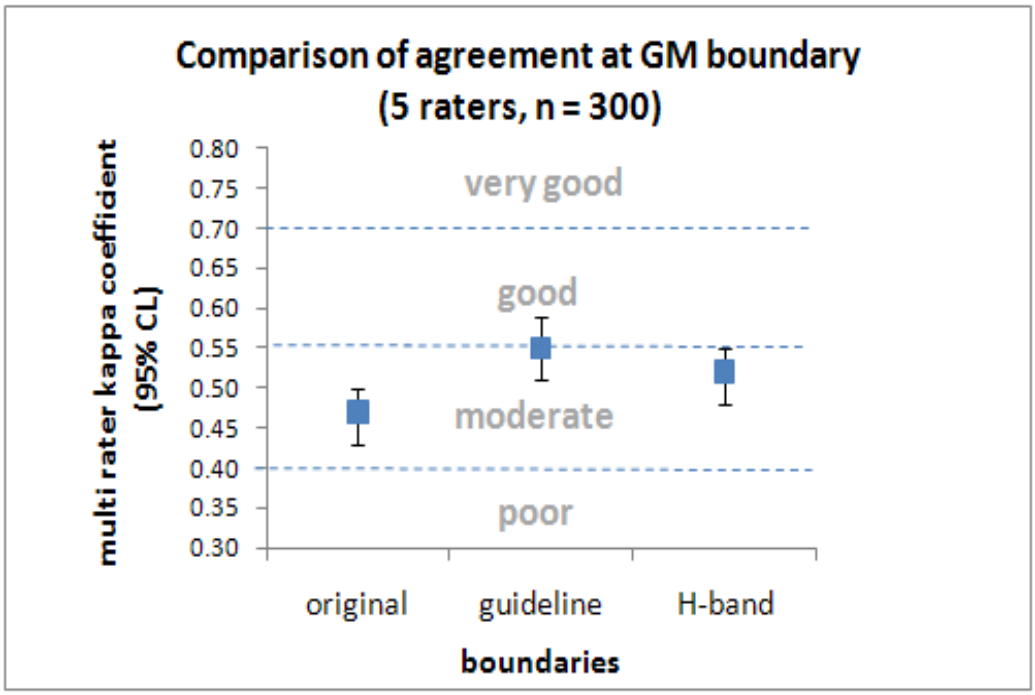



Figure 4: Agreement between classifications at the Good-Moderate boundary. Five methods were compared classifying 300 cases as <good or >good. With the original class boundaries a kappa coefficient of 0.47 ±0.03 was obtained. Through application of the class boundaries consistent with the harmonisation guideline a kappa coefficient of 0.55 ±0.03 was obtained. Further adjustment of the boundaries was therefore merited. The boundaries of the two outlying countries were then shifted to fall inside the harmonisation band producing a kappa value of 0.52 ±0.03 which would not be significantly different from that associated with the harmonisation guideline. Further change in the class boundaries would thus bring only a trivial improvement in agreement.

In this stage of the exercise, consideration should be given to removing small numbers of sites (<10%) from the common dataset that show persistently high levels of disagreement in classification between countries (e.g. a very wide standard deviation in nEQRs). This would be analogous to reducing the prediction error in national EQR versus nCM relationships described in Step 3 above for Option 2.

*Option 2*

The approach suggested here deals with the scenarios where a common dataset is either very small or lacking, or the methods used to acquire the data in the common dataset prohibit the application of all national methods to this data.

Generate a set of synthetic national EQR values for each country by evenly distributing the full range of possible values of the normalised common metric across 300 sites. Then generate a realistic set of EQR values for each country

using the relationship between the nCM and the national EQR for each country to which an offset is applied that is drawn randomly from the error distribution at each predicted point, where the error is normally distributed, has a mean of zero and a standard deviation equivalent to the regression prediction error (Figure 5). The random offset should be applied subject to the constraint that the product does not exceed the range limits of the nCM. This requires the use of software that can randomly generate values drawn from a specified distribution with known parameters. 'ZRandom' an Excel plug-in is suitable for this purpose, although there are a number of alternatives. Then follow the process as for Options 1 and 3 described above.

