Citation for published version:

Kelly, M., Bennett, C., Coste, M. et al. A comparison of national approaches to setting ecological status boundaries in phytobenthos assessment for the European Water Framework Directive: results of an intercalibration exercise. *Hydrobiologia* 621, 169–182 (2009). https://doi.org/10.1007/s10750-008-9641-4

Peer reviewed version

General rights:

A comparison of national approaches to setting ecological status boundaries in phytobenthos assessment for the European Water Framework Directive: results of an intercalibration exercise

Martyn Kelly[1*], Cathy Bennett[2], Michel Coste[3], Cristina Delgado[4], François Delmas[3], Luc Denys[5], Luc Ector[6], Claude Fauville[7], Martial Ferreol[6], Malgorzata Golub[8], Amelie Jarlman[9], Maria Kahlert[10], John Lucey[11], Bernadette Ní Chatháin[12], Isabel Pardo[4], Peter Pfister[13], Joanna Picinska-Faltynowicz[14], Christine Schranz[15], Jochen Schaumburg[15], Juliette Tison[3], Herman van Dam[16] and Sirje Vilbaste[17].

[1]Bowburn Consultancy, 11 Monteigne Drive, Bowburn, Durham DH6 5QB, U.K.

[2]SEPA, Greyhope Road, Torry, Aberdeen, AB11 9RD, U.K.

[3] REQUE Research Unit, Cemagref - Bordeaux Centre, 50, Avenue de Verdun, Gazinet, France.

[4]Departamento de Ecología y Biología Animal Facultad de Biología Campus Lagoas-Marcosende 36200 Vigo, Spain.

[5]Research Institute for Nature and Forest, Kliniekstraat 25, 1070 Brussel, Belgium.

[6]Centre de Recherche Public – Gabriel Lippmann, 41 rue de Brill, L-4422 Belvaux, Luxembourg.

[76]Laboratory of Freshwater Ecology - URBO FUNDP Rue de Bruxelles, 61 B-5000 Namur, Belgium.

[8]Institute of Environmental Protection, ul. Kolektorska 4, 01-692 Warsaw, Poland.

[9]Jarlman HB, Stora Tvärgatan 33, SE 22352 Lund, Sweden.

[10]Department of Environmental Assessment, Swedish University of Agricultural Sciences, POBox 7050, SE 75007 Uppsala, Sweden.

[11]Environmental Protection Agency, Butts Green, Kilkenny, Ireland.

[12]Shannon River Basin District Project, Mulkear House, Newtown Centre, Annacotty, Co. Limerick, Ireland.

[13]ARGE Limnologie, Hunoldstr. 14, A-6020 Innsbruck, Austria.

[14]Institute of Meteorology and Water Management, Wrocław Branch, Parkowa 30, 51-616 Wrocław, Poland.

[15]Referat 8.4, Bayerisches Landesamt für Umwelt (Bavarian Environmental Agency), Lazarettstraße 67, 80636 München, Germany.

[16]Consultant for Water and Nature, P.O. Box 37777, NL-1030 BJ Amsterdam, Netherlands.

[17]*Estonian University of Life Sciences (former Estonian Agricultural University), Riia 181, 51014 Tartu, Estonia.*

*Corresponding author. Telephone: +44 191 377 2077.
Email address: **MGKelly@bowburn-consultancy.co.uk** (M. Kelly)

Abstract

The European Union (EU)'s Water Framework Directive (WFD) requires that all Member States participate in intercalibration exercises in order to ensure that ecological status concepts and assessment levels are consistent across the EU. This paper describes one such exercise, performed by the countries in the Central / Baltic Geographical Intercalibration Group stretching from Ireland in the West to Estonia in the East and from the southern parts of Scandinavia to the northern regions of Spain and Italy (but excluding alpine regions, which were intercalibrated separately). In this exercise, methods used to measure ecological status of rivers using benthic diatoms were compared. Ecological status is estimated as the ratio between the observed value of a biological element and the value expected in the absence of significant human impact. Approaches to defining the 'reference sites' from which these 'expected' values were derived varied from country to country. Minimum criteria were established as part of the exercise but there was still considerable variation between national reference values, reflecting typological differences that could not be resolved during the exercise. A simple multimetric index was developed to compare boundary values using two widely-used diatom metrics. Boundary values for high/good status and good/moderate status set by each participant were converted to their

equivalent values of this intercalibration metric using linear regression. Variation of ± 0.05 EQR units around the median value was considered to be acceptable and the exercise provided a means for those Member States who fell significantly above or below this line to review their approaches and, if necessary, adjust their boundaries.

Key words

Diatoms, algae, monitoring, intercalibration, eutrophication, pollution

Introduction

The Water Framework Directive (WFD: European Union, 2000) establishes a framework for the protection of all waters (including inland surface waters, transitional waters, coastal waters and groundwater) in Europe. The environmental objectives of the WFD require all surface water bodies to achieve “good ecological status” (defined as having a biota consistent with only slight alterations from that expected in the absence of human impacts) by 2015. Each Member State (MS) has had to establish methods for assessing ecological status for a range of biological elements, as defined in Annex V of the WFD. One of these biological elements is ‘macrophytes and phytobenthos’.

The WFD also requires that all MS participate in an intercalibration exercise in order to ensure that high and good ecological status concepts are consistent across the EU. In order to achieve this, each MS is required to establish Ecological Quality Ratios (EQRs – observed status / expected status) for the boundaries between high (H) and good (G) status and for the boundary between good (G) and moderate (M) status, which are consistent with the WFD normative definitions of those class boundaries given in Annex V of the WFD. All 27 MS of the EU are involved in this process, along with Norway, who has joined the process on a voluntary basis. Expert groups have been established for lakes, rivers and coastal/transitional waters, subdivided into 14 Geographical Intercalibration Groups (GIGs -groups of MS that share the same water body types in different sub-regions or ecoregions). The outcome of the process will be published by the European Commission (van der Bund *et al.*, 2008).

This paper describes an intercalibration exercise performed on river phytobenthos responses along a nutrient enrichment / organic pollution gradient in the Central / Baltic GIG (CB GIG), the largest of the five GIGs established for the intercalibration exercise, covering most of central Europe, from the Atlantic coast to the Baltic states (other GIGs encompass northern parts of Britain and Ireland plus Scandinavia, Alpine regions, the Mediterranean and Eastern Continental areas – see ECOSTAT, 2004). Twelve MS belonging to CB GIG are taking part in the phytobenthos IC exercise: Austria (AT), Belgium (BE), Estonia (EE), France (FR), Germany (DE), Ireland (IE), Luxembourg (LU), Netherlands (NL), Poland (PL), northwest Spain (ES), Sweden (SE) and the United Kingdom (UK). Czech Republic, Denmark, Italy, Latvia and Lithuania are also part of CB GIG but have not been involved in this exercise. The two administrative regions of Belgium that joined the intercalibration, Flanders (BE-F) and Wallonia (BE-W) have different methods for assessing ecological status and are treated separately here.

