OPINION

River macrophyte indices: not the Holy Grail!

BENOÎT O. L. DEMARS*, JACQUELINE M. POTTS†, MICHELÈ TRÉMOLIÈRES‡, GABRIELLE THIÉBAUT§, NATHALIE GOUVELIN*– AND VINCENT NORDMANN*–**

*The James Hutton Institute, Aberdeen, U.K.
†Biomathematics & Statistics Scotland, Aberdeen, U.K.
‡Université de Strasbourg, Strasbourg, France
§Université de Rennes, Rennes, France
–Ecolor, Fénétrange, France
**Ecole Nationale du Génie de l’Eau et de l’Environnement de Strasbourg, Strasbourg, France

SUMMARY

1. Recent studies have demonstrated that there is generally no unambiguous relationship between plant species composition and specific environmental conditions in rivers. Nevertheless, indices of environmental pressures based on macrophytes are flourishing, because of the requirements of the Water Framework Directive (WFD).

2. We first reviewed nine such indices against 13 criteria for bioindicators. Then, using data from France and England, we tested whether the IBMR (Macrophyte Biological Index for Rivers) and LEAFPACS (predictions and classification system for macrophytes) methods could reliably indicate nutrient and hydromorphological pressures. Finally, we used an improved bootstrapping method to estimate accuracy.

3. Currently, most indices lack ecological meaning for a variety of reasons, including partial sampling (backwaters are excluded); reliance on list of taxa (there are identification difficulties) rather than structure and functions; correlation rather than causation; application within a limited biogeographical area; reliance on ‘expert’ judgement; high precision but poor accuracy; poorly defined reference conditions; lack of independent tests; and an inability to discriminate reliably between the target pressures of interest from confounding background variables.

4. IBMR was a far better indicator of pH (or HCO$_3^-$-pCO$_2$) than it was of soluble reactive phosphorus, SRP (or SRP-NH$_4$). While there was a highly significant correlation between IBMR and SRP after removing the effect of pH, the relationship was weak ($r^2 = 0.08$, $n = 215$, $P < 0.001$).

5. LEAFPACS is a multi-metric method summing up five individual indices, each compliant with the WFD. Its individual metrics were not better correlated with nutrient and hydromorphological pressures (with $r^2 < 0.1$, $n = 62$, $P < 0.05$) than was the IBMR. The meaning of the overall metric is questionable.

6. There are problems in determining the precision of the indices, owing to uncertainties in recording, but they are less than the uncertainties in determining accuracy (because species optima and tolerances are sometimes poorly known).

7. Reliable information is needed to improve the state of our rivers. Macrophyte indices are able to detect statistically significant pressures from a large population of sites but cannot be applied at specific sites, as required by the WFD, owing to large uncertainties and low explanatory power. Typically, more than 90% of the variability in macrophyte indices is attributed to factors other than human pressure. The WFD would be better served by a simpler, holistic approach based on our current mechanistic understanding of river processes. These findings are likely to apply also to other taxonomic groups (macroinvertebrates, diatoms, etc.).
fish) used in the assessment of purported ecological quality and to palaeolimnological measures of reference status.

Keywords: aquatic plants, biomonitoring, multiple stressors, review, uncertainty

Introduction

Human water security and river biodiversity is globally threatened (Vörösmarty et al., 2010). Many nations have taken steps to protect their freshwater resources through legislation such as the Environmental Quality Standards for Surface Water in China (GB 3838-2002), the Clean Water Act in the United States of America or the European Water Framework Directive (WFD, 2000/60/EC).

The WFD has stimulated many scientific studies on biomonitoring, striving to evaluate the ecological conditions of water courses relative to (near) natural ‘reference’ conditions (i.e. with ‘no, or only very minor, anthropogenic alterations’, WFD, Annex V). Although the WFD defines ecological status as an expression of the quality of the structure and functioning of aquatic ecosystems associated with surface waters (Art. 2.21.), the current approach generally relies on linking environmental pressures to taxonomic-based indices. This is, perhaps, a reflection of Annex V defining ecological status as the basis of taxonomic composition and abundance of independent biological quality elements (supported by hydro-morphological and physicochemical elements) without referring to the interconnections among these biological elements.

This paradox has opened the door to the world of single and multi-metrics, with claims that macrophyte indices are able to detect single pressures (Dodkins, Rippey & Hale, 2005) that may not be detectable by other bioindicators (Johnson et al., 2006a,b). Such claims may be incompletely substantiated, however, because the authors (Dodkins et al., 2005; Johnson et al., 2006a,b) did not distinguish reliably between the effects of inorganic nutrients (N and P) and pH alkalinity, a problem that had been recognised a century ago (see West, 1910; Hutchinson, 1975, p. 391), but only recently investigated thoroughly in rivers (Demars & Thiébaut, 2008; Demars & Trémolières, 2009).

