

# Exploration of sample size and diatom-based indicator performance in three North American phosphorus training sets

Euan D. Reavie · Steve Juggins

Received: 22 March 2011 / Accepted: 12 September 2011 / Published online: 23 September 2011  
© Springer Science+Business Media B.V. 2011

**Abstract** Three large training sets were investigated to determine optimal sample sizes for diatom-based inference models. The sample sets represented (1) assemblages from Great Lakes coastlines, (2) phytoplankton from the pelagic Great Lakes and (3) surface sediment assemblages from Minnesota lakes. Diatom-based weighted average models to infer nutrient concentrations were developed for each training set. Training set sample sizes ranging from 10 to the maximum number of samples were created through random sample selection, and performance of each model was evaluated. For each model iteration, diatom-inferred (DI) nutrient data were related to stressor data (e.g., adjacent agricultural or urban development) to characterize the ability of each model to track human activities. The relationships between model performance parameters (DI-stressor correlations and model  $r^2$ , error and bias) and sample size were used to determine the minimum sample size

needed to optimize models for each region. Depending on the training set, at least 40–70 samples were needed to capture the variation in diatom assemblages and environmental conditions to such a degree that non-analog situations should be rare and so should provide an unambiguous result if the model was applied to any sample assemblage from the region. It is recommended that one exercises caution when dealing with smaller training sets unless there is certainty that the selected samples reflect the regional variability in diatom assemblages and environmental conditions.

**Keywords** Diatoms · Stressors · Training sets · Inference models · Sample size · Models

## Introduction

Application of diatom-based calibration models has become commonplace in monitoring and paleoecological studies. Such models are typically built using a training set of sample locations (e.g., lakes, wetlands), and for each location, the environmental variables of interest (e.g., phosphorus, pH, salinity, agricultural stress), and the corresponding local diatom assemblage (e.g., in surface sediments, rock scrapes, epiphyton), are sampled. The environmental and species matrices of data derived from the training set are then integrated to determine the environmental characteristics of the diatom taxa. When weighted averaging (WA) approaches are applied, the resulting model

---

Handling Editor: Piet Spaak.

---

E. D. Reavie (✉)  
Natural Resources Research Institute, University  
of Minnesota Duluth, 1900 East Camp Street,  
Ely, MN 55731, USA  
e-mail: ereavie@nrri.umn.edu

S. Juggins  
School of Geography, Politics and Sociology, Newcastle  
University, Newcastle upon Tyne NE1 7RU, UK  
e-mail: Stephen.Juggins@ncl.ac.uk

essentially comprises the diatom species and their optima and tolerances for environmental parameters, supplemented by any rules and recommendations for model application such as data transformations, downweighting by species tolerance and removal of outliers. The diatom-based model can then be used to quantitatively infer environmental conditions from the species composition in a new sample. This approach was originally used in paleolimnology to “reconstruct” past environmental conditions from downcore species assemblages (e.g., Birks et al. 1990; Battarbee et al. 2001), but some have extended these efforts to contemporary monitoring (e.g., Dixit and Smol 1994; Reavie et al. 2006).

Diatom-based training sets are an essential prerequisite to model development and training set design, and size has been investigated by several authors. Using the correlation between observed and diatom-inferred (DI) environmental values as a guide to model performance (Birks et al. 1990), sample sets as small as 46 (Hall and Smol 1992) and 33 (Bennion 1994; Tibby 2004) have been demonstrated to provide “good” predictive ability. It is generally expected that training sets with more samples will have higher numbers of taxa and better definition of environmental conditions in the region of interest, and as a result, a model based on a larger dataset should provide more reliable inferences. However, the substantial effort involved in developing a training set necessitates that an optimal sample size be estimated using model performance criteria. Three considerations are used to determine the adequacy of a training set as the number of samples therein increases.