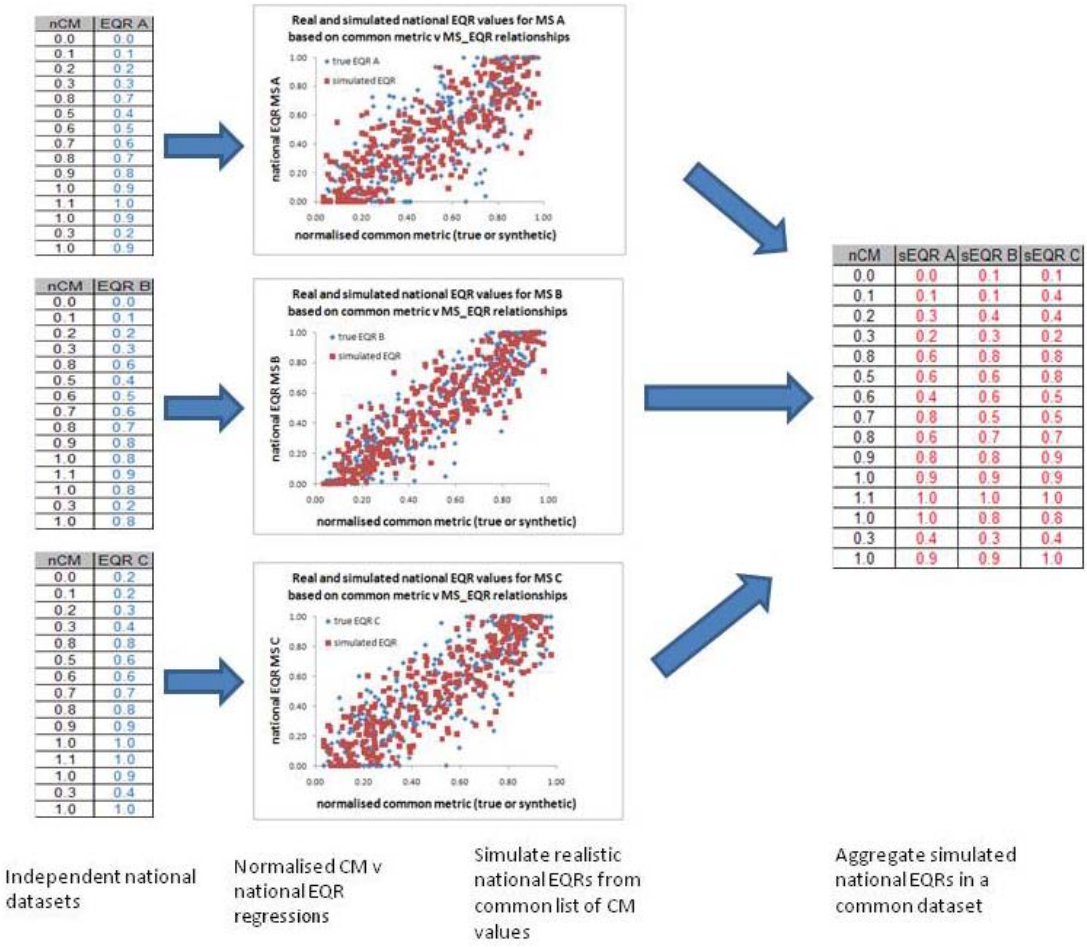



Figure 5: Generating synthetic common dataset for class comparisons in Option 2. True values of the national EQR in each dataset are shown in blue. The red values are modelled from a synthetic set of nCM values and achieve a realistic scatter by constraining them to the same error distribution as shown by the relevant national EQR versus nCM regression.

*Step 8 Optimising placement of class boundaries*

With reference to the harmonisation band move the class boundaries of those countries currently outside the band until they fall on the edge of the band. Re-calculate the kappa value. At this point the new kappa value will fall between the original value and the optimum value that would be obtained if all boundaries were moved to the harmonisation guideline (Figure 4). Unless the harmonisation band is unduly wide (e.g. > ±0.08) the new kappa value is likely to overlap the guideline value since the boundaries of those methods that generate the greatest disagreement have now been adjusted. At this point making further changes to the class boundaries would produce a negligible increase in agreement. If, however, the kappa value remains different from the guideline value the boundaries of the most outlying countries should each be shifted by 0.01 units towards the middle of the harmonisation band. Harmonisation is achieved when the kappa value does not differ significantly from the guideline value (i.e. the 95% confidence limit of the kappa values overlap). Report this value and supporting metrics, such as the average absolute class difference, maximum bias and percentage of class agreement, which can be used to assist in the communication of results.