Annex V of the WFD identifies four characteristics of ‘macrophytes and phytobenthos’ (taxonomic composition, abundance, likelihood of undesirable disturbances and presence of bacterial tufts) that need to be considered when setting status class boundaries. Most MS in CB GIG have chosen to develop separate assessment methods for macrophytes and phytobenthos and, in addition, to use diatoms as proxies for phytobenthos, focussing on taxonomic composition. A variety of approaches have been adopted (Table 1), most of which are based on existing weighted-average metrics although a few MS have developed new methods for the WFD based on the relative abundance of those species which are characteristic of reference sites compared with that of species which are associated with impacted conditions.

The metrics used by MS convert the response to a pressure gradient into a continuous variable which then has to be converted into an EQR, computed from Observed (O) and Expected (E) status values. The WFD defines high, good and moderate status in terms of their deviation from the biota expected at the reference state ("no or only very minor alteration to the water body resulting from human activities") and, therefore, a national method, if it is to be compliant with the WFD, has to be able to express each status class in terms of change from the reference state. The intercalibration was performed at the scale of individual diatom samples; in practice several MS will base final classifications of water bodies on multiple samples from one or several sites and, in some cases, on non-diatom algae too.

Methods

Sample collection and analysis

Samples were collected and analysed following standard methods (CEN, 2003; 2004; Kelly et al., 1998). At least 300 specimens were identified using standard Floras (primarily Krammer and Lange-Bertalot, 1986, 1997, 2000, 2004) and counted. The resulting lists of taxa plus relative abundances were then used to compute the national metric and the intercalibration metrics.

Test datasets

Data used in this exercise came from two sources (depending upon the MS): either from existing monitoring networks or from research projects associated with method development for the WFD. All data were stored in a central relational database, managed by Scottish Environment Protection Agency (UK). The database comprises three main components: raw diatom data, supporting chemical data and sample information. There were, however, difficulties in obtaining comparable environmental data due to differences in both determinands (e.g. total phosphorus versus soluble reactive phosphorus) and sampling strategies (e.g. spot measurements versus annual means) which, in turn, limited the number of comparisons that could be made. A summary of the number of sites available in each quality class (including reference sites) from each MS is presented in Table 2. The EU working group on ecological status ('ECOSTAT') defined a number of river types within each GIG based on the area, altitude and predominant geology of the catchment (ECOSTAT, 2004). Preliminary studies showed, however, that this "intercalibration typology" was not helpful for separating diatom assemblages and this typology was not used for subsequent analyses (an additional benefit of pooling these types was that analyses were based on larger datasets).

Member States were also asked to screen all their candidate reference site data according to agreed catchment land use and chemical reference thresholds (Table 3: WFD CIS Guidance Document No. 10, 2003) and to identify these samples in their national datasets (Table 2). Member States were also asked to indicate if they used more stringent criteria (or different but equivalent ones).

Development of the Intercalibration Common Metric

In order to compare status class boundaries developed in each MS, national metrics first had to be converted to a common scale. The mechanism for doing this was to develop an ‘intercalibration common metric’ (ICM: see Buffagni *et al*., 2005) – a metric with a statistically-significant relationship with all of the national metrics so that EQR values computed using national metrics can be quoted on a common scale.

Initial comparisons were made between national metrics and four of the most widely-used metrics: the Indice de Polluosensibilité (IPS: Coste, in CEMAGREF, 1982); Trophienindex (TI: Rott et al., 1999) and Saprobienindex (SI: Rott *et al*., 1997) and a revised form of the Trophic Diatom Index (TDI: Kelly *et al*., 2008). In most cases there were high correlations with national metrics; however, two types of response were observed along the pressure gradient, with two metrics (TI, TDI) being particularly responsive at low levels of nutrient / organic pressure (moderate to high EQRs) and the other two (IPS, SI) being more responsive at higher pressure levels (low to moderate EQRs). Rather than use any metric in isolation, a simple multimetric, composed of two of the candidate ICMs was tested. The TI was chosen over the TDI as the ‘sensitive’ metric as this had a slightly better performance when compared to ambient nutrient concentrations, whilst the IPS was chosen over the SI as the complementary metric as this metric was already widely used as a national metric within the GIG. The metrics were converted to EQRs as follows:

IPS: this metric measures ‘general water quality’, and is used widely to integrate effects across the entire water quality gradient (Hering et al., 2006), with low values corresponding to high pressure levels and, therefore, low EQRs. Therefore:

EQR_IPS = Observed (O)/Expected (E),

where: Expected = median IPS value of reference sites for a national dataset. Different reference values for each national type could be used, if appropriate, and the two MS without reference sites used expert judgement to select reference sites from neighbouring countries (the latter were not included in the calculation of a mean reference value based on all MS data).

TI: this is a trophic index, with low values corresponding to low nutrient concentrations (= high ecological quality and needs to be inverted so that high values represent high EQR values, therefore,

EQR_TI = (4-O)/(4-E)

(4 is the maximum possible value of the TI). Expected values were calculated as for EQR_IPS.

Two options for combining the metrics were considered (Table 5): where metrics indicating the same stressor are combined in a multimetric index, then the average of these metrics is the most appropriate value to use (based on the assumption that it shows the stronger relationship across the entire gradient). However, if the metrics indicate different stressors, then the minimum value of the two metrics would be appropriate. The response of the TI and IPS to a nutrient / organic gradient is assumed to be a composite of a number of ecological and physiological processes, with interspecific competition for inorganic nutrients prevailing at low pressure levels (high

EQRs) whilst factors such as tolerance to ammonia toxicity, capacity for heterotrophic growth and survival in environments with low oxygen concentration and redox potential prevailing at high levels of nutrient / organic stress (low EQRs). If this is the case, then the nutrient / organic gradient could be viewed as a combination of stressors, and the minimum of EQR_TI and EQR_IPS might be an appropriate measure.

The performance of the ICM was evaluated using linear regression models. The objective is to predict values of the ICM from values of each national metric. Four properties were used to evaluate the relationship:

- A visual examination of scatterplots to check for a linear response between the ICM and national metrics;
- The root mean square error (RMSE • residual standard error: a measure of prediction error - Wallach & Goffinet, 1989);
- The coefficient of determination (r^2); and,
- The closeness of the slope of a Model I regression of the ICM against the national metric to 1 (to maximise sensitivity of predictions across the entire EQR scale).

The coefficient of determination (r^2) measures association between two variables and gives little indication of the predictive power of that relationship. It is also dependent, to some extent, on the length of the gradient over which the coefficient is applied. RMSE, on the other hand, gives a better indication of the predictive power of the relationship, regardless of gradient length. Using both, along with visual examination and slope, provides a robust basis for evaluating relationships between national metrics and the ICMs. Overall, RMSE was lower using ICM (mean) though ICM (min) gave slopes closer to unity and higher r^2. However, examination of scatterplots showed fewer obvious deviations from linearity using ICM (mean) and this was selected as the most appropriate tool for this exercise.

Conversion of national metrics to the ICM

Values of the national metric representing the High / Good and Good / Moderate ecological status boundaries to corresponding values of the ICM were computed using a conventional (Model 1) linear regression equation. For each MS, EQR values from the national assessment method were plotted against the corresponding EQRs from the ICM and the regression equation and associated statistics were calculated. Conspicuous outliers were removed prior to calculation of the regression equation. Fig. 1 shows an example regression between the EQR values of the national metric and the ICM for one MS. For MS with different boundary values for separate national river types, a single relationship was computed for each national dataset and this relationship was used to convert boundary values for each national type to the ICM. Some MS had national types each with a different reference value. In these cases, EQR values were calculated for each type separately and then all data were pooled before the regression was calculated.

