Optimizing taxonomic resolution and sampling effort to design cost-effective ecological models for environmental assessment

Joseph R. Bennett[1]*[†], **Danielle R. Sisson**[1][†], **John P. Smol**[2], **Brian F. Cumming**[2], **Hugh P. Possingham**[1,3] and **Yvonne M. Buckley**[1,4]

[1]*School of Biological Sciences, University of Queensland, Brisbane, Qld 4072, Australia;* [2]*Department of Biology, Paleoecological Environmental Assessment and Research Lab (PEARL), Queen's University, Kingston, ON K7L 3N6, Canada;* [3]*School of Life Sciences, Silwood Park Imperial College London, Ascot, Berkshire SL5 7PY, UK; and* [4]*Zoology Department, School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland*

Summary

1. Predictive models relating ecological assemblages to environmental conditions are widely used in environmental impact assessment and biomonitoring. Such models are often parameterized using comprehensive ecological sampling and taxonomic identification efforts.

2. Limited resources mean that expensive sampling and analytical procedures should be planned to maximize information gain and minimize unnecessary expense. However, there has been little consideration of cost-effectiveness in parameterizing predictive models using ecological assemblages and no explicit consideration of cost-effectiveness in balancing investment in the crucial aspects of sample size and taxonomic resolution.

3. Using lacustrine diatom (Bacillariophyceae) assemblages from four large-scale (*c.* 77 000–1·3 million km^2) data sets containing between 207 and 493 lakes, we address the following questions: (1) how does taxonomic resolution affect information content; (2) how does sample size affect information content for different taxonomic resolutions; and (3) what are the most cost-effective strategies for constructing environmental assessment models using diatom assemblages across a range of budgets? We use weighted averaging regression models for pH, phosphorus, salinity and lake depth and realistic data collection costs to examine the relationship between cost and model information content (R^2 and root mean squared error of prediction).

4. For diatom-based models, finer taxonomic resolutions almost always provide more cost-effective information content than collecting more samples, with (morpho)species being the most appropriate taxonomic resolution for nearly all budget scenarios. Information content exhibits an asymptotic relationship with sample size and budget, with greatest information gain during initial sample size increases, and little gain beyond *c.* 100 samples. Smaller sample sizes can also achieve surprising predictive power in some cases, suggesting low-cost regional models may be achievable. However, caution is necessary in such an approach, because spatial dependencies in predictions may be missed and analogues with predicted assemblages may be poor.

5. *Synthesis and applications.* We demonstrate the utility of explicitly considering cost estimates to determine optimal sampling effort and taxonomic resolution for ecological assemblage models. For large, regional biomonitoring programmes, cost-effective sampling could save millions of dollars. Our framework for determining optimal trade-offs in ecological assemblage models is easily adaptable to other taxa and analytical techniques used in biomonitoring and environmental assessment.

*Correspondence author. E-mail: j.bennett5@uq.edu.au
[†]These authors contributed equally.

Introduction

Predictive models linking environmental conditions with the composition of ecological assemblages are frequently used for biomonitoring and environmental assessment. Large-scale programmes using these models, such as the US National Lakes Assessment (NLA) (USEPA 2013) and the European Union's Water bodies in Europe – Integrative Systems to Assess Ecological Status and Recovery (WISER) Program (WISER Consotrium 2014), can incur substantial costs. Given limited funds and the need to balance costs between monitoring and mitigation efforts, there is great incentive to improve the cost-effectiveness of biomonitoring and environmental assessment programmes.

Efforts to make quantitative assessments and monitoring more efficient have often focused on either taxonomic sufficiency or sample size. Taxonomic resolution and sample size both affect model information content (i.e. the accuracy of model predictions) and project costs (e.g. King & Richardson 2002; Marshall, Steward & Harch 2006). Collecting an insufficient sample size or using an identification resolution that is too coarse can mean information content is inadequate for generating reliable results for decision making. However, it is costly and time consuming to collect samples and undertake detailed identifications. If sample size and taxonomic resolution are beyond what is necessary to construct acceptable models, resources that could be better spent in other areas such as direct management are wasted.

Studies of taxonomic sufficiency in aquatic ecosystems have led to a variety of recommendations (Carter & Resh 2001; Jones 2008). Across several taxonomic groups, Bevilacqua *et al.* (2012) found that the utility of higher-level taxa for replicating species-level patterns in community dissimilarity was highest in taxa that had relatively low species-level diversity. For macroinvertebrate assemblages, Marshall, Steward & Harch (2006) found that family-level and genus-level analyses were sufficient to observe patterns in community richness and turnover that were recorded using species-level analyses, while Siqueira *et al.* (2012) found that small subsets of taxa could be isolated that could act as reliable surrogates for broader community-level compositional patterns. Heiri & Lotter (2010) found that analyses focusing largely on genus-level taxonomy for chironomid (Order Diptera, Family Chironomidae) assemblages were likely sufficient to predict environmental changes and that coarser taxonomic resolution may have the advantage of decreasing misidentification and discrepancies. However, King & Richardson (2002) and Greffard, Saulnier-Talbot & Gregory-Eaves (2011) found that substantial information could be lost in moving from species-level to coarser taxonomic resolution in ordination-based analysis of chironomid assemblages, suggesting that some species have environmental preferences that are not captured at the genus level.

For microalgae, Heino & Soninen (2007) found that information loss in ordinations and taxonomic richness models was greater at reduced taxonomic resolution than it was for macroinvertebrates. However, Carneiro, Bini & Rodrigues (2010) found that species-level and genus-level spatiotemporal patterns in microalgae assemblages were highly similar, and Wunsam, Cattaneo & Bourassa (2002) found that genus-level analyses could potentially outperform species-level analyses for ordination-based modelling of community responses to some environmental variables.

Analyses regarding sample size have often focussed on determining an adequate subsample of organisms counted per sample (e.g. King & Richardson 2002; Payne & Mitchell 2009). However, despite the fact that sample collection costs can be a substantial component of environmental assessments (e.g. Hughes & Peck 2008), there have been few studies regarding the number of study sites necessary to accurately model aquatic assemblage relationships with environmental variables. Using diatom-based salinity models with species-level resolution, Wilson, Cumming & Smol (1996) found relatively little information gain after a sample size of *c*. 100 lakes. Reavie & Juggins (2011) found that 40–70 sample lakes were the minimum necessary for diatom-based models of nutrient-related environmental stress, while Weilhoefer & Pan (2006) found that at least five samples were necessary to characterize diatom community composition within a hydrologically complex wetland.