1. *How many samples are needed to characterize the species present?* The number of samples necessary to adequately characterize the organism assemblages from particular regions has been explored by several researchers. Species representation is critical to diatom-based models because one is less likely to encounter non-analog conditions during model application, especially when inferring condition from species assemblages in sediment cores. Not surprisingly, as the sample region increases in size and complexity, more samples are needed. For instance, Weilhöfer and Pan (2006) found that composition of only five surface sediment diatom samples was sufficient to characterize the diatom species richness in a wetland in the Columbia River floodplain. Conversely, Bowen and Freeman (1998) in a study of medium-sized rivers in Alabama, determined that 70 electrofisher samples were needed to adequately describe the fish species richness in the system. Rarefaction analysis has been used to determine the number of samples needed to obtain most of the diatom species from a region (Birks and Line 1992). Similarly, sufficient samples are needed to characterize the environmental conditions for a region, although measured environmental conditions across regional training sets have generally been shown to be significantly less variable than the algal assemblages. For example, species gradients were much larger than environmental gradients in the coastal Great Lakes (Reavie et al. 2006).
2. *How many samples are needed to provide accurate estimates of species coefficients for environmental variables?* Each location in a training set has a distinct diatom assemblage comprising several species of varying abundance. In addition to describing the species present, it is important that representative abundance data are collected so that species responses across environmental gradients are well defined. These abundance data are used to calculate the point of maximum abundance (optimum) and expected range (tolerance) along the environmental gradient of interest. Well-defined species coefficients are particularly important when weighted averaging approaches are applied because the optimum and tolerance for each taxon in an assemblage are selectively weighted based on the taxon’s abundance in a sample. One way to determine whether assemblage structures have been captured is to examine model performance criteria. For example, Wilson et al. (1996) determined that approximately 100 samples resulted in a nearly asymptotic minimum model error from a training set of 219 British Columbia Lakes used to develop a salinity model.
3. *How many samples are needed to adequately track human impacts?* Maximizing distribution measures such as species richness, diversity or evenness, and optimizing model performance criteria such as  $r^2$  and RMSE do not fully define a model’s power. Despite apparent model performance, such as a high observed-versus-inferred  $r^2$

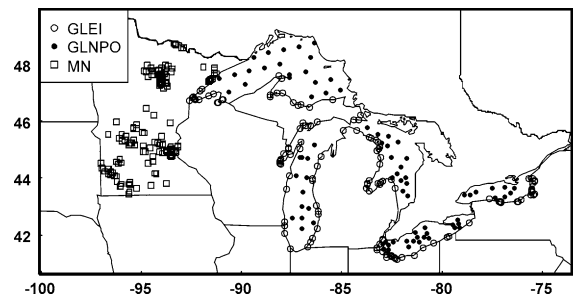
or a low prediction error, models intended as tracers of human impacts need to be sensitive to stressors. A small sample set may adequately describe the species that occur in that region, and it may even reveal an apparently robust diatom-based model, but an indicator model with little correlation to the stressors of interest has little value for aquatic management. Strong relationships between diatom-inferred condition and corresponding anthropogenic stressors have been described (e.g., Reavie et al. 2006), and diatom-inferred Great Lakes coastal water quality is better able to reflect stressors than snapshot water quality measurements, but to date, no study has investigated the effect of sample size on the strength of the DI-stressor relationship.

Sampling strategies to capture the number of species in a region have been given good treatment in other volumes (e.g., Hayek and Buzas 1997), so the current study investigates issues 2 and 3 above in the context of three large diatom-based training sets collected from North American lakes. To provide accurate estimates, training sets need to capture both the species and their typical abundances that occur across an environmental gradient. Small training sets are likely to suffer from poor model precision (and analog issues during model application) due to an inability to adequately characterize variability in the ambient assemblages. The current work investigates the effect of training set sample size on diatom-based model performance and the ability to track human stressors in adjacent watersheds.

## Methods

### Sample collection and preparation

Three diatom training sets (Fig. 1) were evaluated to determine how sample size influenced diatom-based model performance (Table 1). Each training set comprised two matrices: a diatom dataset of samples and species percent abundance of the total diatom count for a given sample and an environmental dataset containing one or more measured water quality nutrient variables. One of the three datasets also had a corresponding matrix of stressor variables, i.e., quantified data reflecting anthropogenic activities



**Fig. 1** North American map with sample locations for the three training sets

(e.g., agricultural and urban development) in watersheds adjacent to diatom sample locations. A surrogate stressor dataset was created for Great Lakes pelagic training set as described below. Diatoms were collected from substrates as summarized in Table 1, and additional details are provided in the respective articles and standard operating procedures. For all samples, diatoms were prepared by digestion with a strong acid and/or base, washing with deionized water and streaking for observation on microscope slides. Preparation details are provided in the respective publications (Ramstack et al. 2003; Reavie et al. 2006; USEPA 2010). Counts were performed using oil immersion at 1000× or greater magnification. In cases where multiple taxonomists were involved in counts for a particular training set, collaboration was maintained by phone, email and taxonomic workshops.