The EE national dataset had a curvilinear response to the ICM. A second order polynomial equation was fitted to this dataset:

$$\text{ICM} = a + b_1(\text{national metric as EQR}) + b_2(\text{national metric as EQR})^2$$

The UK national dataset also had a curvilinear response to the ICM but this was less pronounced and a linear equation was computed for the portion that included high, good and moderate status. ICM values for the H/G and G/M boundaries are presented as the predicted value ± the confidence limits of the regression line.

Comparison of national boundaries

In order to test whether status class boundaries were consistent across the CB GIG, it was necessary to define the range of variation between national methods that was considered to be 'acceptable'. This is, obviously, a subjective process and the approach adopted here was a convention used for intercalibration exercises for other biological elements. The acceptable range of boundary values was considered to be the median boundary value ± 0.05 EQR units for all MS that fulfilled an agreed list of criteria. If status class boundaries are approximately equally spread along the EQR gradient then 0.05 EQR units represents approximately 25% of the distance between two adjacent boundaries. The criteria used to select those MS that were used to calculate the acceptable band were:

- The assessment system and boundary values had been approved by the competent authorities within the MS;
- The national dataset contained at least six reference samples (representing at least four sites) screened according to ECOSTAT and CB GIG guidelines;

- There was a statistically-significant linear relationship with the ICM (see above).

These criteria excluded PL (whose national assessment had not been formally adopted at the time of the exercise), BE-F and NL (both of whom did not have any reference sites), BE-W (who use a predicted reference value for their national EQR values), EE and UK (whose national metrics had a curvilinear relationship with the ICM), and IE (whose data gave a low slope when the ICM was plotted against the national metric).

In order to meet the requirements of the intercalibration exercise, MS with boundary values falling below the acceptable range of boundary values defined by the outcome of the exercise must harmonise their national boundary values such that equivalent values of the ICM fall within the acceptable range. MS with boundary values occurring above the acceptable band were not required to adjust their boundary values as this would force MS to impose less stringent boundaries and would be contrary to the spirit of the WFD.

Statistical methods

Statistical analyses were performed using the R software package (R Development Core Team, 2005). The ordination technique Detrended Correspondence Analysis (DCA: Hill, 1979) was run using the Vegan package within R (Oksanen et al., 2007) after nomenclatural differences within the national datasets had been resolved.

Results

Comparison of reference conditions

Reference data were analysed in two ways: first, the IPS, SI, TI and TDI were calculated for all samples, after which the biological data for all reference sites were submitted to DCA.

Fig. 2 a-d shows the variation in values of the four metrics between MS, ignoring both intercalibration and national typologies. DE, ES, PL and SE tended to have lower values for SI, TI and TDI and higher values for IPS than other MS, whilst BE-W, EE and LU tended to have higher values for SI, TI and TDI and lower values for IPS. Other MS were neither consistently high nor consistently low.

The mean value of the TI was 2.1 (Fig. 2c), which means that variation in reference samples alone extends across about 50 per cent of the entire metric scale. The TI was designed to be particularly sensitive to inorganic nutrients, and the mean value of the IPS, a metric which operates across a longer nutrient/organic gradient was 17.4, although LU had a mean value of 15.8 and one LU reference sample had an IPS value of 11.9.

The ordination of reference samples showed no clear separation of MS based on community composition alone; however, there were significant correlations between the first axis of the DCA and the four metrics. The strongest relationship was with the TDI (r = -0.757; Fig. 3) suggesting that there is still a significant response to nutrients even after the reference screening process, at least in some of the MS. It was not possible to ascertain whether this was due to deficiencies in the reference screening process or to variation in background nutrient concentrations.

Comparison of boundaries

Fig. 4 shows the national boundary values expressed as ICM for the MS participating in the exercise for high/good and good/moderate ecological status respectively, along with the 95% confidence intervals for the predictions. The band of acceptable values for the high/good boundary has been superimposed on these charts. Seven MS fall within the acceptable band for H/G and six for G/M. A few other MS are marginally above or below one or both boundaries (i.e. the upper or lower 95th confidence limit overlaps with the acceptable band) while in each case four MS fall outside the acceptable bands for the H/G and G/M boundary. Table 5 shows a detailed breakdown of results, taking national typologies into account.

The order of national boundaries (using the mean boundary value for MS with >1 national river type) for the high / good boundary was as follows:

BE-W > ES > BE-F > **DE > AT > IE > EE > SE > UK > PL** > FR > LU > NL

(MS in bold fall within the acceptable band).

For the good / moderate boundary, the order was:

IE > BE-F > BE-W > EE > ES > **UK > DE > AT > FR > LU > SE** > PL > NL

The open circles in Fig. 4 indicate the adjusted high/good and good/moderate boundary values that were submitted by Be-W, NL and SE following the outcome of the intercalibration exercise.

Discussion

The WFD represents both a significant step forward in the way in which surface water bodies in Europe are managed but, at the same time, it sets a number of challenges for legislators. The requirement for intercalibration of methods between MS is one of the significant challenges, particularly as many national methods had evolved independently over a number of years to fulfil particular monitoring needs within national programmes.

Methods

In contrast to some other biological elements such as benthic invertebrates (Buffagni et al., 2005), sampling and analysis methods for diatoms are relatively consistent across Europe (Kelly et al., 1998), allowing European standards (CEN, 2003, 2004) to be developed. All those participating in the intercalibration exercise had methods which conformed to these which, in effect, removed one potentially important source of variability from the exercise. Most national metrics were based on the weighted average equation of Zelinka and Marvan (1961) which meant that they were based on similar calculations and all assumed that the primary pressure responsible for deviations from reference conditions was due to nutrients and/or organic pollution and are not sensitive to other pressures (e.g. acidification, toxic pollutants). There were also some fundamental differences in philosophy with regard to the construction of national methods – between, for example, the IPS and TI/TDI – that complicated the intercalibration process. Three MS (BE-F, ES, NL) developed new metrics which were, in many ways, closer to the spirit of the WFD in that they measured deviation from the reference state in terms of species composition without presuming the cause of this deviation; however, these were tailored quite specifically to conditions in localised regions of Europe. The ICM provides a common benchmark that can be applied across Europe albeit with less sensitivity to local conditions than many of the national metrics. Overall, the legal requirement to intercalibrate probably contributed to a conservative approach to method development as radical approaches to ecological status assessment are, by their nature, more difficult to compare with other methods.

The WFD refers to ‘macrophytes and phytobenthos’ rather than to diatoms alone. Most MS chose to develop separate methods for macrophytes and phytobenthos, and a separate intercalibration exercise is being performed on macrophytes (see Birk et al., 2006). Most MS also assumed that ‘diatoms’ were proxies for ‘phytobenthos’. The exceptions were DE and AT, both of which also have separate methods for non-diatom phytobenthos (Schaumburg et al., 2003; Rott et al., 1997, 1999) and UK and IE who tested the assumption that diatoms were valid as proxies (Kelly, 2006; Kelly et al., 2008). None measured phytobenthos abundance: although this is included in the definition of ecological status in the WFD, there is little evidence that such data provides information about nutrient dynamics over and above that provided by taxonomic composition (Biggs & Close, 1989; Islam et al., 2007).