Funding is a key constraint in environmental assessment programmes, and decisions on sample size and taxonomic resolution should be made together to derive the best model possible within the available budget. However, explicit consideration of cost trade-offs has been rare in analyses of taxonomic sufficiency and sample size in aquatic environmental assessment. Several studies (e.g. Ferraro *et al.* 1994; Thompson, Riddle & Stark 2003) have considered sample sorting techniques and taxonomic sufficiency for benthic invertebrates, using preparation and identification time as a measure of cost. Marshall, Steward & Harch (2006) estimated actual costs of sample preparation and taxonomy in concluding that family-level analyses were most cost-effective for analysing aquatic macroinvertebrates among 29 Australian lakes. To our knowledge, there has been no explicit consideration of cost-effectiveness in balancing sample size and taxonomic effort in aquatic environmental assessment.

Here, we use large-scale data sets and explicit cost estimates to determine the optimal combinations of

sample size and taxonomic resolution for environmental assessment models with a commonly used bioindicator. Specifically, we use assemblage vs. environment models based on diatom (Bacillariophyceae) assemblages from data sets totalling >1200 lakes to answer the following questions: (1) how does taxonomic resolution affect information content; (2) how does sample size affect information content for different taxonomic resolutions; and (3) what are the most cost-effective strategies for constructing environmental assessment models using diatom assemblages across a range of budgets? In addressing these questions, we demonstrate a technique for determining the most cost-effective analytical strategy for a variety of applications in biomonitoring and environmental assessment.

Materials and methods

STUDY ORGANISMS

We chose to use diatoms for our analyses because they are used extensively in biomonitoring and environmental assessment, for projects related to water quality, climate change, ecological integrity, fisheries, airborne contaminants and fire history (Smol & Stoermer 2010). Specific environmental variables for which diatoms are used include: pH (Dixit *et al.* 1993; Dixit & Smol 1994; Ginn, Cumming & Smol 2007), dissolved organic carbon, acid neutralizing capacity, monomeric aluminium (Dixit *et al.* 1993), salinity (Wilson, Cumming & Smol 1996), phosphorus (Hall & Smol 2010), nutrient availability (Berthon, Bouchez & Rimet 2011) and heavy metals (Morin *et al.* 2008). In our data sets, the large sample sizes and established relationships between diatoms and environmental variables offer a 'gold standard' against which lower levels of taxonomy and sample size can be compared.

DATA SETS

To examine changes in information content across a range of taxonomic effort and sample size, we incrementally decreased taxonomic resolution and sample size in large-scale data sets for modelling pH, total phosphorus (TP), salinity and lake depth. For pH models, we used two data sets: 493 lakes from a *c.* 671 000 km^2 study area in north-eastern North America (Ginn, Cumming & Smol 2007; Bennett *et al.* 2010) and 488 lakes from a *c.* 1·3 million km^2 study area in north-western Europe (Battarbee *et al.* 2001). For both data sets, pH was measured using handheld pH metres, using methods that were checked for consistency by Ginn, Cumming & Smol (2007) and Battarbee *et al.* (2001). For TP models, we used a subset of 233 lakes from the USEPA Environmental Monitoring and Assessment Program (EMAP; http://www.epa.gov/emap/) located within a *c.* 410 000 km^2 area of the north-eastern North American data set used for pH models (Bennett *et al.* 2010). We used the EMAP lakes for TP analysis because TP was collected using standardized protocols in EMAP surveys. Three lakes from the EMAP subset were not included because TP had not been recorded for these lakes. TP measurements were a combination of suspended and dissolved phosphates, measured in μg L^{-1}. For salinity and lake depth analyses, we used data from 207

lakes in a *c.* 77 000 km^2 area in the southern interior of British Columbia, Canada (Wilson, Cumming & Smol 1996). Salinity was measured as specified in Wilson, Cumming & Smol (1996) and calculated as the summed concentration (g L^{-1}) of Na, K, Ca, Mg, dissolved inorganic carbon, SO$_4$ and Cl. Lake depth was measured at the deepest point in the lake using sonar and a depth probe. TP, salinity and lake depth were log-transformed to improve model fit between measured and predicted values. For each of the four data sets, a single sediment core was collected from the deepest part of every lake as per common protocol in large-scale lacustrine environmental assessments (e.g. Dixit & Smol 1994), and a surface-sediment sample was taken from the top 1–5 cm from the core. This sampled diatom species living in the lakes from recent years. Minimum counts of between 200 and 500 valves per sample (depending on the individual data set) were identified to the lowest feasible taxonomic resolution (usually morphospecies) using protocols outlined in Wilson, Cumming & Smol (1996), Battarbee *et al.* (2001) and Ginn, Cumming & Smol (2007). All variables explained significant variation in community composition in redundancy analyses (RDAs; Table S1, Supporting information). Diatom assemblages were analysed at the following taxonomic resolutions: morphospecies (hereafter, 'species'), genus and family. We reduced taxonomic resolution to genus and family levels using Round, Crawford & Mann (1990) and online diatom data bases (Spaulding, Lubinski & Potapova 2010; Guiry & Guiry 2013). Taxa that had a minimum ≥1% relative abundance in any lake and were found in >1 lake were used in analyses.

STATISTICAL ANALYSES

We used weighted averaging (WA) regression and calibration to construct models relating diatom assemblages to environmental variables. WA regression estimates a species' optimum for an environmental variable as the abundance-weighted average of the environmental variable at sites where the species is present, while WA calibration estimates an environmental variable using the abundance-weighted optima of the species present on a site (Juggins & Birks 2012). We chose to use WA because it is a commonly used technique in diatom-based environmental assessment and is known to perform well compared to other popular techniques based on maximum likelihood and compositional dissimilarity (Juggins & Birks 2012). We used WA with inverse deshrinking of fitted environmental values and downweighting by species tolerances, which offered slightly improved predictive power in initial tests. However, results in initial tests using other WA techniques were very similar and would not have affected our conclusions.

To answer question #1 (*How does taxonomic resolution affect information content?*), we examined model predictive power as a measure of information content across different taxonomic resolutions using the full data sets. We used the coefficient of determination (R^2) between the measured and predicted (using diatom optima) environmental variables and the root mean squared error of predicted vs. the measured environmental variables (RMSEP) to measure model information content. R^2 measured the variance in the test data that were explained by the model (i.e. the strength of the linear relationship between the actual and predicted values) and was useful for general comparisons among all models; RMSEP allowed estimation of overall model accuracy, including any systematic bias. Measurements of model performance tend to be over-optimistic when predicting

environmental variables using the same data being used to model species optima (Juggins & Birks 2012). Therefore, we used bootstrap cross-validation to derive realistic R^2 and RMSEP to examine model performance on full data sets at different taxonomic resolutions. Bootstrapping is a standard statistical technique to correct for overly optimistic fit estimates in ecological models (Vaughan & Ormerod 2005) and is commonly used in environmental assessment models (e.g. Juggins & Birks 2012). Data were resampled 1000 times with replacement and bootstrap-corrected R^2 and RMSEP were calculated. Mean bias was also recorded. In this case, bias is a measure of the directionality of the error between the predicted and the actual measurement of the environmental variable. WA models with bootstrap cross-validation were constructed using the 'rioja' package in R (Juggins 2012; R Core Team 2013).