The number of macrophyte indices continues to grow relatively unchecked. On the basis of the previous studies on species composition (Demars & Harper, 1998, 2005; Demars & Edwards, 2007a, 2009; Demars & Thiébaut, 2008; Demars & Trémolières, 2009), however, we hypothesised that even the best of these indices would be unable to detect reliably specific stressors, such as N and P enrichment or physical impairments, within the main channel of rivers. In this Opinion piece, we first suggest a list of criteria against which macrophyte indices should be assessed. We then test two such indices, the IBMR (Haury et al., 2006) and LEAFPACS (Willby, Pitt & Phillips, 2009), with research data collected in France and England. We then present an improved method for calculating the ecological accuracy of a macrophyte index (often implicitly confused with the precision of an index). Finally, we present our perspective, from the viewpoint of a river ‘general practitioner’ on diagnosing symptoms of river health, drawing inspiration from the writing of Brian Moss (2001, 2002, 2007, 2008, 2010).

Aquatic macrophyte indices: a critical review

The focus of this article is on bioindicators of human pressures. Ecosystem structure and function will be discussed in our concluding reflections. Ideally, biomonitoring tools should: (i) allow the characterisation of the state of the river ecosystem through its biochemical, physiological or ecological characteristics; (ii) be applicable over as broad a biogeographical area as possible; (iii) indicate reliably human impact of a particular type and (iv) be derived objectively from a sound theoretical concept in ecology (modified from Blandin, 1986, p. 224; Doledec, Statzner & Bournard, 1999).

These criteria are the cornerstones against which the scientific credibility of macrophyte indices should be judged. In addition, there are other issues to consider such as (i) the difficulty of species identification; (ii) survey methods (snorkelling versus river bank assessment); (iii) survey area (inclusion of backwaters); (iv) gradients in environmental variables (strength and independence); (v) reference conditions in western Europe (criteria used); (vi) derivation of species scores; (vii) number of taxa involved; (viii) multiple pressures and (ix) conflicts of interest. Some of these criteria are common to all taxonomic indices and have been the subject of previous critical reviews (e.g. Moss, 2008).

An assessment of how key macrophyte indices match up to particular criteria is illuminating (Table 1). Only one assessment system is designed to determine the ecological
diagnosis of backwaters, all the others assess only the main stem. This is very surprising as backwaters are often taxonomically the most diverse (e.g. Bornette, Amoros & Lamouroux, 1998; Williams et al., 2003), yet the most altered (many were drained) and sensitive (especially to nutrient enrichment) areas of the river system. Moreover, even though NGOs and conservation agencies may include the floodplain in their definition of rivers, their actions (protection and restoration) may not guarantee the quality of ecosystem structure and function as requested by the WFD (Irvine, 2009).

Half of the indices consider only vascular plants, while the others include bryophytes and algae (Table 1). The species identification of vascular plants and bryophytes is often difficult because of the lack of fruiting or flowering specimens and hybridisation (e.g. Lansdown, 2007, 2009). The identification of macroalgae is generally restricted to genus despite the fact that ecology can vary substantially within a genus (e.g. Spirogyra; Simons & van Beem, 1990).

As an analogy, what would happen if we were to lump together all the species of Potamogeton or Ranunculus subgenus batrachium?

Plant traits are only considered by two indices, in contrast to continual developments in macroinvertebrate and fish studies (Dolédec & Statzner, 2010; Resh & Rosenberg, 2010; Statzner & Bécè, 2010). Plant traits have the advantage of providing an index more independent of biogeography (Ali, Murphy & Abernethy, 1999) or providing insights into underlying biological mechanisms determining species distribution and index scores (Demars & Trémolières, 2009).

Another issue is that many indices are based on species scores derived subjectively. How can an ‘expert’ provide species indicator values (optimum) for dissolved N and P?

Table 1 River macrophyte indices with the presence (●) absence (○) of selected criteria

<table>
<thead>
<tr>
<th>Indices*</th>
<th>Damage rating RTSI</th>
<th>Ecological diagnosis TIM IBMR CBAS GLM RMNI RMHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Survey in main channel and backwaters</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>2. Taxa: algae, bryophytes and vascular plants</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3. Species trait</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>4. Biological mechanisms</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>5. Independence of biogeography</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>6. Objective species or trait indicator value and tolerance</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>7. Site index score with ecological accuracy</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>8. Reference conditions based on typology or site specific conditions</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>9. Tested on independent data</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>10. Bioindication claim(s)</td>
<td>Gross pollution N, P</td>
<td>NH4, PO4 Scouring, Ground-water, Sedimentation</td>
</tr>
<tr>
<td>11. Independence of confounding factors</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>12. Peer reviewed</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>13. Used by public agencies or private consultancies</td>
<td>?</td>
<td>○</td>
</tr>
</tbody>
</table>