Comparison of reference conditions

EQRs embody the core concept of the WFD: the comparison of the observed state of the biota at a point in space and time with that expected in the absence of anthropogenic disturbances. This means that the intercalibration exercise is simultaneously evaluating two important concepts: the sensitivity of the national metric to the pressure gradient and the value of the denominator. The WFD implies that it should be possible to derive the biological reference conditions for a water body from the hydromorphological and physico-chemical attributes that defines its type. Establishing that reference sites are truly free from anthropogenic influences was a time-consuming task for all MS, and was critically dependent upon the availability of suitable data. The criteria by which reference sites were selected varied between MS (see Table 4). The consensus within CB GIG was that land use criteria provided the most sensitive means of selecting reference sites although such data were not available for all MS and there was always a measure of expert judgement in the process. There was not, however, always a correlation between the rigour of the screening process and the values of metrics at reference sites (Fig. 2 a-d) and there was also still a nutrient-related gradient within the reference samples. It is not clear from this exercise whether these differences are due to underlying differences in the unimpacted state between MS or whether they reflect failures to screen data adequately.

Limitations of the intercalibration exercise

The WFD requires MS to establish biologically-relevant ‘typologies’ for water bodies, each of which should have a unique reference value. In theory, this allows MS to take account of hydrological and geological factors which affect the fauna and flora when setting the reference values. A simple Europe-wide typology should in principle, improve the precision of the intercalibration exercise - allowing comparisons between ecologically-similar sites within the GIG; however, the ‘intercalibration typology’ (ECOSTAT, 2004) proved unhelpful because it was unable to differentiate between the diatom assemblage found at 'reference conditions'. Abandoning the intercalibration typology had the added benefit of allowing those MS with fairly small datasets to pool all their samples into a single large dataset rather than split these between spurious 'types', with the concomitant loss of precision in the exercise.

A further limitation is that the exercise implicitly considered only the response of national methods to a general gradient of eutrophication and organic pollution, from which sites suffering from pressures such as acidification and toxic pollution had been removed.

Practical consequences

The results of the first round of Intercalibration, completed in 2007, will be published in the European Commission's technical report (van de Bund et al., 2008). MS will be legally bound by the boundary values published in this report and will use these boundaries to inform the first river basin management plans to be published in 2009. The river basin management plans will outline the programmes of measures required for water bodies that fail to meet good ecological status to ensure that good status is achieved by 2015.

MS that submit intercalibration data or finalise their national boundary values subsequent to the publication of the results of the first intercalibration exercise must ensure that their national boundary values (expressed as ICM) fall within the acceptable band for high/good and good/moderate status boundaries.

Conclusions

The intercalibration exercise presented a number of challenges for those taking part. It also proved to be a valuable forum for exchanging views and ideas on how the WFD should be interpreted and allowed MS to reconsider provisional classification boundaries based on expert judgement against reference data from the GIG region (Denys 2006). Although it forced a re-evaluation of national approaches in some cases, the final results were surprisingly consistent, with most MS falling in or very close to the 'acceptable band' for high/good and good/moderate status boundaries. The upper and lower limits of the 'acceptable bands' now become the *de facto* limits for all MS within CB GIG and those MS who fell below are required by the European Commission to adjust their national boundaries or present a reasoned justification for retaining their present limits. It also sets limits for those MS that fall within the geographical limits of CB GIG but which had not established national methods by the time that the exercise was conducted.

A criticism of exercises such as this is that they encourage a conservative approach to the development and application of methods. MS with established and extensive monitoring programmes contribute the bulk of data, centring the acceptable boundary ranges around those of their national classification. The WFD is, in many ways, an ambitious and visionary piece of legislation (Moss, 2007); however, the need to 'intercalibrate' is one factor that discourages radically new approaches to environmental monitoring as methods need to produce data that can be compared with that produced by other MS. Furthermore, the 'acceptable band' represents no more than a consensus of national viewpoints and whilst the exercise was able to ensure that all were broadly compatible with the normative definitions, and to set minimum requirements that national methods should meet, the outcome is perhaps best described as ‘acceptable practice’ rather than ‘best practice’. However, compared with the state of algal-based monitoring of European rivers just one or two decades ago (Whitton et al., 1991; Whitton & Rott, 1996), the results of this exercise must be seen as a major step in the right direction.

Conclusions

1. Intercalibration of diatom-based monitoring techniques used by EU members to assess ecological status was facilitated by the adoption of similar sampling and analytical methods and metrics; it is, however, complicated by the adoption of different conceptual frameworks.

2. Although reference sites in all participating countries were subject to the same minimum screening criteria, a significant nutrient gradient remained in the dataset. Part of this may be due to differences in background nutrient concentrations in different stream types but the study was unable to separate out different stream types in a meaningful manner.

3. Use of a common intercalibration metric enabled all high/good and good/moderate ecological status boundaries to be converted to a common scale and compared. As a result of this, countries with particularly high or low boundary values were able to adjust to a value that was in line with those adopted by other countries participating in the exercise.

Acknowledgements

All participants in this exercise acknowledge support from national governments and regulatory agencies. Thanks to Sebastian Birk (University of Essen), John Murray-Bligh (Environment Agency, UK) and Marian Yallop (University of Bristol, UK) for comments on a draft of this manuscript. Views and opinions expressed in this paper do not necessarily represent the final policy positions of any of the governments involved nor of the European Commission.

References

AFNOR, 2000. Détermination de l'Indice Biologique Diatomées (IBD). Norme NF T, 90-354.

Bernhardt, E.S., Likens, G.E., 2004. Controls on periphyton biomass in heterotrophic streams. Freshwater Biology 49, 14-27.

Biggs, B.J.F., 1996. Patterns in Benthic Algae of Streams in Stevenson, R.J., Bothwell, M.L., Lowe, R.L. (eds): Algal Ecology: Freshwater Benthic Ecosystems, Academic Press, San Diego.

Biggs, B.J.F., Close, M.E., 1989. Periphyton biomass dynamics in gravel bed rivers – the relative effects of flows and nutrients. Freshwater Biology 22, 209-231.

Birk, S., Korte, T., Hering, D., 2006. Intercalibration of assessment methods for macrophytes in lowland streams: direct comparison and analysis of common metrics. Hydrobiologia 566, 417-430.

Buffagni, A., Erba, S., Birk, S., Cazzola, M., Feld, C., Ofenböck, T., Murray-Bligh, J., Furse, M. T., Clarke, R., Hering, D., Soszka, H., van de Bund, W., 2005. Towards European inter-calibration for the Water Framework Directive: procedures and examples for different river types from the E.C. Project STAR. 468pp. Istituto di ricerca sulle acque, Rome.

CEN, 2003. Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers. EN 13946: 2003. Comité European de Normalisation, Geneva.

CEN, 2004. Water quality - Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. EN 14407:2004. Comité European de Normalisation, Geneva.

Coste, M., 1982. Etude des méthodes biologiques d'appréciation quantitative de la qualité des eaux. CEMAGREF rapport Q.E. Lyon A.F. Bassin Rhône-Méditérannée-Corse, 218pp.