To answer question #2 (*How does sample size affect information content for different taxonomic resolutions?*), we examined model performance across taxonomic resolutions and sample sizes, using fully independent cross-validation with >20% of original samples (100 samples for pH data sets, 50 samples for other data sets) reserved for testing each model run. Where sufficient data are available, testing models against independent subsamples is the most rigorous method for testing model fit (Vaughan & Ormerod 2005). At every combination of taxonomic resolution level and sample size increment, 100 random sets of test and validation sample sets were generated, and R^2 and RMSEP were calculated for each set. We used random subsampling as opposed to even sampling along gradients (Reavie & Juggins 2011; Telford & Birks 2011), to represent the conservative scenario that environmental gradients are not known prior to sampling and potentially not fully represented in sampled lakes. For pH data sets from North America and Europe, random subsample sizes started at 375 lakes, then decreased in increments of 25–100 lakes, 10–50 lakes and five lakes until a subsample size of 10 lakes. For TP, subsamples started at 175 lakes, with increments diminishing as above. For salinity and lake depth, subsamples started at 150 lakes.

In addition to assessing model information content, we examined a subset of models for spatial autocorrelation in residuals for predicted vs. measured values of the independent subsamples. We did so because spatial autocorrelation in residuals can indicate violation of statistical independence among samples (Fortin & Dale 2005) and can suggest that additional spatially autocorrelated environmental or biological processes are affecting community patterns (Telford & Birks 2009; Beale *et al.* 2010). We examined spatial autocorrelation in residuals using spline correlograms (Bjørnstad & Falck 2001), which are modified correlograms using a smoothing spline to examine patterns in autocorrelation across distances, and bootstrapping to derive confidence bands for estimated autocorrelation. Spline correlograms offer several improvements over traditional correlograms and have been used extensively in ecological models (e.g. Keitt *et al.* 2002; Seabloom *et al.* 2005; Bennett 2014). Since examining correlograms across the 29 700 test models was impossible, we chose a small random subset of models at several intervals for each variable. Further details on spatial autocorrelation analyses are provided in the Supporting Information.

To answer question #3 (*What are the most cost-effective strategies for constructing environmental assessment models using diatom assemblages across a range of budgets?*), we estimated the costs of

sampling and analysis using the 2007 and 2012 USEPA NLA as a basis for estimates. The NLA is an extension of the sampling programme used in the eastern North America data set of Bennett *et al.* (2010) and uses similar protocols to the European samples collected for pH and the western North American data set collected for salinity and lake depth. We separated costs into sample collection and environmental analysis costs (fixed values per lake) and costs associated with identification (a fixed amount for sample preparation, and a variable amount for identification, depending on taxonomic resolution). In 2007 and 2012, the NLA used a fixed cost for sampling of $5800 (all values USD) per lake (R. Mitchell, pers. comm.). Costs for measuring pH and depth were considered to be zero, as pH and depth were measured *in situ* using a metre or probe. Cost of TP and salinity chemical analyses were calculated based on the average of 10 randomly chosen published rates in the USA. Estimated cost was $30 for TP and $80 for salinity. Sample preparation cost was estimated at $35 per sample (D. Charles, pers. comm.). Species-level identification cost, including sample preparation, ranged from $250 per sample (lowest value, 2012 estimate) to $500 per sample (highest value, 2007 estimate) (R. Mitchell, pers. comm.). We used the value of $500 per sample ($465 per sample, with preparation cost subtracted) as the most conservative and likely most realistic estimate (R. Mitchell, pers. comm.).

To determine genus- and family-level identification costs, experienced diatomists were surveyed for their estimates of the proportion of time taken to identify to genus or family level, compared to a baseline of species-level identification. Out of 40 experts contacted, 29 responded. The average estimated percentage of species identification time was used to calculate the costs of identification to genus and family levels, as a proportion of the species-level identification cost. Reductions in costs associated with higher taxonomic resolutions were only applied to identification time, as slide preparation time was assumed to be consistent among taxonomic resolutions. Based on the responses received, the mean percentage of identification time compared to species level was 44% for genus level and 29% for family level. Using a $500 baseline cost for sample preparation and identification to species level, identification to genus would thus cost $239·60 and family $169·85, including sample preparation.

The total cost per subsample was calculated as follows:

$$\text{Total cost} = N(F + A + S + T) \qquad \text{eqn 1}$$

where N is the number of lakes in the subsample, F is the field cost in collecting the sample ($5800), A is the analytical cost for the environmental variable of interest, S is the slide cost for preparing the sample for identification under the microscope ($35) and T is the cost of identifying diatoms to the desired taxonomic resolution (proportion of time spent identifying × $465). To determine relationships between total cost and model information content at each taxonomic resolution, total costs were plotted against mean RMSEP for independent subsamples. Although the sample collection cost of $5800 is a realistic cost paid for an actual large-scale biomonitoring programme, this component of the total cost is likely to be variable, depending on regional conditions. To determine whether the results outlined below were robust to different sample collection costs, we ran additional cost scenarios with sampling costs estimated to be $1000, $500, $250 and $100 per sample.

Results

Q1. HOW DOES TAXONOMIC RESOLUTION AFFECT INFORMATION CONTENT?

Relaxing taxonomic resolution from species to genus and family levels had a consistently negative influence on model information content. For North American pH measurements, bootstrap-corrected R^2 for predicted vs. actual values decreased by 0·09 from species to genus and by 0·13 from genus to family (Fig. 1). RMSEP was 26% higher for genus than species and 22% higher for family than genus. Across all variables, R^2 decreased by a mean of 0·12 from species to genus and 0·07 from genus to family, while RMSEP increased by a mean of 20% from species to genus and 10% from genus to family (Figs S1–S4, Supporting information). Log salinity exhibited the greatest loss in information content as taxonomic resolution was relaxed, with a decrease in bootstrap-corrected R^2 of 0·21 and a 44% increase in RMSEP from species- to genus-level analysis (Fig. S3, Supporting information). Log TP exhibited the smallest loss in information content (Fig. S2, Supporting information); however, the TP models had the poorest fit among all variables (species-level: $R^2 = 0.35$; $P < 0.0001$).