The kappa value to be reported envisages that all outlying countries relocate their boundaries to fall inside the harmonisation band, *including* those countries that are precautionary enough for their boundary to fall above the band. Such countries are not compelled to harmonise their boundaries and the kappa value reported may be seen as hypothetical pending the decision of such countries on whether to lower their relevant class boundary. It should also be noted that while the lower class boundaries (Moderate-Poor, Poor-Bad) are not addressed here, adjustments to these may be required at a national level in the light of changes to upper class boundaries so as to ensure method integrity and compliance with the normative definitions.

*Step 9 Translation of normalised EQRs*

*Options 1 and 3*

Determine where on the original national EQR scale the nEQR value is equivalent to after harmonisation. To achieve this, the class boundaries must be multiplied by the median value of the population of benchmark sites belonging to that country (i.e. the value that was used in Step 2 to achieve the normalisation for the EQR of this country's sites).

*Option 2*

This step is not relevant since the national class boundaries have not themselves been modified, normalisation having been delivered in this case via the nCM.

*Step 10 Judging acceptability*

It is advocated that there is an absolute lower level of acceptability that applies to all exercises. Values of the generalised kappa coefficient below 0.4 are normally classified as representing poor or low levels of agreement (e.g. Monserud & Leemans 1992)[5]. Synthetic datasets indicate that such values of kappa are associated with large absolute average class differences (~1.5), and low levels of exact class agreement (<35%). Pont & Delaigue (2010) found using simulated adjustments to class boundaries that datasets with pairwise correlations of r <= 0.5 generally converged at kappa values below 0.4 indicating 'poor' levels of agreement amongst such countries, even after optimising boundary placement. On the other hand pairwise correlations of r >= 0.9 were required to achieve kappa values of ~0.7, a value normally regarded as the lower threshold for 'very good' agreement. Consequently, such levels of agreement will probably be outside the scope of most IC exercises.

Beyond this it is considered unrealistic to apply a generic standard across all quality elements in all water body types in all IC types and in all GIGs to determine if the extent of harmonisation that is achieved is acceptable. This reflects differences in various factors, such as the state of understanding of biology-pressure relationships between quality elements, the application of different sampling methods, the maturity of groups of classification methods or classification philosophy, the value of different quality elements for bioindication, stability of taxonomy, the ease of establishing reference conditions, the number of different stressors acting on a water body- or IC-type and the significance of biogeographical gradients. For example, it is reasonable to expect good harmonisation when comparing classifications of benthic invertebrates in small upland streams in northern Europe because there is a long history of using macroinvertebrates in biomonitoring, assessment philosophies are highly convergent, sampling methods are very similar and well established, the biogeographical gradients are short, taxonomy is generally agreed to high level of resolution, the range of different stresses in this environment is comparatively limited and contemporary reference sites are

[5] Monserud, R. & Leemans, R. (1992) Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*, 62, 275-293.

relatively easy to find in different countries. On the other hand, a lower level of harmonisation of classifications using phytoplankton in lowland base-rich lakes in central Europe might be tolerated since phytoplankton have had limited previous use in bioassessment, their taxonomy is poorly resolved, sampling methods vary, the biogeographical gradient is large and lowland base-rich lakes are exposed to multiple stressors and there are few, if any, reference sites remaining.

Acceptability must therefore be judged on a case by case basis considering levels of agreement that have been achieved by, for example, that quality element in other GIGs and other water body types.

## 4 Ecological characterisation

To support the quantitative comparison of class boundaries it is essential that GIGs produce a narrative for each IC type that clearly establishes where the harmonised High-Good and Good-Moderate boundaries lie in terms of their ecological characteristics. Common biological metrics are likely to be invaluable in this process. The outcome should be a mapping of upper class boundaries onto an ecological quality gradient that is defined based on relevant and readily understandable structural and functional attributes of each quality element.