Coste, M., Ayphassorho, H., 1991. Étude de la qualité des eaux du Bassin Artois-Picardie à l'aide des communautés de diatomées benthiques (application des indices diatomiques). Rapport CEMAGREF. Bordeaux – Agence de l'Eau Artois-Picardie, Douai.

Denys, L., 2006. Validation and revision of the reference concept for river phytobenthos in Belgium – Flanders proposed for the European Water Framework Directive, based on diatom assemblages from reference sites in the Central-Baltic GIG region. Advies Instituut voor Natuur- en Bosonderzoek INBO.A.2006.163, Brussel, 8pp.

ECOSTAT [Working Group 2.A Ecological Status], 2004. Overview of common Intercalibration types. Final Version 5.1, 38pp.

European Union, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 20000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L327: 1-73.

Hill, M.O., 1979. DECORANA - A FORTRAN program for Detrended Correspondence Analysis and Reciprocal Averaging. Ecology and Systematics, Cornell University, Ithaca, New York.

Hendrickx, A., Denys, L., 2005. Toepassing van verschillende biologische beoordelingssystemen op Vlaamse potentële interkalibratielocaties overeenkomstig de Europese Kaderrichtlijn Water : partim 'fytobenthos'. Rapport van het Instituut voor Natuurbehoud, IN.R.2005.06. Instituut voor Natuurbehoud: Brussel : Belgium, 107 pp.

Hering, D., Johnson, R.K., Kramm, S., Schmutz, S., Szoszkiewicz, K., Verdonschot, P., 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshwater Biology 51: 1757-1785.

Islam, A. Sharifi, M., Kelly, M.G., 2007. Evaluation of the Trophic Diatom Index for assessing water quality in the River Gharasou, western Iran. Hydrobiologia 589: 165-173.

Johnson, R. K., Goedkoop, W., Willén, E., Larsson, D., 2003. Typanpassning av referenssjöar och vattendrag: Kritisk granskning av biologiska kvalitetsfaktorer med bedömningsgrunder. Report 2003:17. Department of Environmental Assessment, Swedish University of Agricultural Sciences, Uppsala: 47pp.

Kelly, M.G., 2006. A Comparison of diatoms with other phytobenthos as indicators of ecological status in streams in northern England, 139-151. In: Proceedings of the 18th International Diatom Symposium (edited by A. Witkowski). Biopress, Bristol.

Kelly, M.G., Whitton, B.A., 1995. The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. Journal of Applied Phycology 7, 433-444.

Kelly, M.G., Juggins, S., Guthrie, R., Pritchard, S., Jamieson, J., Rippey, B. Hirst, H., Yallop, M., 2008. Assessment of ecological status in U.K. rivers using diatoms. Freshwater Biology 53, 403-422.

Kelly, M.G., King, L., Jones, R.I., Barker, P.A., Jamieson, B.J., 2008. Validation of diatoms as proxies for phytobenthos when assessing ecological status in lakes. Hydrobiologia (in press).

Krammer K. & Lange-Bertalot H., 1986. Die Süsswasserflora von Mitteleuropa 2: Bacillariophyceae. 1 Teil: Naviculaceae. 876 pp. Stuttgart: Gustav Fischer-Verlag.

Krammer K. & Lange-Bertalot H., 1997. Die Süßwasserflora von Mitteleuropa,II:2. Bacillariophyceae. Teil 2: Bacillariaceae, Epithemiaceae, Surirellaceae. 2te Auflage, mit einem neuen Anhang. 594 pp. Gustav Fischer Verlag, Stuttgart.

Krammer K. & Lange-Bertalot H., 2000. Die Süsswasserflora von Mitteleuropa 2: Bacillariophyceae. 3 Teil: Centrales, Fragilariaceae, Eunotiaceae. 2nd edition. Gustav Fischer Verlag, Stuttgart.

Krammer K. & Lange-Bertalot H., 2004. Süsswasserflora von Mitteleuropa 2, Bacillariophyceae. Teil 4 : Achnanthaceae. Kritische Ergänzungen zu Achnanthes s.l., Navicula s. str., Gomphonema. 468 pp. Spektrum Akademischer Verlag/Gustav Fischer, Heidelberg.

Leclercq, L., Maquet, B., 1987. Deux nouveaux indices chimique et diatomique de qualité d'eau courante. Application au Samson et à ses affluents (bassin de la Meuse belge). Comparaison avec d'autres indices chimiques, biocénotiques et diatomiques. Institut Royal des Sciences Naturelles de Belgique, document de travail 28.

Lenoir, A., Coste, M., 1996. Development of a practical diatom index of overall water quality applicable to the French National Water Board network. In: WHITTON BA and ROTT E (eds.) Use of Algae for Monitoring Rivers II. Institut für Botanik. Universität Innsbruck, 29-43.

Moss, B., 2007. Shallow lakes, the water framework directive and life: what should it all be about? Hydrobiologia 584, 381-394.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, R.B., 2007. Vegan: Community Ecology Package version 1.8-6. http://cran.r-project.org/

Pan, Y., Stevenson, R.J., Hill, B.H., Kaufmann, P.R., Herlihy, A.T., 1999. Spatial patterns and ecolocical determinants of benthic algal assemblages in mid-atlantic streams, USA. Journal of Phycology 35, 460-468.

Pardo, I. A. M., Olsen, C., Delgado, L., García, A., Nebra, Domínguez, M., 2005. Implantación da Directiva Marco da Auga 2000/60/CE no Ámbito territorial Galicia-Costa. Technical report.

R Project Core Development Team, 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rott, E.; Hofmann, G.; Pall, K.; Pfister, P., Pipp, E., 1997. Indikationslisten für Aufwuchsalgen. Teil 1: Saprobielle Indikation. Publ. Wasserwirtschaftskataster, BMfLF, 1-73.

Rott, E.; Van Dam, H.; Pfister, P.; Pipp, E.; Pall, K.; Binder, N., Ortler, K., 1999. Indikationslisten für Aufwuchsalgen. Teil 2: Trophieindikation, geochemische Reaktion, toxikologische und taxonomische Anmerkungen. Publ. Wasserwirtschaftskataster, BMfLF, 1-248.

Schaumburg, J., Schmedtje, U. Köpf, B. Schranz, C. Schneider, S. Meilinger, P. Stelzer, D. Hofmann, G. Gutowski, A., Foerster, J., 2005. Makrophyten und Phytobenthos in Flüssen und Seen. Leitbildbezogenes Bewertungsverfahren zur Umsetzung der EG-Wasserrahmenrichtlinie. Informationsbericht Heft 1/05. Bayerisches Landesamt für Wasserwirtschaft, München.

Schiefele, S., Schreiner, C., 1991. The use of diatoms for monitoring nutrient enrichment, acidification and impact of salt in rivers in Germany and Austria. In: Whitton, B.A., Rott, E. and Friedrich, G. (eds.) Use of Algae for Monitoring Rivers. Institut fur Botanik, Univ. Innsbruck, 103-110.

Slàdeček V, 1986. Diatoms as indicators of organic pollution. Acta Hydrochim. Hydrobiologia 14(5), 555-566.