*Indices’ names and references: Damage rating (Haslam, 1982); River Trophic Status Indicator (RTSI; Ali et al., 1999); Ecological diagnosis (Amoros et al., 2000); Trophic Index of Macrophytes (TIM; Schneider & Melzer, 2003); Indice Biologique Macrophytique en Rivière (IBMR; AFNOR, 2003; Haury et al., 2006); CCA Based Assessment System (CBAS; Dodkins et al., 2005); General linear model (GLM, Demars & Trémolières, 2009); River Macrophyte Nutrient Index (RMNI, Willby et al., 2009); River Macrophyte Hydraulic Index (RMHI, Willby et al., 2009).
without measuring these nutrients? The species scores tend to reflect alkalinity, as acknowledged in the original studies (see Demars & Thiébaut, 2008). It is surprising that in 2011 we are still largely relying on expert judgement from the late 1970s and 1980s [e.g. the trophic rank scores of Newbold & Palmer (1979), later refined by Newbold & Holmes (1987) and Holmes et al. (1999), which are still at the heart of the French and British indices]. It is true that one report (Dawson et al., 1999) confirmed that the ‘trophic rank scores’ were more or less aligned with the water inorganic N and P, but warned that these scores may be confounded by other chemical, physical and spatial factors, as found later (Demars & Thiébaut, 2008; Demars & Edwards, 2009). Only three indices have derived objective scores based on statistical analyses linking species to environmental data, only one of which had measurements of sediment as well as water column nutrient concentrations (Table 1). There is no guarantee that these scores reflect causal relationships because they were derived from field data where environmental variables are rarely independent.

Nearly all systems of bioindication produce final scores, but only three indices had meaningful units: units of species turnover (equivalent to Whittaker β diversity) or the same unit as the environmental pressure (e.g. μg P L⁻¹). Most scores are to be taken as a ‘golden number’ (here we refer to the golden ratio, see e.g. Livio, 2002, too often over-interpreted) and have the effect of disconnecting conservation and regulatory organisations from nature. Generally, uncertainties are inadequately reported, if at all.

More than half of the indices do not conform to the WFD because they do not assess the ecological condition relative to (near) natural state. The deviation from reference conditions approach was devised much earlier than the WFD (e.g. Haslam, 1982) and well publicised (e.g. Haslam, 1987). It was not applied because of the difficulty in finding reference conditions. A combination of historical, palaeolimnological and modelling approaches may be necessary (see e.g. Brush & Hilgartner, 2000; Andersen, Conley & Hedal, 2004; Baattrup-Pedersen et al., 2008). Alternatively, the plant trait approach may allow us to look for reference conditions at the European scale. The old river typology, with the creation of too many detailed categories (see Demars & Edwards, 2009), has been recently superseded by site-specific reference conditions.

The definition of reference condition, the number of reference sites and their spatial distribution are crucial if one wants to understand what deviation from reference conditions means. So far, reference conditions represent a very limited number of sites (when indicated in publications) which are already somewhat degraded, and thus, there is a real danger that rivers will not be protected adequately (Moss, 2008). The risk of political compromise at the expense of environmental protection is ongoing. For instance, Pardo et al. (2012) recognised as potentially in reference condition a catchment with up to 50% intensive agriculture and with excessive nutrient concentration bearing little resemblance to a (near) natural system, independently of the geological context and catchment size (e.g. Meybeck, 1982; Perakis & Hedin, 2002; Demars & Edwards, 2007b).

Presently, only two indices have been tested quantitatively on independent data sets. All indices (including Dodkins et al., 2005; Willby et al., 2009) have failed so far to identify reliably specific environmental pressures (e.g. N, P enrichment) independent of natural variability (HCO₃, pCO₂, slope). The contrary claims by Dodkins et al. (2005) were based on a statistical misunderstanding (e.g. the claim of a multivariate species optimum is incorrect, because the species optima were calculated independently) and errors in statistical reporting (e.g. the selected predictive variables cannot explain more than 54.9% of the variability in species composition, but only 5.1% with the information provided in their Table 1 and text). Moreover, while significant statistical relationships have been established (with high probability), their explanatory power is generally weak (e.g. Dodkins et al., 2005 multivariate analyses could explain only around 2% of the variability in species composition with individual human pressures as predictors) or confounded with other factors.

Some indices are so complicated that it is difficult to see the ecological meaning of the final results (e.g. Ali et al., 1999; Dodkins et al., 2005; Willby et al., 2009) and see Hatton-Ellis (2008). Simpler systems come closer to what a ‘river doctor’ might be expected to do (e.g. Haslam, 1982; Amoros, Bornette & Henry, 2000). However, over simplistic indices, concentrating only on the pressure of interest, are likely to miss potential confounding effects or alternative causes (e.g. Schneider & Melzer, 2003). Diagnostic methods should strive to be based on mechanistic understanding (e.g. Demars & Trémolières, 2009), so that a programme of measures to restore impacted rivers could be put in place in the river basin management plans, as required by the WFD.