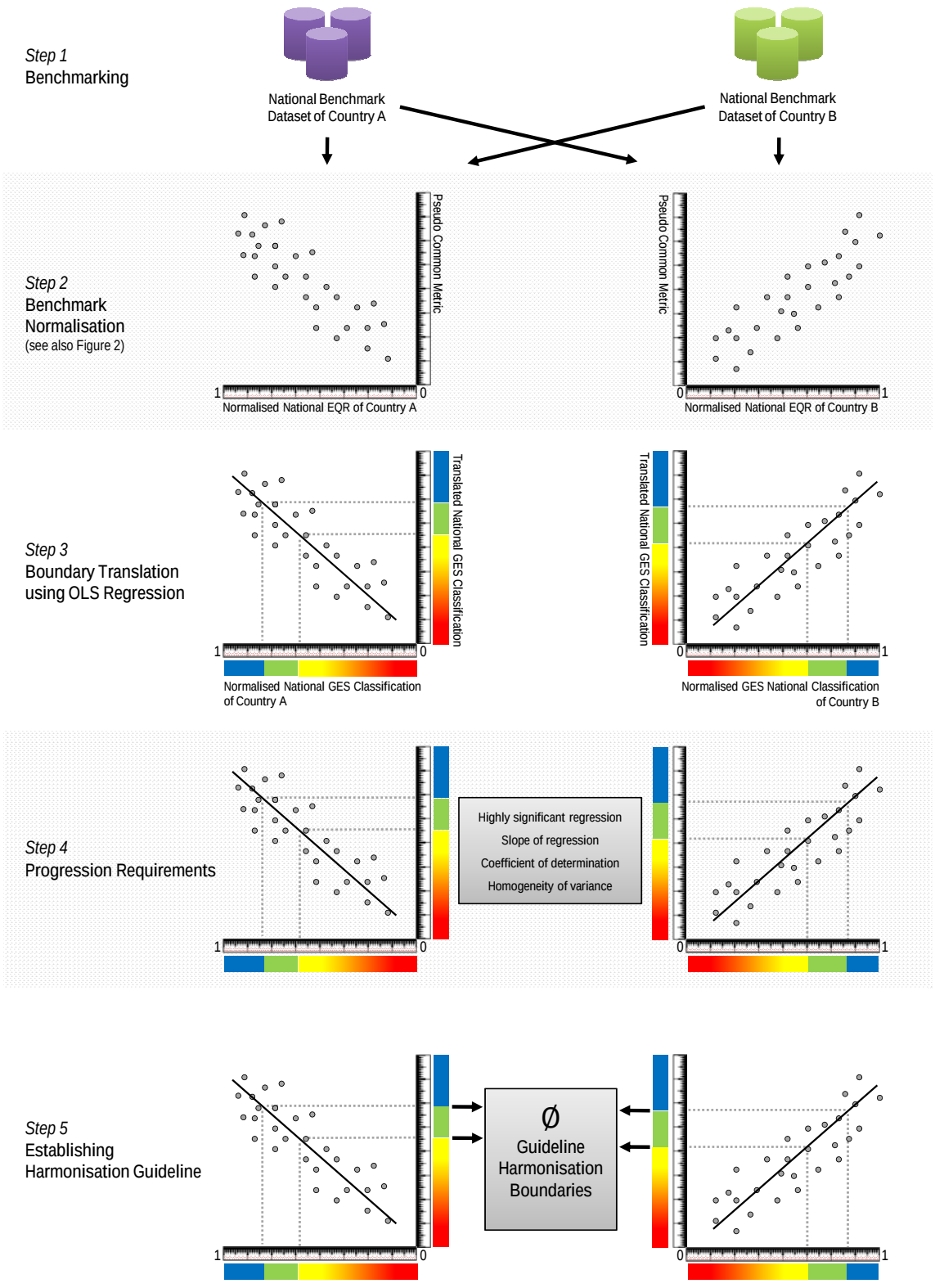