Q2. HOW DOES SAMPLE SIZE AFFECT INFORMATION CONTENT FOR DIFFERENT TAXONOMIC RESOLUTIONS?

Information content followed a similar pattern of decline as sample sizes were lowered for species-level, genus-level and family-level analyses and was highest for species-level analyses at every sample size. For North American pH predictions, confidence intervals for mean RMSEP in species-level, genus-level and family-level analyses overlapped at the smallest sample size (10 lakes; Fig. 2). Predictive ability for North American species-level pH models using 375 lakes was very high for independent data (mean $R^2 = 0.84$; mean RMSEP = 0·46; Fig. 2). This diminished only slightly by 100 sample lakes and then declined more quickly until a minimum mean R^2 of 0·61 and maximum mean RMSEP of 0·77 were reached at a sample size of 10 lakes. Mean bias also increased in a similar fashion, from *c.* 0 at 375 samples to *c.* 0·07 pH units at a sample size of 10 lakes.

For other variables, information content loss as sample size and taxonomic resolution diminished was broadly similar to that of North American pH (Figs S7–S10, Supporting information). However, information content for log salinity and log lake depth diminished slowly until <50 samples (Figs S9 and S10, Supporting information), and confidence intervals for RMSEP of predicted log TP overlapped among species-level, genus-level and family-level analyses below *c.* 75 samples. Predictability of log TP was again considerably lower than that of other variables (Fig. S8, Supporting information). Spatial
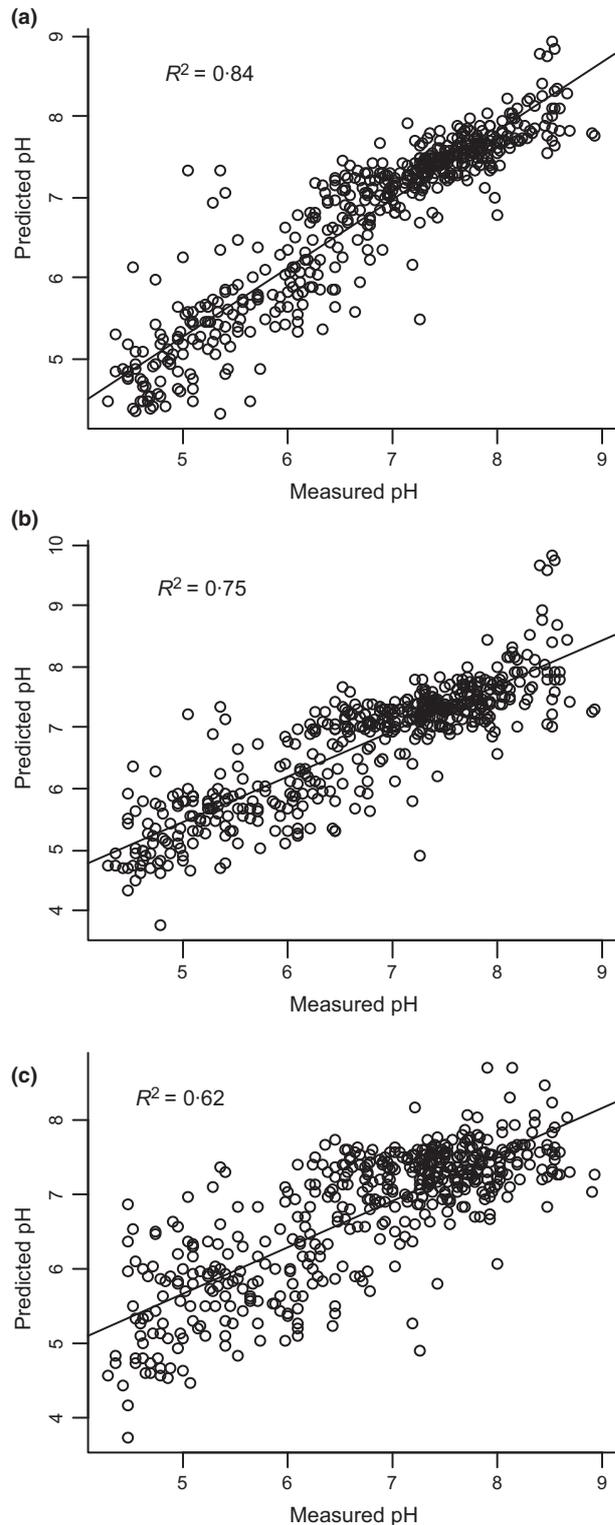

Fig. 1. Predicted vs. actual pH, at (a) species, (b) genus and (c) family taxonomic identification level from the North American pH data set. Figures show bootstrap-corrected R^2 values. RMSEP values were (a) 0·46, (b) 0·59 and (c) 0·72. $P < 0.0001$ across all models.

autocorrelation in residuals between measured and predicted variables was absent or extremely uncommon for most variables, with the exception of European pH (Table
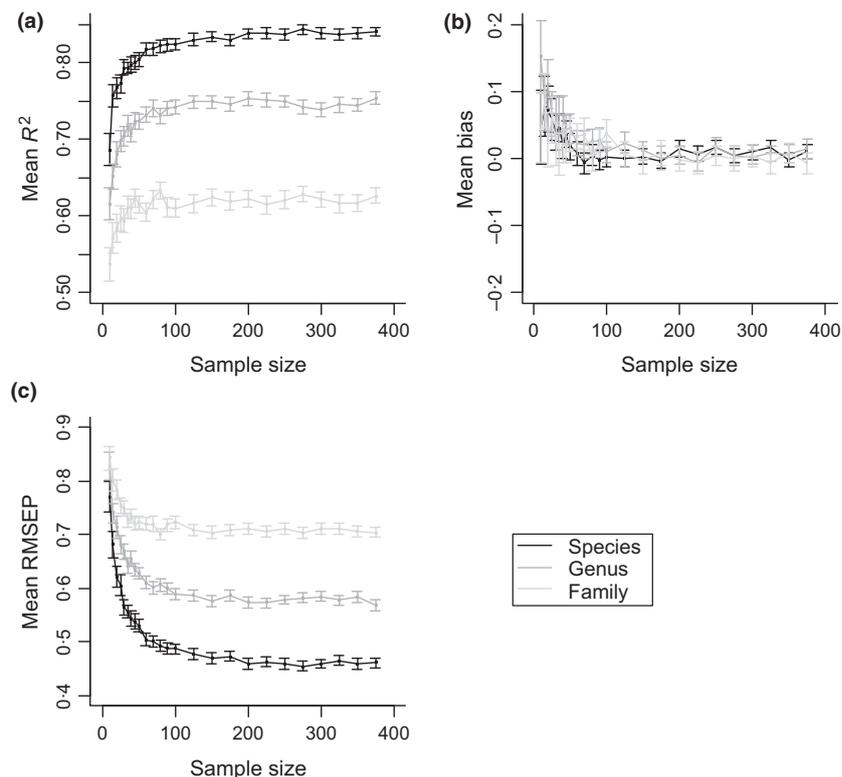
Fig. 2. Mean R^2 (a), mean bias (b) and mean root mean squared error (RMSEP) (c) for predicted vs. measured pH in independent subsamples from the North American pH data set, across 100 iterations at each subsample size. Error bars represent 95% confidence intervals.