*Great Lakes Environmental Indicators (GLEI)* The US Environmental Protection Agency (US EPA) established the Estuarine and Great Lakes (EaGLe) research program to develop new approaches to assess environmental condition. One of the EaGLe projects, the Great Lakes Environmental Indicators (GLEI) project, was specifically designed to develop indicators for the Laurentian Great Lakes. Over 200 diatom samples were collected from embayments, high-energy shorelines and coastal wetlands from 2001 through 2004. Several robust diatom-based models were developed from these data (Reavie et al. 2006, 2008; Kireta et al. 2007; Reavie 2007; Sgro et al. 2007) as confirmed by relating diatom-inferred conditions to anthropogenic stressors in adjacent watersheds. For this investigation, the environmental variable of choice for diatom calibration is total phosphorus (TP), which was shown to be related to patterns in the coastal diatom communities (Reavie et al. 2006).

**Table 1** General information for the three diatom datasets

Dataset name	Diatom sample type	Sample number	Number of non-rare taxa	Species/nutrient transformation	Nutrient variable and range	Stressor variable	Estimated sample number needed for best model performance
Great Lakes Environmental Indicators (GLEI)	Surface sediments, substrate scrapes	206	380	Log/log	Total phosphorus (1–521 $\mu\text{g L}^{-1}$ )	GIS-based agricultural development	70, but slight improvements might be achieved up to 115
Great Lakes National Program Office phytoplankton (GLNPO)	Whole-water collection	302	192	None/log	Total phosphorus (1–99 $\mu\text{g L}^{-1}$ )	Coarse, lake-based stress estimate	45, but slight improvements might be achieved up to 70
Minnesota lakes (MN)	Surface sediments	145	170	Log/log	Total phosphorus (5–664 $\mu\text{g L}^{-1}$ )	Not used	40, but slight improvements might be achieved up to 50

Specific model parameters refer to transformations applied to data as recommended by their developers. The sample numbers needed to determine “best” sample size were determined using statistical comparisons between performance at various sample sizes versus that at maximum sample size

Corresponding stressor data adjacent to each sample location were created from landscape characteristics summarized by Danz et al. (2007). An integrated variable quantifying agricultural activities (agricultural principal component 1 that captured 73% of the variation in 21 agriculture variables) was determined to be the one best related to patterns in the coastal diatom assemblages (Reavie et al. 2006), so that variable was selected as the stressor.

*Great Lakes National Program Office: phytoplankton (GLNPO)* Twice yearly the US EPA conducts surveillance monitoring of the offshore waters of the Great Lakes to fulfill provisions of the Great Lakes Water Quality Agreement. To track environmental conditions and trends, these surveys include phytoplankton collections (Barbiero and Tuchman 2002), and detailed diatom assessments are performed on these samples. The diatom samples in this investigation comprise pelagic collections from spring and summer cruises in 2007 and 2008. Due to ancillary analyses showing strong relationships between pelagic diatom assemblages and nutrient concentrations (unpublished data), TP was selected as the model variable. To date, stressor data corresponding to each sample location have not been developed, so a highly simplified, rank-based surrogate dataset that classified each lake based on its known developmental stress-to-volume ratio was used. Lakes were ordered from 1

(least impacted) to 5 (most impacted): Superior, Huron, Michigan, Ontario and Erie, respectively (Environment Canada and USEPA 2009). The relevance of this simplified classification is debatable, but for this study, it provides a semi-independent stressor dataset to use for comparisons of measured and DI TP data.

*Minnesota lake set (MN)* The MN training set of lakes has been augmented and developed for more than a decade (Heiskary and Swain 2002; Ramstack et al. 2003; Edlund and Kingston 2004; Reavie et al. 2005), and it has been applied several times in state-based paleolimnology initiatives to determine historical nutrient shifts and eutrophication trends (e.g., Kingston et al. 2004; Reavie and Baratono 2007). In keeping with previous intentions for this model (Heiskary and Swain 2002), the environmental variable of interest is TP. Detailed stressor data have not been collected for all of the 145 lakes in the MN training set, so no stressor variable was tested.