Despite all these gaps, nearly all the studies we considered here have been published in peer-reviewed scientific journals. Worryingly, some of these indices were used by regulatory agencies before they had been published in peer-reviewed scientific journals (e.g. IBMR, LEAFPACS).

© 2012 Blackwell Publishing Ltd, Freshwater Biology, 57, 1745–1759
Case studies

Macrophyte Biological Index for Rivers (IBMR)

The IBMR has been one of the most prominent indices used in the European STAR project (Furse et al., 2006). It was first published in 2003 by AFNOR (Association Francaise de Normalisation). While it resulted in part from previously published peer-reviewed papers (e.g., Haury et al., 1996), the method was published in a peer-reviewed journal three years after it was made available to regulatory agencies and citizens (Haury et al., 2006).

The IBMR index is intended to indicate concentrations of phosphate [soluble reactive phosphorus (SRP)] and ammonium (NH₄) and gross organic pollution (saprobity). However, this has not been appropriately tested. All that have been presented hitherto are changes in the IBMR score along rivers. Haury et al. (2006) focussed on small longitudinal changes, attributed to point source effluents, but did not explain the large differences in scores between rivers of different alkalinity. The only exception was represented by the small streams of the Northern Vosges and Alsace (north-eastern France), where the IBMR seemed to perform adequately (Haury et al., 2006). However, it now appears that pH explains more variability in IBMR than NH₄ and SRP. Three French data sets on aquatic plant and nutrients in rivers have been compiled (Grasmück et al., 1995; Demars & Thiebaut, 2008; Demars & Trémolières, 2009). All sites with $5 < \text{pH} < 8.5$ and $\text{HCO}_3 > 40 \text{ meq L}^{-1}$ have been included. There was no significant relationship between IBMR and BOD₅ in the Lorraine data set ($r^2 = 0.04$, $n = 95$, $P > 0.05$; Grasmück et al., 1995), and no significant organic pollution was indicated in the Northern Vosges ($n = 46$) and Alsace streams ($n = 74$) (Robach et al. 1996; Demars & Thiebaut 2008; Demars & Trémolières 2009). This allows us to study the effects of nutrient enrichment (N, P) and carbon availability ($\text{CO}_2$, $\text{HCO}_3$) on IBMR, independently of acidification and gross organic pollution.

As expected, pH explained more variability in IBMR than SRP (Fig. 1). If we were to remove the effect of pH on the IBMR, SRP is still significantly related to the IBMR but with a very weak $r^2 = 0.08$ (Fig. 1). Variance partitioning shows that pH is a much better predictor than SRP (Fig. 1). The predictive power of NH₄ and SRP was much weaker than $\text{pCO}_2$ and HCO₃ in the Northern Vosges and Alsace (Fig. 1). Variance partitioning also shows that the independent effect of SRP and NH₄ is extremely weak (Fig. 1). This more thorough analysis seriously questions the results and conclusions presented earlier by Haury et al. (2006).

Aquatic plant growth rate (Robe & Griffiths, 1994; Boedeltje, Smolders & Roelofs, 2005; Lambert & Davy, 2011), macrophyte occurrence and diversity (James et al., 2005; Lambert & Davy, 2011) and a river plant index (Mean Trophic Rank, Demars & Edwards, 2009) have previously been negatively related to nitrate (NO₃) concentrations. The above compilation of the three French data sets is also very valuable because there was a strong NO₃ gradient (20–9000 µg N L⁻¹) unrelated to PO₄ and NH₄ ($P > 0.16$; all data log transformed) and only weakly related to pH ($r = 0.38$, $P < 0.001$). However, the French index (IBMR) was barely negatively related to log-transformed nitrate ($r^2 = 0.02$, $P < 0.029$).

Predictions and classification system for macrophytes – LEAFPACS

LEAFPACS is based on five indices, each complying with the WFD. That is, the scores of the indices are calculated relative to what would be expected under reference conditions [called Ecological Quality Ratios (EQRs)]. LEAFPACS included two compositional indices for specific pressures (nutrient enrichment, RMNI; hydromorphology, RMHI), two richness indices [number of aquatic taxa, TAXA; number of ‘functional’ groups (FG)] and an index of the abundance of filamentous green algae (ALG). From these, a single ‘golden number’ (LEAFPACS) is calculated, to provide an overall quality assessment of the river. The two compositional indices are thought to reflect two specific pressures independently of other environmental variables: nutrient enrichment and hydromorphological impairment.

Nutrients (N, P) explained only 5% of the variability in plant composition (after accounting for other variables such as alkalinity and slope) in the British data used to develop LEAFPACS (Willby et al., 2009, p. 56), yet the authors concluded that ‘there remains a strong empirical basis for a focus on the assessment of nutrient-related pressures using macrophytes’. This clearly reflected what had been found previously (e.g. Dodkins et al., 2005), but our conclusion is the opposite: the relationship is too weak to derive a nutrient metric to be applied at specific sites or waterbodies according to the WFD (Demars & Edwards, 2009).