S2, Supporting information), for which 83% of models exhibited significant spatial autocorrelation in residuals, and log TP, for which 17% of models exhibited significant spatial autocorrelation in residuals.

Q3. WHAT IS THE MOST COST-EFFECTIVE STRATEGY FOR CONSTRUCTING ENVIRONMENTAL ASSESSMENT MODELS USING DIATOM ASSEMBLAGES ACROSS A RANGE OF BUDGETS?

For nearly all budgets across all variables, species-level identification was the most cost-effective option (Figs 3 and S11–S14, Supporting information). Cost savings at lower taxonomic resolutions did not allow enough additional samples to be collected to overcome the sacrifice in information content. However, confidence intervals for mean RMSEP overlapped across taxonomic resolutions for European pH and log lake depth at the lowest budgets (*c.* <\$100 000) (Figs S11 and S14, Supporting information) and for log TP across most budgets (Fig. S12, Supporting information).

Using lower field cost scenarios, genus- and family-level analyses became slightly more cost-effective than species-level analyses at low budgets for some variables (Figs S15–S19, Supporting information). For example, RMSEP was lowest for genus-level European pH predictions at budgets *c.* <\$10 000 with field cost scenarios of \$250 per sample and \$100 per sample. However, for nearly all budget scenarios and variables, species-level analyses were still more cost-effective, regardless of field costs. In the limited budget ranges where genus- and family-level analyses were

more cost-effective, confidence intervals on RMSEP for a given budget overlapped with those of species-level analyses. For log TP analyses, confidence intervals for species-level, genus-level and family-level analyses overlapped in several budget scenarios, again reflecting the greater uncertainty in log TP analyses at any taxonomic
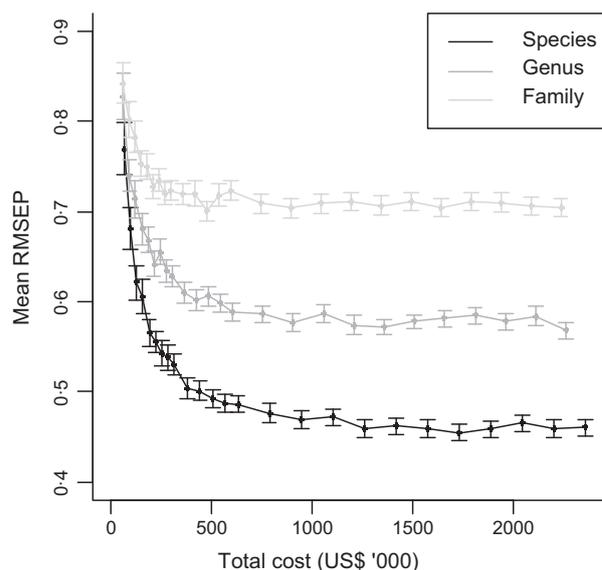

Fig. 3. Root mean squared error of prediction (RMSEP) vs. combined cost of sampling, slide preparation and specimen identification (US\$) for North American pH models. Each point represents a sample size increment, from 10 to 375. Error bars represent 95% confidence intervals.

resolution. Only for log depth prediction with field costs of $100 per sample and budgets <$10 000 was information content clearly greater for genus-level predictions than species-level predictions.

Across all variables, the rate of improvement in information content was greatest at the lowest budgets. For the North American pH data set, mean RMSEP for species-level identification decreased from >0·74 pH units at a cost of *c*. $63 000 and a sample size of 10 to 0·48 pH units at a cost of *c*. $630 000 (Fig. 3) and a sample size of 100. Beyond this level, additional spending would garner little improvement in RMSEP. For example, for a budget of *c*. $1 million, between 150 and 175 samples could be collected resulting in a mean RMSEP of 0·46–0·47 for species-level identification. Other variables followed similar patterns, with little improvement in RMSEP beyond budgets of *c*. $500 000 to $600 000 (Figs S11–S14, Supporting information).

Discussion

Quantitative models relating biological assemblages to environmental variables are vital components of many environmental assessments and biomonitoring programmes, including those predicting ecological impacts of pollution (e.g. Turak *et al.* 1999; Donohue *et al.* 2009), and paleoenvironmental reconstructions (Juggins & Birks 2012). The costs of constructing such models can be substantial and must be balanced against use of resources for management. Thus, it is crucial to choose the most cost-effective taxonomic resolution and sample size that will satisfy study objectives. Our analyses show that, for large regional diatom data sets, lowering taxonomic resolution and sample size both result in information loss. However, near-maximum information content can be achieved with considerably less than original expenditures.

For nearly all budget scenarios and variables, analysis to species level was most cost-effective. Even for scenarios where sample collection costs were artificially lowered, species-level analyses were almost always either more cost-effective, or as cost-effective, as other strategies. Our baseline estimate of field costs represents realistic costs of sample collection for large-scale data sets that include both remote and non-remote sites. Although field costs will vary from region to region, they are usually a major component of costs for larger environmental assessments (Hughes & Peck 2008) and are likely to be considerably higher than our lowest artificial estimate of *c*. $100 per sample. Thus, species-level analyses are more cost-effective than coarser-resolution analyses in nearly all conceivable cost scenarios for diatom-based environmental assessment models.

Increasing sample size (and cost) beyond *c*. 50–100 samples (Figs 2 and S7–S10, Supporting information) appeared to be inefficient, adding little information content. This sample size is considerably smaller than the size of the individual data sets, suggesting that substantial savings could be achieved in the collection and enumeration of regional data sets. For example, collection and analysis of a similar data set to our 493-lake North American pH data set would cost *c*. $3 million, using our cost estimates. Negligible loss in prediction error (RMSEP) could be achieved with a budget of *c*. $1 million and 150–175 samples, or even *c*. $630 000 and 100 samples. Similar savings could be achieved in other data sets (Figs S11–S14, Supporting information).