#### Model development

Weighted averaging (WA) calibration and regression were applied in the development and testing of all models in this investigation. WA is a standard application that employs modern diatom–environmental relationships to infer environmental conditions

from a given diatom assemblage, such as that from a particular period in a sediment core (ter Braak and van Dam 1989; Birks et al. 1990). The WA method assumes symmetric responses of diatom species along environmental gradients and has been repeatedly demonstrated to provide robust inferences of condition (Battarbee et al. 2001). Although other popular modeling approaches exist and could be applied to these data, for the sake of comparison among datasets, this investigation maintains a consistent approach, as described below. Diatom-based models were developed using the R programming language (version 2.10.0, R Development Core Team 2010) with the package “rioja” (version 0.5-6, Juggins 2009).

Rare taxa were removed from each dataset by eliminating those that never achieved at least 1% relative abundance and occurred in fewer than five samples. Transformation of species data was applied as necessary based on dataset-specific recommendations (Table 1). Diatom-based WA models were iteratively created and tested using repeated calls to rioja modeling functions, as follows using the GLEI dataset as an example. Starting with the complete dataset of 206 samples, 33% of samples (68) were selected to serve as an independent test set using a stratified pseudorandom sampling procedure. Stratified sampling was applied because it mimics the approach that would be taken by the developer of a training set; i.e., samples were selected to ensure that they reflected the environmental gradient of interest. In this case, the full set of samples was divided into 10 bins ranked by lowest to highest TP concentrations, and samples were randomly selected from each bin (6 or 7 samples per bin). Samples not selected (138) became the model (calibration) set. Progressive iterations created model datasets from 10 to 138 samples in steps of 5 samples, also using stratified subsampling to mimic the sampling distribution of the full calibration dataset. This procedure was repeated 20 times for each model size to provide an estimate of variation in performance. For the GLEI dataset, a total of 520 models were generated and tested on independent sample sets. The R script for these analyses is available from the primary author on request.

Weighted averaging models were developed using both simple WA, which uses the species environmental optima for predictions, and tolerance downweighting which down-weights the contribution of each taxon according to its tolerance. After preliminary

evaluations of downweighting approaches, the most suitable application of tolerance downweighting was based on the effective number of occurrences of taxa (Hill’s  $N_2$ ) and no replacement of very small tolerances. However, WA with tolerance downweighting yielded consistently poorer model performance, and this method is not further discussed. Inverse deshrinking was applied for all models. TP predictions for each sample in the independent dataset were then inferred using the calibrated model. After each model iteration, diatom-inferred TP data were regressed against the stressor data for the independent sample set to quantify the strength of the relationship between stressors and DI data.

Five parameters, described below, were used to evaluate model performance relative to the number of samples in the calibration dataset. These parameters were generated based on analysis of relationships between observed and inferred data for the independent test sets (e.g., the 68 samples in the GLEI example above). Relationships between sample size and performance statistic were modeled using generalized additive models (GAM) with spline smoothing, with the degree of smoothing chosen by generalized cross validation (Wood 2006). We also compared the performance statistic at each model size to the maximum size using a series of pairwise  $t$  tests with  $P$  values adjusted for multiple comparisons using Bonferroni correction (Quinn and Keough 2002).

*Squared correlation coefficient ( $r^2$ ) of the observed-inferred relationship* This parameter is generated through the comparison of observed environmental measurements (e.g., measured phosphorus) and diatom-inferred values (e.g., DI phosphorus) for that variable. It was expected that as sample sizes increased,  $r^2$  would also increase because a larger sample size should increase the likelihood of capturing the representative diatom assemblages from a region. Assuming samples are being collected from locations that are evenly distributed along the environmental (e.g., nutrient) gradient, and a threshold for  $r^2$  should be reached where additional samples provide little or no additional information (environmental measurements and diatom assemblages) to improve model performance.