We tested LEAFPACS (and its individual components) on nitrate-rich, largely groundwater fed, calcareous lowland rivers draining eastern England with the data from
Demars & Harper (2002, 2005). The survey method (including all taxa, i.e. emergent and aquatic) is comparable to the LEAFFACs method, because both are based on the Mean Trophic Rank survey method (Holmes et al., 1999). Demars & Harper’s data (2002, 2005) were carefully collected (comparative survey) to obtain strong gradients of nutrient concentrations and hydromorphological impairment (because of the presence of weirs), largely independent from each other, as well as to assess the effect of the longitudinal connectivity. We know from these data that the longitudinal connectivity and hydromorphological effects were significantly affecting aquatic species composition, but SRP and NH₄ concentrations had no detectable effects (Demars & Harper, 2005).

The nutrient index (RMNI) failed to correlate with SRP (Fig. 2) in the water column and average bioavailable P in

Fig. 1 Test of IBMR (Indice Biologique Macrophytique en Riviere) across 215 sites (left-hand panels) in rivers of north-east France against soluble reactive phosphorus (SRP), pH and SRP after removing the effect of pH ($r^2 = 0.08$, $n = 215$, $P < 0.001$), inset summarises the variance partitioning; IBMR in a subset of 120 sites (right-hand panels) against predicted IBMR by SRP and NH₄ or HCO₃ and pCO₂, and against SRP after removing the effects of HCO₃ and pCO₂ ($r^2 = 0.06$, $n = 120$, $P < 0.001$), inset summarises the variance partitioning. The boundaries of water quality classes are indicated in the top left graph. All the chemical variables were log transformed (except pH) prior to data analyses. All analyses were simple linear (partial) regressions performed in CANOCO and tested with 1000 random Monte Carlo permutations.
the sediment (Fig. 2; iron oxide paper strip method; Sharpley, 1993). The abundance of green filamentous algae did correlate with SRP, but very weakly considering the very strong gradients (Fig. 2). There was no relationship between the hydraulic index (RMHI) and silt cover (Fig. 2), but weak relationships with mean water velocity ($r^2 = 0.10$, $n = 34$, $P = 0.07$) and mean depth (Fig. 2). There was also no relationship between the hydraulic index and river width/depth ratio (Fig. 2), a more integrative and independent measure of hydromorphological modification (Kemp, Harper & Crosa, 1999; A. Baattrup-Pedersen, pers. comm.).

Willby et al. (2009, p. 75–76) rightly questioned the validity of such simple tests to detect eutrophication (ecological response to nutrient enrichment). In so doing, however, they inevitably call into question the approach of macrophyte indices. The only way to progress is to resolve the interplay of the main components of the

![Image of graphs showing correlations between variables]

**Fig. 2** Test of LEAFPACS (predictions and classification system for macrophytes) component indices across 62 sites (34 sites for sediment P) in lowland calcareous rivers of eastern England. All indices are adjusted Ecological Quality Ratios ($^4$EQRs) and represent the deviation from expected site reference conditions: nutrient index ($^4$EQR$_{RMNI}$), algal index ($^4$EQR$_{ALGAE}$) and hydraulic index ($^4$EQR$_{RMHI}$). The boundaries of river quality classes (identical for all indices) are indicated in the top left graph. SRP was log transformed prior to data analyses. All analyses were simple linear regressions.
ecosystem, not just macrophytes (e.g. Wade et al., 2002; Jones & Sayer, 2003; Davidson et al., 2010). Clearly, if the relationships are too weak, the index should not be used. If the relationships are strong, then they need to be tested on independent data and the causality of the correlations must be ascertained, for example, with relevant plant traits, as recently done for pCO₂ (Demars & Trémolières, 2009).

Carefully crafted studies (comparative surveys) can be more powerful than the 6500 sites of LEAFPACS because we were able to test for the effect of a very strong gradient in SRP (10–3500 µg L⁻¹) under similar conditions of alkalinity and slope, and the effect of hydromorphological impairment by sampling up- and downstream of weirs. The conservation agency’s data (JNCC, 1-km-long survey) used in LEAFPACS were previously found inadequate to reflect the physical characteristics of lowland rivers (Demars & Edwards, 2009). Finally, the r² and probabilities reported by Willby et al. (2009) are generally not better than our detailed studies (Demars & Harper, 2005; Demars & Thiebaut, 2008; Demars & Edwards, 2009; Demars & Trémolières, 2009).

A brief look at LEAFPACS results (Fig. 3) questions the meaning of the final LEAFPACS ‘golden number’ (see Hatton-Ellis, 2008). In our test data set, it represents only a moderate status according to the WFD, that is, moderate quality of the structure and functioning of rivers, even though the number of observed taxa and functional groups greatly exceeded reference conditions for nearly all sites, except for sites situated <10 km from the source of the river.