Across all variables, the gain in information content was most rapid with initial budget increases (Figs 3 and S11–S14, Supporting information). Surprising predictive power could be achieved at relatively low budgets (and sample sizes) using randomly chosen data subsets, for pH and salinity in particular (Figs 3 and S11–S14, Supporting information). This suggests that when budgets are limited, even small sample sizes (<25) could allow prediction of regional patterns for some variables. If management decisions do not depend on near-optimal information content, this could be an acceptable strategy for study design.

However, some caution is necessary in using models based on small sample sizes, for several reasons. Environmental assessments are often used to inform regulatory decisions and in such cases, near-maximum achievable accuracy may be important. For example, paleolimnological assessments of lakewater pH change in the USA were key components influencing environmental policy decisions but were subjected to extensive scrutiny and quality control considerations (Smol 2008). Although our results showed that confidence intervals on information content can be reasonably low at small sample sizes, some of our individual runs performed worse than others. For example, the lowest R^2 for actual vs. predicted North American pH among models using a 10-lake subsample was 0·34 (RMSEP = 1·32); for those using 100-lake subsamples, the lowest R^2 was 0·73 (RMSEP = 0·56). An individual 10-sample study that performed relatively poorly might have unacceptable accuracy, while an individual 100-sample study that performed relatively poorly might still have acceptable accuracy.

Choosing a small number of samples to infer patterns across large study areas also risks inadequate characterization of spatial autocorrelation, which can violate assumptions of independence among samples (Fortin & Dale 2005), and potentially lead to spurious models where spatially autocorrelated non-causal variables mask the effects of causal variables or biological processes such as dispersal limitation (Telford & Birks 2009; Beale *et al.* 2010). Choosing the number and configuration of samples to detect spatial autocorrelation in models is not straightforward (Fortin, Drapeau & Legendre 1989; Gilbert & Bennett 2010); however, sampling design should be sufficient to construct a distance matrix for examining spatial autocorrelation throughout the area where environmental response will be predicted. Our data sets allowed us to test both predictive power and residual spatial autocorrelation for large independent data sets and to show that

our models were highly accurate. Prediction residuals showed little or no spatial autocorrelation, with the exception of TP and European pH models. For TP, spatial autocorrelation in model residuals and relatively low model information content suggest that other spatially autocorrelated variables (e.g. pH) are important additional drivers of community composition. For European pH predictions, previous studies (Bennett *et al.* 2010; Juggins 2013) have suggested that pH is a strong driver of community composition in these data. Residual spatial autocorrelation may be due to unmeasured environmental variables related to the broad climatic and geological gradients in the data set, which spans from Wales to northern Finland. In situations where patterns are less understood and sample sizes are low, testing spatial autocorrelation in residuals in independent data could be difficult.

While we found remarkably consistent patterns in information content and cost-effectiveness across sample sizes and taxonomic resolutions, some differences among the data sets stand out. For log TP, R^2 was generally lower than that of other variables, mean RMSEP did not decrease as sharply as sample sizes increased, and information content among taxonomic groups largely overlapped (Fig. S12, Supporting information). This may be due to the mobility of phosphorus in the environment and its seasonal and interannual fluctuations in lakes (Wetzel 2001). Diatoms from surface sediments integrate water-quality conditions over a period potentially spanning several seasons, so any short-term changes in phosphorus may be homogenized by the diatoms in surface sediments. A single sample used for TP analysis may therefore introduce considerable variability into model predictions. In such cases where causal links between measured variables and sampled assemblages are occluded, caution is warranted in model interpretation (Juggins 2013). Salinity, lake depth and pH exhibit less annual and interannual variability than phosphorus and consequently are easier to model from surface-sediment assemblages (Hall & Smol 2010).

Salinity models also showed somewhat distinct patterns, with greater information content loss at lower taxonomic resolutions, and apparent downward bias for genus-level prediction of high-salinity lakes (Fig. S3, Supporting information). This is due to large variations in salinity optima within a single genus. For example, species within the genera *Craticula, Cyclotella* and *Cymbella* had high variation in salinity optima, and this variation was homogenized in genus-level analyses. In contrast, in the pH data set, species within dominant genera such as *Eunotia* and *Cyclotella* tended to share relatively consistent pH optima, and consequently, information loss was lower when moving from species-level to genus-level analyses. These differences suggest the possibility of more recent evolution of salinity niches than pH niches in diatom assemblages. However, characterization of the diatom phylogeny is still relatively coarse (Theriot *et al.* 2010),

and the potential utility of sub-generic phylogenetic relationships for understanding diatom environmental niches is currently unknown.

Our framework for optimizing cost-effectiveness in environmental assessment models is adaptable to a variety of analyses where model information content can be broken down into sampling, taxonomic and analytical components. These include predictive models using additional taxonomic groups (e.g. Payne & Mitchell 2009; Heiri & Lotter 2010), and multiple-variable or composite-variable predictions (e.g. Reavie & Juggins 2011). These also include techniques using compositional dissimilarity to detect impacts on communities (e.g. Marshall, Steward & Harch 2006; Bevilacqua *et al.* 2012) or infer environmental variables based on multivariate analogues (e.g. Simpson 2012). The ability to discriminate among test sites or the strength of correlations between assemblage and environmental distance matrices can be measured against the costs of various taxonomic resolution and sample-size scenarios. For this type of analysis, sampling a larger number of sites may be necessary to allow fuller coverage of the range of possibilities in assemblage composition and decrease the risk of encountering no-analogue situations (Juggins & Birks 2012).

In projects with hierarchical designs including multiple sites and samples per site (e.g. King & Richardson 2002), the contributions of number of sites and number of samples within a site to cost and information content can be compared and optimized. The cost of an individual sample would be lower than in analogous single-sample scenarios, due to reduced per-sample logistical costs. However, the proportional contribution of an individual sample to total information content could also be lower. Likewise, in projects involving several taxonomic groups and analytical techniques (e.g. USEPA 2013), sampling may be optimized considering shared logistical costs among the sampled taxonomic groups. If project goals are specific (e.g. differentiate among sites, model environmental variables), some taxonomic groups may provide more information than others, and the types and number of analyses may be optimized according to information content and cost.

CONCLUSIONS AND RECOMMENDATIONS

An acceptable level of information content in environmental assessment models depends on the tolerance of a decision maker for uncertainty, the magnitude of differences a decision maker needs to detect, the amount of money available for sampling and identification, and the biological and environmental characteristics of the study area. Financial costs can have a major impact on deciding an acceptable level of uncertainty and can be the limiting factor in choosing an analytical strategy. Our results suggest that species-level analyses are almost always the most cost-effective option for diatom-based environmental assessment models and that 50–100 sample lakes provide

sufficient information content for regional models, provided adequate spatial and environmental coverage. Beyond this level of sampling, information gains are small and cost-effectiveness appears low. However, there may be hidden sacrifices in collecting the minimum number of samples necessary to reach near-maximum information content, particularly in cases where predictions for new sites depend on close assemblage analogues (e.g. use of dissimilarity-based models), or where study goals (e.g. 'examine ecosystem health') are more nebulous than characterization of environmental relationships. More samples will always provide more information. However, the time and resources used in obtaining such information must be balanced with the cost of making and implementing an environmental decision (McDonald-Madden *et al.* 2010). By explicitly modelling costs and information content as we have done, decision makers can choose when enough information has been obtained to direct additional resources towards mitigation or direct management.