*Root mean squared error of prediction (RMSEP) for the observed-inferred relationship* This partner to  $r^2$  reflects the error associated with model predictions. In

other words, RMSEP provides a measure of the spread of predicted values around the idealized 1:1 line of the observed-inferred regression. As sample sizes increased, it was expected that RMSEP would get smaller as more regional measures of diatom populations and environmental conditions would refine the resulting model coefficients.

*Mean bias in the residuals from the observed-inferred relationship* Diatom-inferred residuals around the 1:1 line are usually shown to have some bias. For instance, Reavie and Smol (2001) recognized that diatom-inferred values from an Ontario TP model tended to underestimate measured nutrient concentrations at the upper, eutrophic end of the nutrient gradient. Often such an underestimation of condition is balanced by overprediction at the lower (e.g., oligotrophic) end of the gradient (e.g., Bradshaw and Anderson 2001), resulting in an overall mean bias near zero. It was expected that mean bias would tend to be near zero under most model conditions, but that the direction of any bias would be less predictable with smaller sample sizes.

*Maximum bias in the residuals from the observed-inferred relationship* Maximum bias measures the tendency of a model to over- or under-predict at particular parts of the environmental gradient. It is computed by dividing the gradient into 10 equal sections and calculating the mean of the residuals within each of these sections. The maximum bias is taken as the maximum of the mean biases of the 10 sections. This parameter provides a useful heuristic estimate of the near-worst case precision of any given diatom-inferred value. As for RMSEP, maximum bias was expected to decrease as sample sizes increased. Maximum bias values can be positive or negative, so these values were standardized as absolute values.

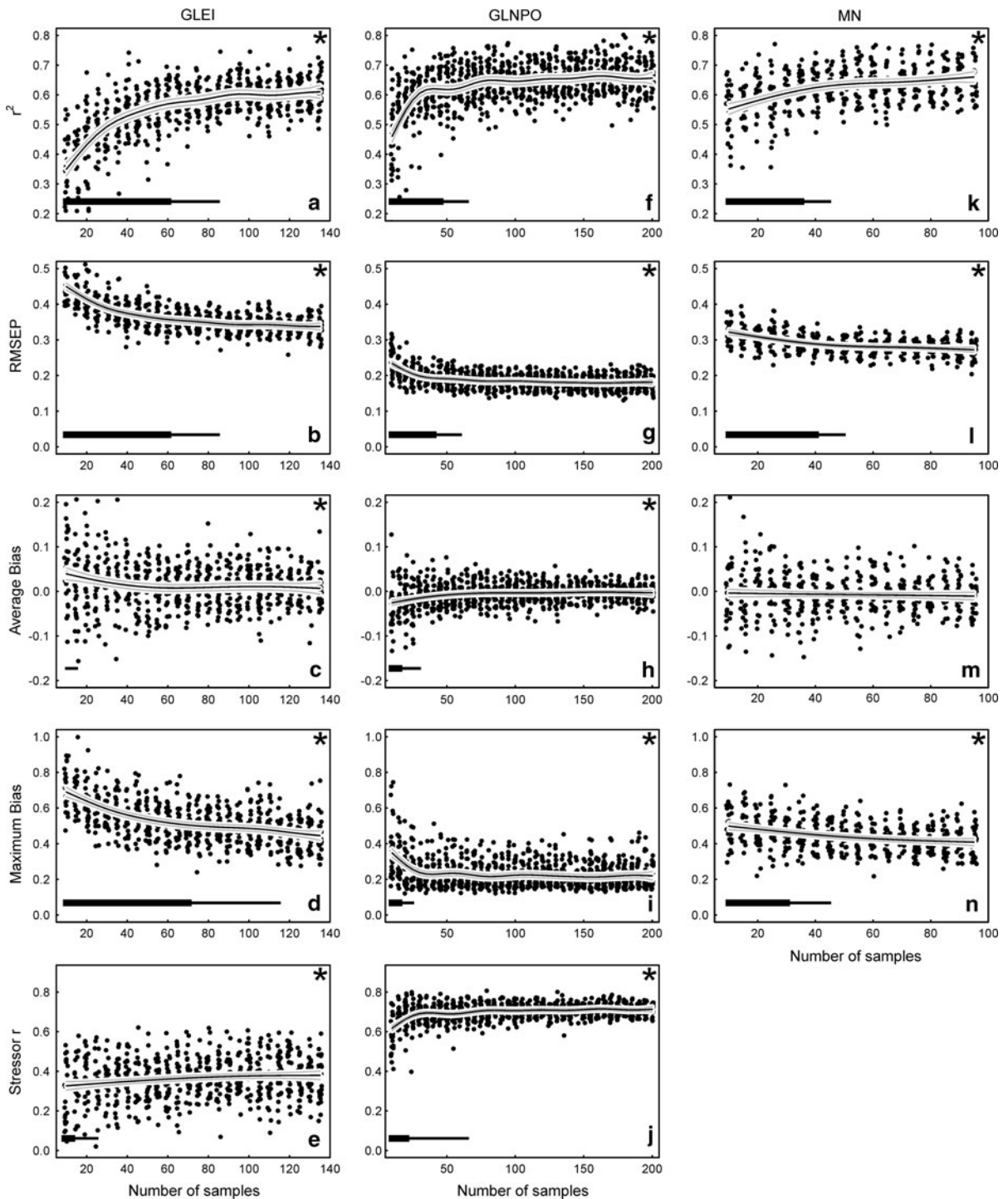
*Pearson product–moment correlation ( $r$ ) between diatom-inferred values and stressor measures* Stressor correlations with diatom-inferred water quality provide quantitative evidence whether a diatom-based model has the ability to track anthropogenic activities. It was expected that as sample size increased the correlation would also increase due to larger models providing more refined species coefficients, which in turn generates more accurate diatom-inferred water quality data which are related to anthropogenic influences.