The number of taxa calculated by the LEAFPACS method for reference conditions was only 3.8–6.2 per site, with 4.4 to 5.5 FG, with more FG than species at two sites (which is logically impossible). The difference in expected versus observed species richness is easily explained by site distance from source of the river (r² = 0.53, n = 62, P < 0.001), generally the best predictor of species richness in lowland rivers (e.g. Demars & Harper, 1998; Demars, 2002). Clearly, the data at hand in LEAFPACS were not appropriate to generate a regression model to predict species richness.

Moreover, the FG are only botanical (attribute-based) groups not specifically designed to link to ecosystem functions (see Demars, 1997; Willby, Abernethy & Demars, 2000). Functional groups of plants for macroinvertebrates have already been defined elsewhere (e.g. Harper et al., 1995; Kemp et al., 1999; Demars et al., 2012). As a result, the richness indices only come as weak weighting factors of the compositional indices.

Similarly, the definition of reference condition for algal cover (<2.5% cover) is completely arbitrary and inappropriate, as it does not take into account frequency magnitude of peak flows or grazing intensity (e.g. Biggs, 2000; Hillebrand, 2002; Jones et al., 2002). Clearly, these aspects of structure and functions in LEAFPACS are not appropriate.

There is generally a lack of appropriate uncertainty estimation in the results of indices. For example, the precision of the species scores (optimum values) is presented to two decimal places, but there is no measure of tolerance attached to the scores. While species cover and tolerance weighting may not change the site score significantly (Willby et al., 2009, p. 72), it would allow better quantification of the uncertainty of the site score. Another example is the number of sites required to classify a waterbody (typically a few kilometres long) with 95% confidence in the middle of an ecological class. This uncertainty does not reflect our ability to detect deviation from reference conditions, but only the precision in LEAFPACS scores. Clearly, when the individual components of LEAFPACS are tested (Fig. 2), it shows that more than 90% of the variability in environmental quality cannot be explained by the indices. Hence, we seem to have highly precise results with little idea of their accuracy.

The design of LEAFPACS is dominated by a pressure diagnosis attempt, rather than the structure and functional quality of rivers. Clearly, the nutrient index is failing to indicate reliably nutrient concentrations, yet the authors (and others e.g. Pardo et al., 2012) go on to derive environmental standards for nutrients in rivers. These
environmental standards are not appropriate (far too high) for both highland and lowland rivers (e.g. Demars & Harper, 2002; Demars & Edwards, 2007b), and their application could have detrimental effects on receiving waterbodies (e.g. Moss et al., 2003).

Site-specific uncertainties

Current limitations and opportunities

At present, site scores (index) are generally calculated using the weighted average of the species optima \((u_i)\) as follows (e.g. Kohler & Schneider, 2003):

\[
I_j = \frac{\sum_{i=1}^{n} u_i A_{ij} W_i}{\sum_{i=1}^{n} A_{ij} W_i}
\]

(1)

with \(n\) number of species, \(A_{ij}\) abundance of species \(i\) in site \(j\), \(t_i\) tolerance (ecological amplitude) and weight \(W_i\) equal to \(t_i^{-1}\) or \(t_i^{-2}\). The site uncertainties (\(SC_j\)) are then based on the degree of scatter of the species scores around the site average score \((I_j)\) as follows:

\[
SC_j = \sqrt{\frac{\sum_{i=1}^{n} (u_i - I_j)^2 A_{ij} W_i}{(n - 1) \sum_{i=1}^{n} A_{ij} W_i}}
\]

(2)

This estimation is simple but too conservative, because it does not take into account the uncertainties in estimating the species optima and tolerances. In addition, the method of weighted averaging only performs well if the species optima are evenly distributed over a large interval around the value of the environmental variable (ter Braak & Barendregt, 1986). In situations where all species recorded at the site have the same optimum, this assumption clearly does not hold and the rate of scatter will report zero uncertainty. The rate of scatter will also be zero in the case where only one species is recorded at a site; in this situation, a better estimate of the site uncertainty would be provided by the tolerance of the species. In addition to the precision of the site scores, there is also the issue of bias. If there is an uneven distribution of optima in the species pool, weighted averaging can give biased results. If the response curves for different species have unequal maxima, as some species are never as common as others, this can also affect the performance of the weighted averaging method (ter Braak & Barendregt, 1986).

Other sources of uncertainty (e.g. between surveyors, seasons or year of survey) have been investigated using replicate sampling and were found to have very little effect on the resulting macrophyte indices (e.g. Staniszewski et al., 2006; Pentecost, Willby & Pitt, 2009; Willby et al., 2009). These tests, however, only quantify the precision of the indices, not their ecological accuracy (Clarke et al., 2006).