The size and geographic scope of our data sets, number of variables analysed and relatively consistent patterns suggest that our recommendations are applicable to a variety of diatom-based assemblage vs. environment models commonly used in environmental assessments and palaeoecology. Future research can use a similar approach to determine the most cost-effective balance between sample collection and taxonomic effort in other bioindicators such as aquatic invertebrates and for additional sampling scenarios including multiple samples per site and multiple indicators. In addition to sample collection costs, the number of individual organisms enumerated per sample could also be analysed. In our case, samples were enumerated to accepted standards for environmental assessment data sets. Additional cost savings may have been possible with fewer diatom valves counted in the original samples. In the case of aquatic invertebrates, since individual organisms are often hand-picked from samples, the number of organisms isolated and counted may be an important component of sample cost and information content (King & Richardson 2002). The framework we have provided could easily accommodate such additional cost considerations.

Acknowledgements

Data accessibility

- North American pH and TP data, with the exception of Nova Scotia sites from Ginn, Cumming & Smol (2007), are archived in the USEPA EMAP website (USEPA 2013) and the Diatom Paleolimnology Data Cooperative website (DPDC 2013).
- European pH data are archived on the website for the EDDI project (Battarbee *et al.* 2001).
- Data from Nova Scotia sites and salinity/lake depth data set are provided in the Supporting Information.

References

Battarbee, R.W., Juggins, S., Gasse, F., Anderson, N.J., Bennion, H., Cameron, N.G. *et al.* (2001) European Diatom Database (EDDI): an information system for paleoenvironmental reconstruction. *ECRC Research Report*. Environmental Change Research Centre, University College London. http://craticula.ncl.ac.uk/Eddi/jsp/

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010) Regression analysis of spatial data. *Ecology Letters*, **13**, 246–264.

Bennett, J.R. (2014) Comparison of native and exotic distribution and richness models across scales reveals essential conservation lessons. *Ecography*, **37**, 120–129.

Bennett, J.R., Cumming, B.F., Ginn, B.K. & Smol, J.P. (2010) Broad-scale environmental response and niche conservatism in lacustrine diatom communities. *Global Ecology and Biogeography*, **19**, 724–732.

Berthon, V., Bouchez, A. & Rimet, F. (2011) Using diatom life-forms and ecological guilds to assess organic pollution and trophic level in rivers: a case study of rivers in south-eastern France. *Hydrobiologia*, **673**, 259–271.

Bevilacqua, S., Terlizzi, A., Claudet, J., Fraschetti, S. & Boero, F. (2012) Taxonomic relatedness does not matter for species surrogacy in the assessment of community responses to environmental drivers. *Journal of Applied Ecology*, **49**, 357–366.

Bjørnstad, O.N. & Falck, W. (2001) Nonparametric spatial covariance functions: estimation and testing. *Environmental and Ecological Statistics*, **8**, 53–70.

Carneiro, F.M., Bini, L.M. & Rodrigues, L.C. (2010) Influence of taxonomic and numerical resolution on the analysis of temporal changes in phytoplankton communities. *Ecological Indicators*, **10**, 249–255.

Carter, J.L. & Resh, V.H. (2001) After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *Journal of the North American Benthological Society*, **20**, 658–682.

Core Team, R. (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

DPDC (2013). Diatom Paleolimnology Data Cooperative. http://diatom.ansp.org/dpdc/

Dixit, S. & Smol, J.P. (1994) Diatoms as indicators in the Environmental Monitoring and Assessment Program-Surface Waters (EMAP-SW). *Environmental Monitoring and Assessment*, **31**, 275–307.

Dixit, S., Cumming, B., Birks, H.J.B., Smol, J.P., Kingston, J., Uutala, A., Charles, D. & Camburn, K. (1993) Diatom assemblages from Adirondack lakes (New York, USA) and the development of inference models for retrospective environmental assessment. *Journal of Paleolimnology*, **8**, 27–47.

Donohue, I., Donohue, L.A., Ainín, B.N. & Irvine, K. (2009) Assessment of eutrophication pressure on lakes using littoral invertebrates. *Hydrobiologia*, **633**, 105–122.

Ferraro, S.E., Swartz, R.C., Cole, F.A. & Deben, W.A. (1994) Optimum macrobenthic sampling protocol for detecting pollution impacts in the Southern California Bight. *Environmental Monitoring and Assessment*, **29**, 127–153.

Fortin, M.-J. & Dale, M. (2005) *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press, Cambridge.

Fortin, M.-J., Drapeau, P. & Legendre, P. (1989) Spatial autocorrelation and sampling design in plant ecology. *Vegetatio*, **83**, 209–222.

Gilbert, B. & Bennett, J.R. (2010) Partitioning variation in ecological communities: do the numbers add up? *Journal of Applied Ecology*, **47**, 1071–1082.

Ginn, B.K., Cumming, B.F. & Smol, J.P. (2007) Diatom-based environmental inferences and model comparisons from 494 Northeastern North American lakes. *Journal of Phycology*, **43**, 647–661.

Greffard, M.-H., Saulnier-Talbot, É. & Gregory-Eaves, I. (2011) A comparative analysis of fine versus coarse taxonomic resolution in benthic chironomid community analyses. *Ecological Indicators*, **11**, 1541–1551.

Guiry, M.D. & Guiry, G.M. (2013) *AlgaeBase*. National University of Ireland. http://www.algaebase.org/.

Hall, R.I. & Smol, J.P. (2010) Diatoms as indicators of lake eutrophication. *The Diatoms: Applications for the Environmental and Earth Sciences* (eds J.P. Smol & E.F. Stoermer), pp. 122–151. Cambridge University Press, Cambridge.

Heino, J. & Soninen, J. (2007) Are higher taxa adequate surrogates for species-level assemblage patterns and species richness in stream organisms? *Biological Conservation*, **137**, 78–89.

Heiri, O. & Lotter, A.F. (2010) How does taxonomic resolution affect chironomid-based temperature reconstruction? *Journal of Paleolimnology*, **44**, 589–601.

Hughes, R.M. & Peck, D.V. (2008) Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. *Journal of the North American Benthological Society*, **27**, 837–859.