Van der Molen, D.T. (Ed.), 2004. Referenties en concept-maatlatten voor rivieren voor de Kaderrichtlijn Water. STOWA-rapport 2004/43. STOWA, Utrecht. 365pp.

Van de Bund et al (editors), 2008. *Water Framework Directive Intercalibration Technical Report*. Joint Research Centre, Ispra.

Wallach, D., Goffinet, B., 1989. Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecological Modelling 44, 299-306.

Warton, D., 2005. SMATR: (Standardised) Major Axis Estimation and Testing Routines. R package version 2.0, translated to R by John Ormerod.

WFD CIS Guidance Document No. 10, 2003. Rivers and Lakes – Typology, Reference Conditions and Classification Systems. Working Group 2.3 - REFCOND Guidance Document No 10. Common Implementation Strategy for the Water Framework Directive (2000/60/EC). ISBN: 92-894-5614-0, ISSN: 1725-1087.

WFD CIS Guidance Document No. 14, 2005. Guidance on the Intercalibration Process 2004-2006. Common Implementation Strategy (CIS) for the Water Framework Directive (2000/60/EC). ISBN: 92-894-9471-9, ISSN: 1725-1087.

Whitton, B.A., Rott, E., 1996. Use of Algae for Monitoring Rivers II. Institut fur Botanik, Univ. Innsbruck, 196pp.

Whitton, B.A, Rott, E., Friedrich, G., 1991. Use of Algae for Monitoring Rivers. 193pp. Institut fur Botanik, Univ. Innsbruck.

Zelinka, M., Marvan, P., 1961. Zur Präzisierung der biologischen Klassifikation des Reinheit fliessender Gewässer. Archiv für Hydrobiologie 57, 389-407.

Table 1: Member State (MS) national metric/assessment methods for phytobenthos intercalibration.

MS	National metric
AT	Multimetric method consisting of 3 modules/metrics: A) trophic status module (based on Trophienindex: Rott et al. 1999) B) saprobic status module (based on Saprobic Index: Rott et al. 1997) C) reference species module (portion of defined reference and bioregion-specific species in total abundance and species number) Ecological status is evaluated separately for each of the modules and overall phytobenthos classification is equivalent to the worst of the three results (worst-case-scenario).
BE-F	Proportions of impact-sensitive (‘positive’) and impact-associated (‘negative’)indicator taxa (Hendrickx & Denys, 2005)
BE-W	IPS (Coste, in CEMAGREF, 1982; Lenoir & Coste, 1996)
DE	Diatom Module: WFD Diatom Index = Average of the sum of abundances of type specific reference species (following Schaumburg et al. 2005) and Trophienindex (Rott et al., 1999) or (in one special case) Saprobic Index (Rott et al., 1997). Additional metrics are available for cases of acidification or salinisation. There are, in addition, modules for non diatom algae and Macrophytes (Schaumburg et al. 2005) Ecological status is calculated and classified from the average of the three module scores. If a module is absent, status class can be calculated with two modules or, exceptionally, with a single module. For this reason every module is classified separately and can be considered separately for intercalibration purposes. The national classification system needs all modules of the benthic flora occurring in a monitoring section of a water body.
EE	IPS (Lenoir & Coste, 1996); boundaries set following WFD CIS Guidance Document No. 10 (2003).
ES	MDIAT (Diatom multimetric). composed by averaging six indices calculated using OMNIDIA (SHE +SLAD+IDG+TDI+IPS+L&M – see below) along with two sensitive taxa metrics constructed with the reference diatom community of small and medium-sized rivers in Galicia (NW Spain) (FPSS+PABSS). Note: SLAD: Slàdecek (1986); SHE: Schiefele & Schreiner (1991); IDG: Coste & Ayphassorho (1991); IPS: Coste in CEMAGREF (1982); L&M: Leclercq and Maquet (1997); TDI: Kelly & Whitton (1995); FPSS: % richness of sensitive taxa (developed for Galicia,); PABSS: % abundant of sensitive taxa (developed for Galicia)

MS	National metric
FR	IBD (Lenoir & Coste, 1996, AFNOR NF T90-354, 2000)
IE	Revised form of Trophic Diatom Index (TDI) (Kelly et al., 2008)
LU	IPS (Coste, in CEMAGREF, 1982)
NL	EKR based on proportions of positive and negative indicator taxa (Van der Molen, 2004)
PL	Average of Trophic Index (Rott et al., 1999) and Saprobic Index (Rott et al., 1997)
SE	Swedish assessment methods, Swedish EPA regulations (NFS 2008:1) based on IPS (Coste, in CEMAGREF, 1982). See Johnson *et al.* (2003).
UK	Revised form of Trophic Diatom Index (TDI) (Kelly et al., 2008)

Table 2: Summary of the number of reference sites and the number of sites in each quality class from each Member State according to their national methods.

Member State	Reference	High	Good	Moderate	Poor	Bad	Total
AT	18	18	278	168	52	3	**519**
BE-F	0	0	15	22	29	14	**80**
BE-W	37	26	250	121	47	23	**467**
DE	8	8	11	22	11	1	**53**
EE	12	56	8	2	0	0	**66**
ES	18	40	57	41	6	0	**144**
FR	31	49	58	137	53	7	**304**
IE	12	14	16	16	4	1	**51**
LU	44	97	34	41	24	6	**202**
NL	0	26	57	32	18	20	**153**
PL	6	8	4	9	5	0	**26**
SE	16	16	10	15	4	1	**46**
UK	46	314	211	377	139	10	**1051**
Total	**248**	**672**	**1009**	**1003**	**392**	**86**	**3162**

Table 3: Criteria used by Member States to select reference sites. Key: 0: missing info; 1: not used; 2, measured; 3, estimated; 4, field inspection; 5, expert judgement.

	Landuse data (e.g. CORINE)	BOD_5	O_2	Phosphorus fractions	Nitrogen fractions	Comments
AT	1	2	1	2	1	
BE-F	2	2	2	2	2	
BE-W	1	2	2	2	2	
DE	3	2	2	2	2	Hydro morphological degradation, biological data, expert judgement
EE	0	1	1	2	2	
ES	2	2	2	2	2	REFCOND criteria used for invertebrate exercise
FR	2	2	2	2	2	
IE	3	2	2	2	2	
LU	3	2	2	2	2	A land use index was set from ministry of environment CORINE data
NL	5	5	5	5	5	
PL	3	2	1	2	0	
SE	2	1	1	2	1	Assessment of acidification
UK	1	3	3	2	2	

Table 4: Performance characteristics of linear regressions between national metrics and the minimum ('min') and mean ('mean') intercalibration metric (ICM) (based on EQR_TI and EQR_IPS). (* = non-linear response)

Member	RMSE		Slope		r^2	
State	min	mean	min	mean	min	mean
AT	0.072	0.056	0.901	0.654	0.616	0.506
BE-F	0.130	0.111	0.840	0.886	0.591	0.686
BE-W	0.065	0.083	0.640	0.645	0.792	0.705
DE	0.091	0.086	0.694	0.885	0.687	0.803
EE *	0.037	0.083	1.021	1.197	0.888	0.685
ES	0.0116	0.083	1.034	0.874	0.673	0.743
FR	0.105	0.122	0.668	0.826	0.621	0.653
IE	0.123	0.096	0.527	0.401	0.528	0.514
LU	0.110	0.079	0.622	0.719	0.540	0.752
NL	0.119	0.096	0.490	0.541	0.550	0.696
PL	0.037	0.062	1.067	1.030	0.983	0.951
SE	0.098	0.093	1.974	1.865	0.824	0.825
UK	0.095	0.061	0.379	0.233	0.349	0.323

Table 5. High/good and good/moderate boundaries expressed as ICM for all national datasets, sub-divided by national river type, where appropriate. The acceptable band for H/G is 0.839 – 0.939 and for G/M it is 0.654 – 0.754.