The quantification of the ecological accuracy against specific pressures is routine in palaeolimnology (e.g. Birks et al., 1990) but rather exceptional in stream biomonitoring (e.g. Andrén & Jarlman, 2008; Dolédec & Statzner, 2008; Demars & Edwards, 2009). Ecological accuracy is also reported in (RIVPACS) in the form of a deviation from reference conditions (observed/expected, also known as EQRs) in the fauna (see e.g. Clarke, Wright & Furse, 2003). These uncertainties may be propagated to the derived metrics such as BMWP, AWIC or LIFE scores, but the causality of the stress remains untested (Wright, 2000, p. 23). The correlative strength of these indices or their EQRs should at least be tested directly against the related pressure of interest (e.g. Ormerod et al., 2006; Dunbar et al., 2010).

So far as we know, this site-specific ecological accuracy for a given pressure has not been reported for biotic indices in the context of stream biomonitoring, despite the availability of easy to use software specifically designed for this purpose (e.g. Line, ter Braak & Birks, 1994; Juggins, 2003). The reason is probably that most indices are simply too weak and cannot isolate one pressure from another (e.g. De Pauw & Roels, 1988; Demars & Edwards, 2009), although there are some notable exceptions (e.g. Andrén & Jarlman, 2008; Demars & Trémolières, 2009; Statzner & Béche, 2010).

Sample-specific errors of prediction by bootstrapping

Demars & Trémolières (2009) used a similar weighted average method to calculate the plant index following ter Braak & Barendregt (1986):

\[
I_j = \left( \frac{\sum_{i=1}^{n} A_{ij} u_i}{\sum_{i=1}^{n} t_i^{-1}} \right) \left( \frac{\sum_{i=1}^{n} A_{ij}}{\sum_{i=1}^{n} t_i^{-2}} \right)
\]

(3)

The optimum and tolerance were calculated from a calibration data set using a generalised linear model in which the logarithm of abundance is a quadratic function of the environmental variable. This is often referred to as a ‘Gaussian response curve’ (ter Braak & Looman, 1986). We use the test data set assembled by
Demars & Trémolières (2009) to illustrate our point. We focus on the CO$_2$ excess partial pressure ($E_{p}CO_2$) plant index because it was very strongly related to observed $E_{p}CO_2$ in the calibration ($r^2 = 0.73$, $n = 74$, $P < 0.001$) and test ($r^2 = 0.47$, $n = 37$, $P < 0.001$) data sets and supported by a biological mechanism (submerged plants ability to use HCO$_3$). We also report uncertainties in the PO$_4$-P macrophyte index from the same sites for comparative purpose. This index performed less well both in the calibration ($r^2 = 0.50$, $P < 0.001$) and in the test data sets ($r^2 = 0.27$, $P < 0.001$).

An improved bootstrapping approach, programmed in GenStat 13 (VSN International, Hemel Hempstead, UK), was used to investigate the ecological accuracy of the site index. Some additional technical details are provided in Supporting Information, Appendix S1. For each bootstrap replicate, the sites in the calibration data set were re-sampled and, following the method of Demars & Trémolières (2009), a ‘Gaussian’ response function was fitted by Poisson regression. It should be noted, however, that the abundances are in fact ordinal scores so a Poisson model does not strictly apply. At each site in the test data set, the species were then re-sampled by treating the cover abundance scores for each species as though they were individuals of that species. The site score for each bootstrap sample was estimated using the sample-specific estimates of the optima and tolerances. This ensured that uncertainties in estimating the optima and tolerances were propagated. However, the method of forming the bootstrap re-samples at each site underestimates the true variability because species that were absent in the data but for which conditions are potentially suitable at the site never occur in the bootstrap re-samples.

We plotted the plant indices and bootstrap estimates of their variability (95% confidence interval) against the observed values of the test data set (Fig. 4). As previously reported (ter Braak & Verdonschot, 1995), the tolerance weighting did not greatly affect the final site score (plant indices with $t_i$ weighting and indices without $t_i$ weighting were strongly correlated, $r = 0.96$, $n = 74$, $P < 0.001$). It is clear that our relative uncertainties in plant indices, calculated as the standard deviation/expected mean, for both $E_{p}CO_2$ (c. 30%) and PO$_4$-P indices (c. 13%, on a log scale), are indeed larger than is normally reported for macrophyte indices (c. 5–10%, Davey, Garrow & Glennie, 2009; Pentecost et al., 2009; Willby et al., 2009). The standard deviations calculated on transformed data (e.g. log transformed for the PO$_4$-P index) may be presented as a confidence interval either on the log scale (Fig. 4) or on the original data scale. However, Figure 4 also suggests that there is systematic bias, with smaller values of the environmental variable tending to be overestimated and larger values underestimated. The extent of this bias is difficult to quantify because the observed values of the environmental variables are themselves subject to uncertainty (c. 40% for $E_{p}CO_2$ and c. 15% for log-transformed SRP; B.O.L. Demars, unpubl. data) and we did not have replicate measurements. Note that our conservative ecological uncertainties would also be much larger with weaker bioindicators, such as IBMR and LEAFPACS, for SRP presented above.