Jones, F.C. (2008) Taxonomic sufficiency: the influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environmental Reviews*, **16**, 45–69.

Juggins, S. (2012) rioja: Analysis of Quaternary Science Data, R package version (0.8-5). http://cran.r-project.org/package=rioja.

Juggins, S. (2013) Quantitative reconstructions in palaeolimnology: new paradigm or sick science? *Quaternary Science Reviews*, **64**, 20–32.

Juggins, S. & Birks, H.J.B. (2012) Quantitative environmental reconstructions from biological data: data handling and numerical techniques. *Tracking Environmental Change Using Lake Sediments* (eds H.J.B. Birks, A.E.F. Lotter, S. Juggins & J.P. Smol), pp. 431–494. Springer, Dordrecht.

Keitt, T.H., Bjørnstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, **25**, 616–625.

King, R.S. & Richardson, C.J. (2002) Evaluating subsampling approaches and macroinvertebrate taxonomic resolution for wetland bioassessment. *Journal of the North American Benthological Society*, **21**, 150–171.

Marshall, J.C., Steward, A.L. & Harch, B.D. (2006) Taxonomic resolution and quantification of freshwater macroinvertebrate samples from an Australian dryland river: the benefits and costs of using species abundance data. *Hydrobiologia*, **572**, 171–194.

McDonald-Madden, E., Baxter, P.W.J., Fuller, R.A., Martin, T.G., Game, E.T., Montambault, J. & Possingham, H.P. (2010) Monitoring does not always count. *Trends in Ecology and Evolution*, **25**, 547–550.

Morin, S., Duong, T.T., Dabrin, A., Coynel, A., Herlory, O., Baudrimont, M. *et al.* (2008) Long-term survey of heavy-metal pollution, biofilm contamination and diatom community structure in the Riou Mort watershed, South-West France. *Environmental Pollution*, **151**, 532–542.

Payne, R.J. & Mitchell, E.A.D. (2009) How many is enough? Determining optimal count totals for ecological and palaeoecological studies of testate amoebae. *Journal of Paleolimnology*, **42**, 483–495.

Reavie, E.D. & Juggins, S. (2011) Exploration of sample size and diatom-based indicator performance in three North American phosphorus training sets. *Aquatic Ecology*, **45**, 529–538.

Round, F.E., Crawford, R.M. & Mann, D.G. (1990) *Diatoms: Biology and Morphology of the Genera*. Cambridge University Press, Cambridge, UK.

Seabloom, E.W., Bjørnstad, O.N., Bolker, B.M. & Reichman, O.J. (2005) Spatial signature of environmental heterogeneity, dispersal, and competition in successional grasslands. *Ecological Monographs*, **75**, 199–214.

Simpson, G.L. (2012) Analogue methods in palaeolimnology. *Tracking Environmental Change Using Lake Sediments* (eds H.J.B. Birks, A.F. Lotter, S. Juggins & J.P. Smol), pp. 495–522. Springer, Dordrecht.

Siqueira, T., Bini, L.M., Roque, F.O. & Cottenie, K. (2012) A metacommunity framework for enhancing the effectiveness of biological monitoring strategies. *PLoS One*, **7**, e43626.

Smol, J.P. (2008) *Pollution of Lakes and Rivers: A Paleoenvironmental Perspective*, 2nd edn. Blackwell Publishing, Oxford.

Smol, J.P. & Stoermer, E.F. (2010) Applications and uses of diatoms: prologue. *The Diatoms: Applications for the Environmental and Earth Sciences*, 2nd edn (eds J.P. Smol & E.F. Stoermer), pp. 3–8. Cambridge University Press, Cambridge.

Spaulding, S.A., Lubinski, D.J. & Potapova, M. (2010) *Diatoms of the United States*. University of Colorado, Boulder, U.S. http://westerndiatoms.colorado.edu/.

Telford, R.J. & Birks, H.J.B. (2009) Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, **28**, 1309–1316.

Telford, R.J. & Birks, H.J.B. (2011) Effect of uneven sampling along an environmental gradient on transfer-function performance. *Journal of Paleolimnology*, **46**, 99–106.

Theriot, E.C., Ashworth, M., Ruck, E., Nakov, T. & Jansen, R.K. (2010) A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution*, **143**, 278–296.

Thompson, B.W., Riddle, M.J. & Stark, J.S. (2003) Cost-efficient methods for marine pollution monitoring at Casey Station, East Antarctica: the choice of sieve mesh-size and taxonomic resolution. *Marine Pollution Bulletin*, **46**, 232–243.

Turak, E., Flack, L.K., Norris, R.H., Simpson, J. & Waddell, N. (1999) Assessment of river condition at a large spatial scale using predictive models. *Freshwater Biology*, **41**, 283–298.

U.S. Environmental Protection Agency. (2013) Environmental Monitoring and Assessment Program. http://www.epa.gov/emap2/.

Vaughan, I.P. & Ormerod, S.J. (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**, 720–730.

Weilhoefer, C.L. & Pan, Y. (2006) Diatom-based bioassessment in wetlands: how many samples do we need to characterize the diatom assemblage in a wetland adequately? *Wetlands*, **26**, 793–802.

Wetzel, R.G. (2001) *Limnology: Lake and River Ecosystems*, 3rd edn. Academic Press, San Diego.

Wilson, S.E., Cumming, B.F. & Smol, J.P. (1996) Assessing the reliability of salinity inference models from diatom assemblages: an examination of a 219-lake data set from western North America. *Canadian Journal of Fisheries and Aquatic Sciences*, **53**, 1580–1594.

WISER Consotrium. (2014) Water bodies in Europe: Integrative Systems to assess Ecological status and Recovery. http://www.wiser.eu/.

Wunsam, S., Cattaneo, A. & Bourassa, N. (2002) Comparing diatom species, genera and size in biomonitoring: a case study from streams in the Laurentians (Québec, Canada). *Freshwater Biology*, **47**, 325–340.

Supporting Information

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

Table S1. Explained variation in RDAs.

Table S2. Models examined for spatial autocorrelation across variables.

Table S3. Data from Nova Scotia pH lakes and salinity/lake depth data set.

Fig. S1. Linear regressions for predicted vs. actual pH from the European pH data set.

Fig. S2. Linear regressions for predicted vs. actual log total phosphorus.

Fig. S3. Linear regressions for predicted vs. actual log salinity.

Fig. S4. Linear regressions for predicted vs. actual log lake depth.

Fig. S5. Typical spline correlogram showing no significant spatial autocorrelation in residuals.

Fig. S6. Typical spline correlogram showing significant spatial autocorrelation in residuals.