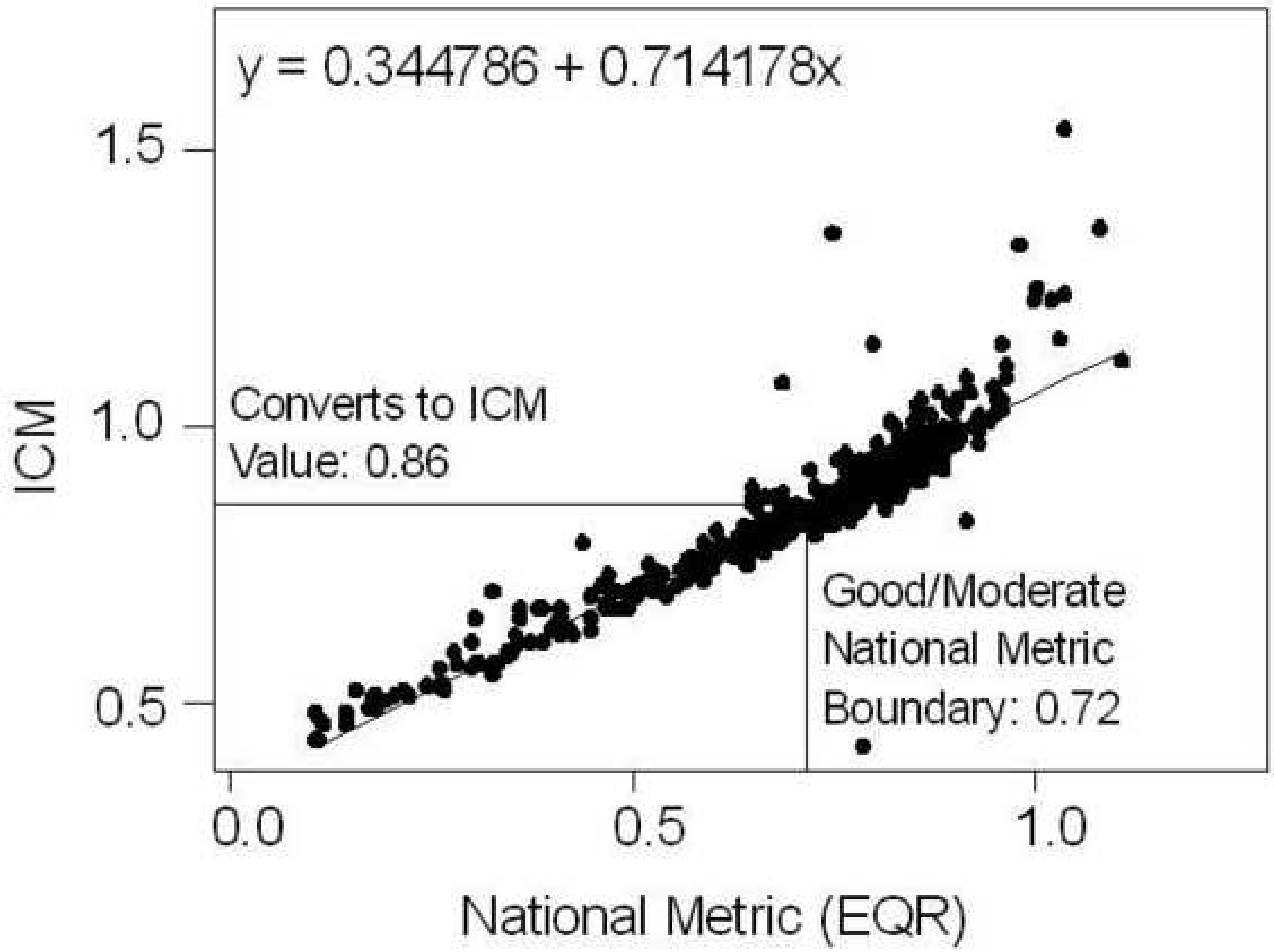
		Properties of regression			Boundaries	
	River Type	**R-squared**	**RMSE**	**slope**	**High/Good**	**Good / Moderate**
AT (Austria)		0.683	0.068	0.758		
	< 500 m				0.917	0.705
	> 500 m				0.917	0.705
	mean				**0.917**	**0.705**
BE-F (Belgium - Flanders)		0.686	0.112	0.886	0.997	0.82
BE-W (Belgium – Wallonia)		0.755	0.116	1.023	1.021	0.796
DE (Germany)		0.803	0.086	0.885		
	R-C1, R-C3				0.930	0.707
	R-C4				0.877	0.707
	R-C5				0.983	0.824
	mean				**0.930**	**0.746**
EE (Estonia)		0.828	0.062	*	0.862	0.779
ES (Spain)		0.718	0.109	1.054	1.001	0.759
FR (France)		0.653	0.128	0.826		
	Type 1				0.830	0.699
	Type 2				0.830	0.699
	Type 3				0.820	0.669
	Type 4				0.830	0.699
	mean				**0.828**	**0.703**
IE (Ireland)		0.566	0.092	0.446	0.911	0.844
LU (Luxembourg)		0.869	0.820	0.961	0.838	0.694

	Properties of regression			Boundaries	
River Type	**R-squared**	**RMSE**	**slope**	**High/Good**	**Good / Moderate**
NL (Netherlands)	0.696	0.096	0.541	0.685	0.577
PL (Poland)	0.951	0.060	1.030	0.849	0.633
SE (Sweden)	0.910	0.066	1.206	0.861	0.668
UK	0.563	0.121	*	0.858	0.752
Median Boundary				0.889	0.704