**Future directions**

Assessing and linking directly (and simply) ecosystem structure and function may be more amenable to
conservation and regulatory government organisations and based on current ecological understanding of rivers (e.g. Allan & Castillo, 2007; Woodward, Friberg & Hildrew, 2010; Friberg et al., 2011), for example, active meandering river channel in lowland rivers (Robertson & Augspurger, 1999) versus straightened and deepened river sections with no backwaters or wetlands (Walter & Merritts, 2008); natural versus modified flow regimes (Renôfalt, Jansson & Nilsson, 2010); good versus no supply of tree propagules and dead wood (Gurnell, 1997); fast versus slow metabolic and nutrient (N, P) cycling rates (Fisher et al., 1998; Young, Matthaei & Townsend, 2008); large versus no population of migrating fish (Milner et al., 2007); diverse versus poor in-stream and riparian structure (e.g. Harper et al., 1995); low versus high ecological integrity (Covich et al., 1999); and no versus well-developed ecohydrological processes in catchments impacted by human activities (sensei Zalewski et al., 2008).

It does not suffice to list the symptoms of waterbodies, one needs to identify the ultimate causal mechanisms and assess the likely effects on the system (Vörösmarty et al., 2010; United States Environmental Protection Agency CADDIS initiative; http://www.epa.gov/caddis/) and, just as a medical general practitioner would do (Lovelock, 1991), suggest a treatment or ‘programme of measures’ to remedy the situation, using as much as possible natural processes.

Freshwater aquatic systems are hot spots of biogeochemical processes (e.g. Battin et al., 2009) and a source of extraordinary biodiversity, much of which is threatened with extinction (Strayer & Dudgeon, 2010). So, clearly, there is no room for weak science as it is already difficult enough to turn robust ecological knowledge into policy and management (e.g. Lawton, 2007).

Acknowledgments

This study was funded by the Scottish Government’s Rural and Environment Science and Analytical Services (RESAS) division. Thanks to Joanna Kemp, Simon Langan, Robert Ferrier, Mark Brewer, David Elston and two anonymous referees for their valuable comments on the draft manuscript, Alan Hildrew for his careful editing, Annette Baattrup-Pedersen for discussions on hydromorphological modifications and their effects on macrophytes, Mike Dunbar and Ralph Clarke for their insightful replies to BOLD’s queries on uncertainties, and finally the delegates of the macrophyte variability workshop 4–5 June 2009 in Llangammarch Wells for fruitful discussions which prompted this study. The authors have no conflict of interest to declare.

References


**Supporting information**

Additional Supporting Information may be found in the online version of this article:

**Appendix S1.** Estimation of species optima and tolerances and improved bootstrapping method to report site ecological uncertainty.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organised for online delivery, but are not copyedited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

*(Manuscript accepted 22 May 2012)*
Appendix S1

Estimation of optima and tolerances

The methods used were similar to those in Demars & Trémolières (2009). A Poisson regression with a log link was fitted to each species. The response function took the form:

\[ \ln(E[y]) = b_0 + b_1x + b_2x^2 \]

where \( E[y] \) is the expected abundance and \( x \) is the environmental variable. Provided that the estimate of \( b_2 \) was negative the optimum (\( u \)) and tolerance (\( t \)) for the species were estimated as follows:

\[ u = -b_1 / 2b_2 \]
\[ t = 1/\sqrt{-2b_2} \]

If this gave an estimate of the optimum that was lower than the minimum or higher than the maximum value of the environmental variable in the data set, the optimum was set to the minimum or maximum value, respectively.

If the estimate of \( b_2 \) was positive a linear model was fitted in which the response function took the form:

\[ \ln(E[y]) = b_0 + b_1x \]

The optimum was then set to the minimum value of the environmental variable if \( b_1 \) was negative and to the maximum value if \( b_1 \) was positive. In the former case the tolerance was set such that 84% of the distribution was below a threshold of \( u+t \) and in the latter case it was set such that 84% of the distribution was above a threshold of \( u-t \).
**Bootstrapping**

A subset of sites from the calibration data set was selected randomly with replacement to create a bootstrap sample of the same size as the calibration data set. This is similar to the approach taken by Birks *et al.* (1990) to estimate the part of the prediction error that is due to the estimation of species optima and tolerances. However, Birks *et al.* (1990) estimated the part of the prediction error due to variation in species abundances at a given value of the environmental variable, by calculating the root mean square error across all samples in the test data set of the difference between the observed and mean bootstrap values. This part of the prediction error is then a constant across all samples. Our approach of resampling the cover abundance scores allows uncertainties to be estimated separately for each site.