**Fig. 2** Relationships between sample size ( $x$ -axis) and diatom-based indicator performance for the three datasets used in this study: Great Lakes Environmental Indicators (GLEI) shoreline sample set, Great Lakes National Program Office (GLNPO) phytoplankton sample set and Minnesota (MN) lake surface sediment sample set. Performance parameters are **a, f, k** squared correlation coefficient for the observed diatom-inferred (DI) relationship for total phosphorus, **b, g, l** root mean square error of prediction, **c, h, m** mean model bias, **d, i, n** absolute value of the maximum model bias and **e, j** correlation coefficient of the DI-stressor relationship. Individual data points are jittered on the  $x$ -axis to avoid overlap. Curves show GAM fits and 95% confidence intervals. *Asterisks in top right* indicate a significant GAM fit as tested against a null model of no change in performance statistic with sample size. *Solid lines at the bottom* of each figure indicate the range of sample sizes with significant differences in performance statistic compared to the respective value for maximum sample size (pairwise  $t$  tests; *thin line*,  $P \leq 0.05$ ; *thick line*,  $P \leq 0.05$  with Bonferroni correction for multiple comparisons)

## Results

A cursory examination of model performance results (Fig. 2) reveals that, as sample size increases, performance data stabilize in two ways. First, model responses to higher sample sizes were mostly asymptotic to a minimum or maximum performance value, as marked by the GAM fits that leveled off with higher sample numbers. Second, although not always present, performance data became more predictable with larger training sets as indicated by tapering, wedge-shaped trends as the spread of performance values narrowed with higher sample sizes. The stabilities of these two conditions were used to estimate optimal sample sizes for each regional dataset. Principally, the apparent asymptotes of the fitted GAM curves provided a means to estimate appropriate sample sizes (Fig. 2; Table 1). Judgments on critical sample sizes for a given training set were mainly based on the point where larger sample sizes no longer exhibited a significant change in performance parameters relative to the largest sample size. Because Bonferroni corrections were used for this significance testing, additional changes in performance data with higher sample sizes, i.e., based on GAM fits that continued to increase or decrease, were recognized where apparent.

**GLEI** For the GLEI training set, model performance was optimized at approximately 60 samples, above which RMSEP ranged from 0.28 to 0.38 and  $r^2$  that ranged from 0.5 to 0.7 (Fig. 2a, b). Based on trends



that were validated using non-Bonferroni-corrected comparisons, model  $r^2$  continued to increase slightly until 85 samples were reached. Average bias showed a

very slight decline over the range of sample sizes, but no significant affect was noted based on Bonferroni-corrected testing (Fig. 2c). Maximum bias declined

significantly until 70 samples were reached (Fig. 2d). The DI-stressor GAM fit increased most rapidly as sample sizes increased from 10 to 15 (Bonferroni-corrected  $P = 0.05$ ), and a plateau was reached at approximately 25 samples (Fig. 2e).