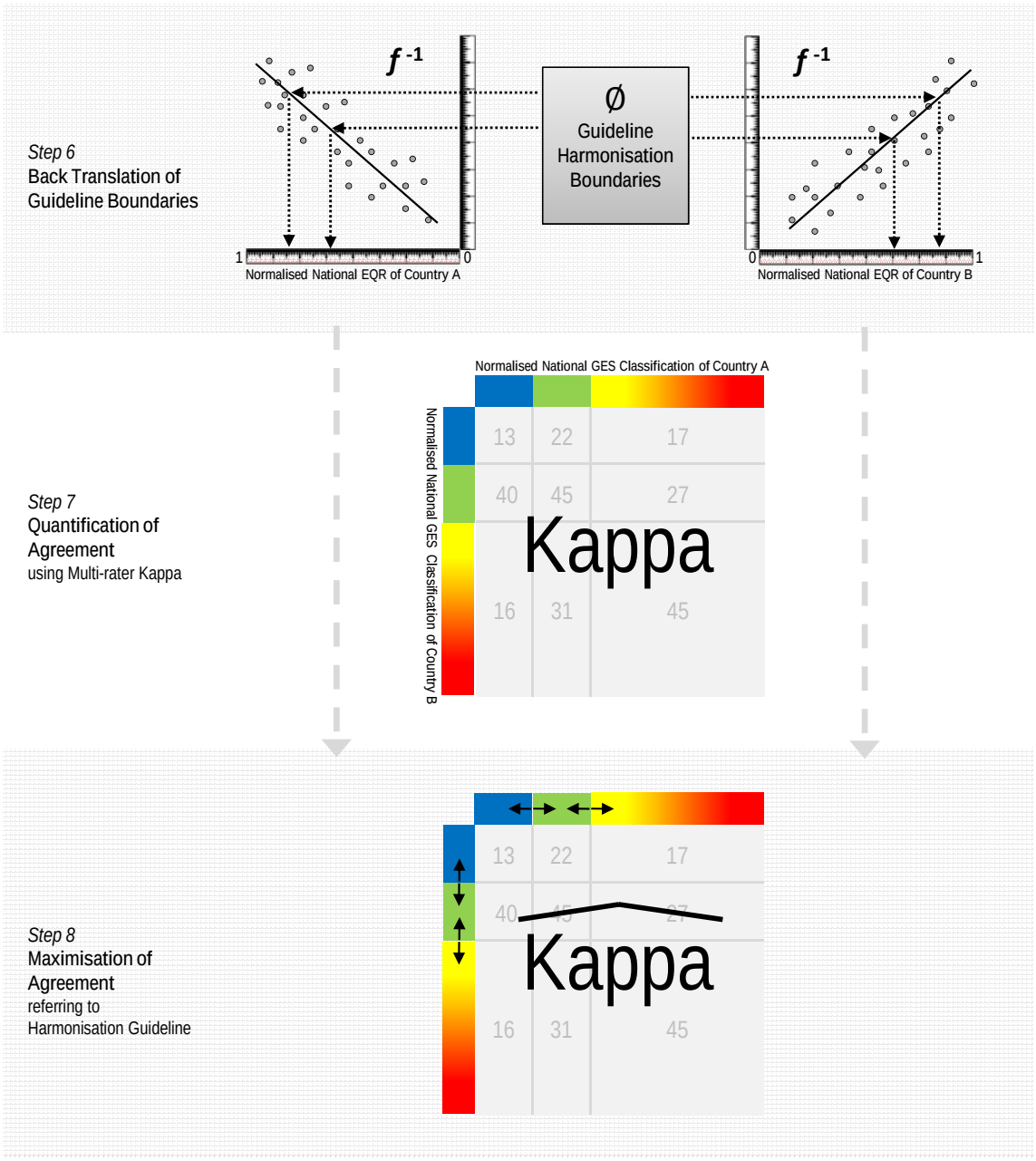