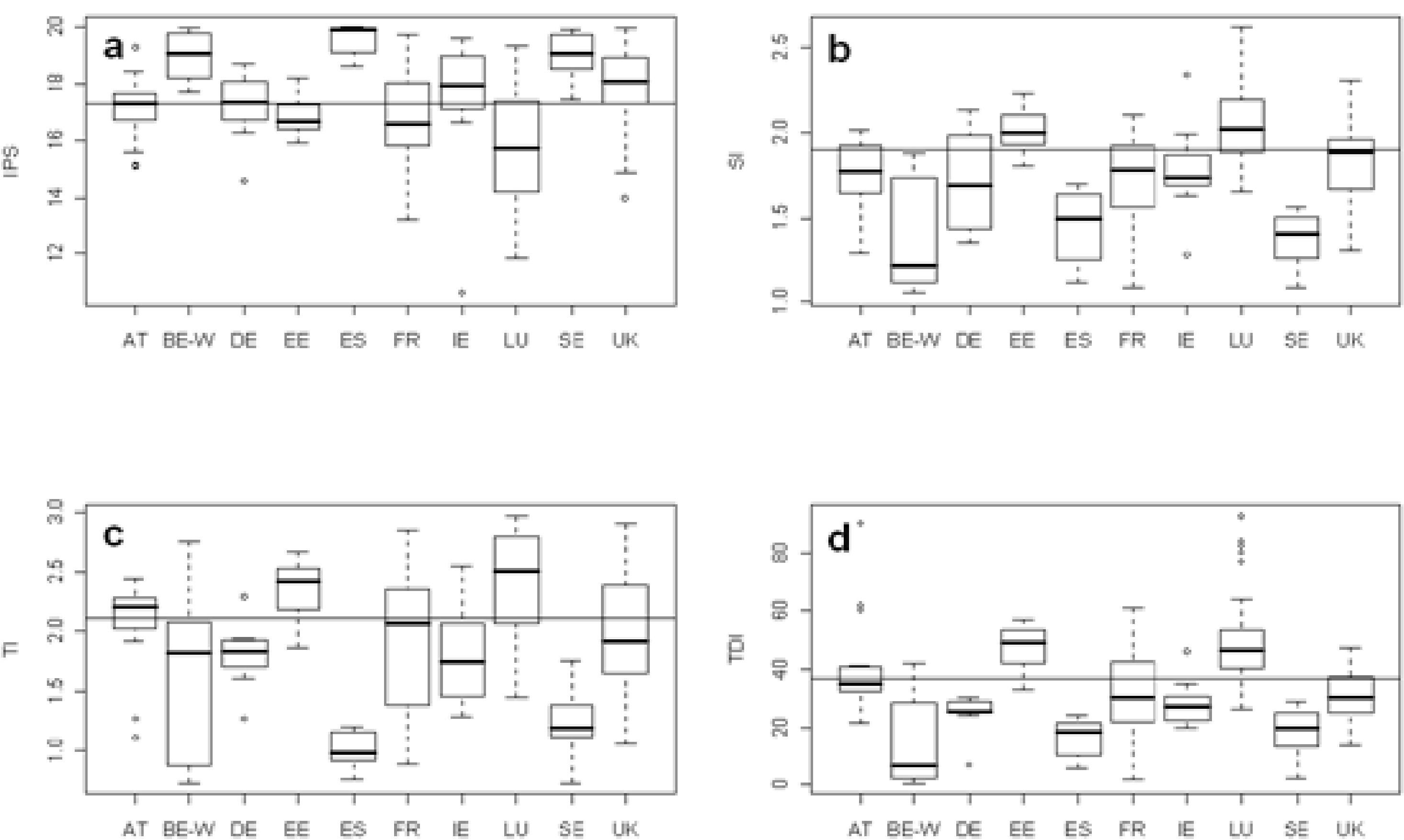
*Curvilinear relationship between National Metric and ICM

List of figures

Figure
Click here to download high resolution image

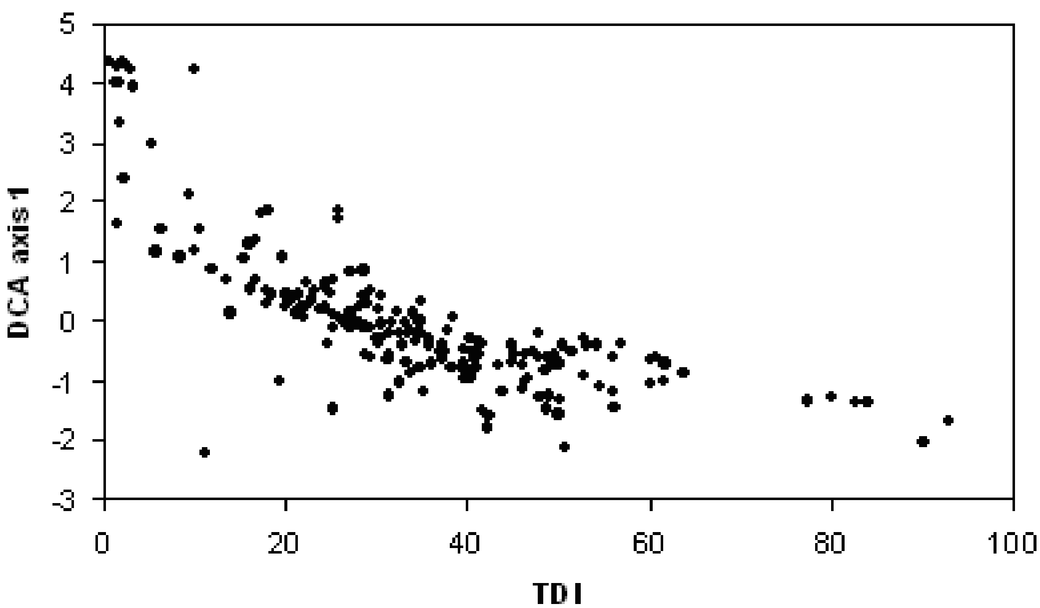


Figure



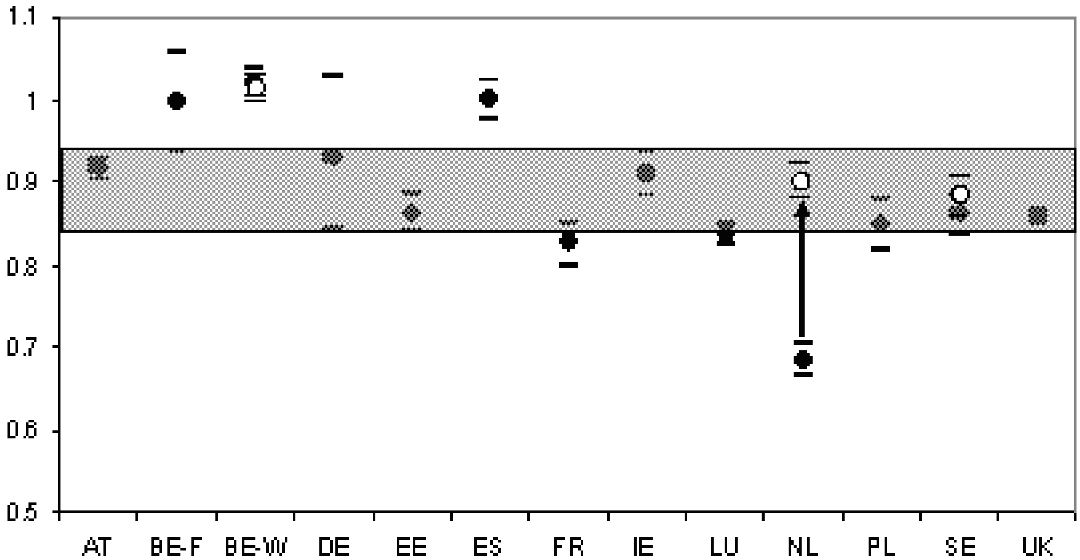
a
IPS
20
18
16
14
12
AT
BE-W
DE
EE
ES
FR
IE
LU
SE
UK
b
SI
2.5
2.0
1.5
1.0
c
TI
3.0
2.5
2.0
1.5
1.0
d
TDI
80
60
40
20
0

Figure



Figure

a) High/Good Boundary



b) Good/Moderate Boundary