Figure 6: Process flowchart for the IC Options 1 and 3
(GES = Good Ecological Status; OLS = Ordinary Least Squares)




*Step 9*
Back Translation of
Normalised EQRs

$$\textbf{National Boundary EQR} = \frac{\text{Harmomised Boundary EQR of a country}}{\text{Median of Benchmark sites of this country}}$$

Figure 6 (continued): Process flowchart for the IC Options 1 and 3
(GES = Good Ecological Status)

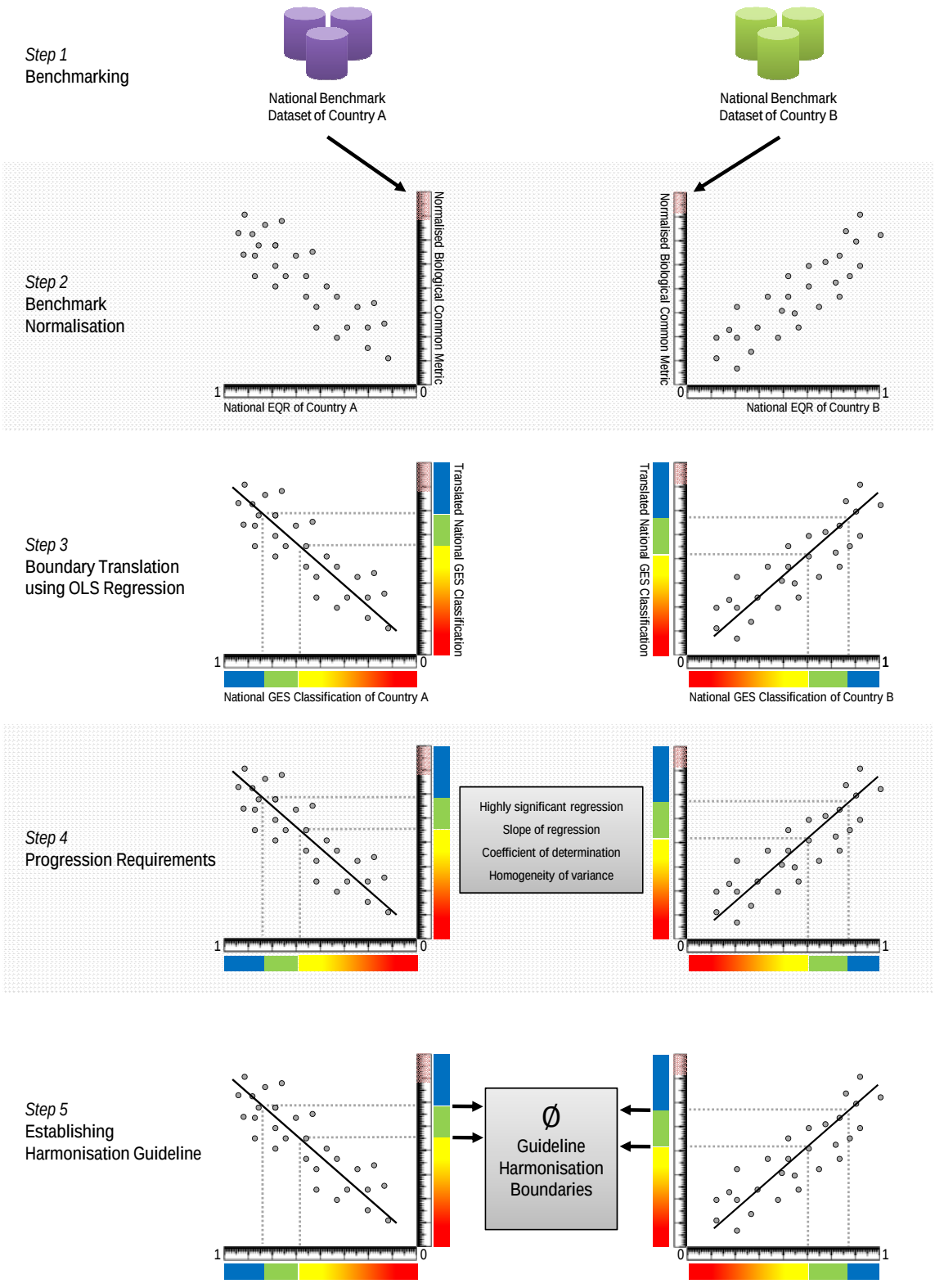



Figure 7: Process flowchart for the IC Option 2
(GES = Good Ecological Status; OLS = Ordinary Least Squares)

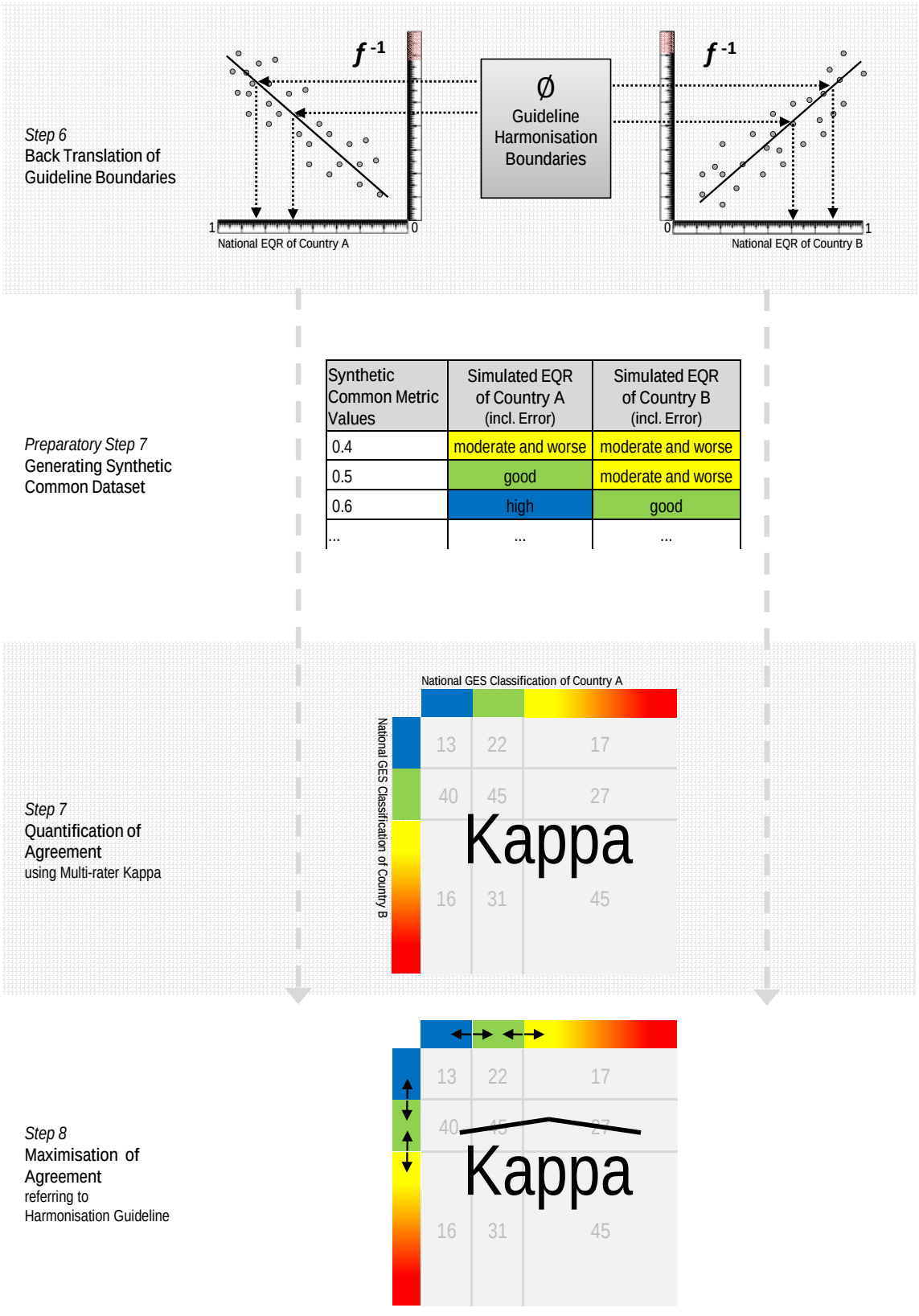



Figure 7 (continued): Process flowchart for the IC Option 2
(GES = Good Ecological Status)