**GLNPO** Model  $r^2$  and RMSEP appear to reach a plateau ( $r^2 = 0.64$ ) above approximately 45 samples (Fig. 2f, g). Average bias had no significant change above 25 samples (Fig. 2h). Maximum bias was optimized at  $\sim 15$  samples (Fig. 2i). The DI-stressor GAM fit reached a plateau of  $r = 0.71$  at approximately 20 samples (Fig. 2j). An interesting pattern was observed in the GLNPO performance results. For maximum bias, certain training sets yielded high values while the majority clustered within the range of 0.15–0.25. It is likely that this artifact was caused by the distinct diatom assemblages that occur in each great lake. We surmise that the independent test sample sets sometimes favored a particular lake. If, for instance, that lake was well represented in the model dataset, TP would be inferred accurately. In contrast, with a more common test set containing a broader mix of lake samples, apparent model performance would be reduced through an average of good and poor DI data.

**MN** The MN lake set was able to provide optimal  $r^2$  and RMSEP results at approximately 40 samples (Fig. 2k, l), although values improved slightly until  $\sim 50$  samples was reached. Average bias was minimal and showed no significant response to increasing sample size, although the spread of values was minimized at approximately 30 samples (Fig. 2m). Maximum bias declined slightly but significantly to a minimum at approximately 30 samples (Fig. 2n).

## Discussion

Sample sizes needed to maximize model performance and indicator-stressor relationships were similar to or smaller than most published training sets. In all cases, at least 40 samples were needed to provide the best model performance, presumably because that sample size captured most of the physical, chemical and biological variability in the training set regions. However, based on more relaxed (non-Bonferroni) statistical testing, more samples are recommended to maximize (or minimize) performance indicators.

Model  $r^2$ , RMSEP and maximum bias appear to be the most conservative determinants of optimal sample size as these parameters generally stabilized at higher sizes. Average bias tended to change little with sample size, and sample numbers needed to maximize correlation with stressors were smaller than that needed to optimize other performance indicators.

Although the MN lake set covers five ecological regions in Minnesota, it is the most geographically constrained set of sample locations compared to the other regional datasets. A narrower environmental gradient in this lake set may be why relatively few samples were needed to provide a robust model. Like the GLEI dataset, the gradient of MN nutrient values was relatively high, and this may have contributed to better model performance with fewer samples, although the gradient length appears to be a weak determinant of model power. Furthermore, although only three training sets are evaluated here, the number of species represented in a model does not appear to determine the number of samples needed to optimize performance.

This study applied nutrient models because it is well known that diatom assemblages respond to nutrient conditions in temperate regions. Further, there is significant management interest in nutrient loads and indicators that can track nutrient-related impacts such as cultural eutrophication. Applications of other variables may change the results presented here. For example, in further testing of the GLEI training set (data not presented), fewer samples than that needed for TP were able to optimize model  $r^2$ , RMSEP and bias when pH was chosen as the environmental variable. This was likely due to smaller temporal variability in pH, so single measurements of pH at each site adequately characterized the prevailing condition and resulted in accurate diatom-pH coefficients.

Based on the current investigation, a sample dataset smaller than the recommended size (Table 1) might provide results with less certainty. Using the GLEI set as an example (Fig. 2a), a 20-sample subtraining set could provide a model  $r^2$  ranging from 0.3 (weaker) to 0.6 (stronger). Similarly, the average bias (Fig. 2c) could range from  $-0.10$  to  $0.15$ , whereas a 60-sample training set provides a narrower range of  $-0.08$  to  $0.10$ . In other words, based on the 20 samples chosen (even if they were selected evenly along the environmental gradient), the apparent GLEI model



performance is questionable in the context of the sample region. Previously published training sets with smaller sample sizes (e.g., fewer than 40 sites) may furnish “good” model performance, but it is recommended in future developments that the physical, chemical and biological variability in the selected sample region is verified to be adequately characterized. Preliminary investigations of a sample region are needed to confirm that the selected training set adequately captures the regional variation that is relevant to diatom indicators, water quality and stressor gradients. Depending on the range and variability of environmental conditions in a regional training set, minimum sample sizes needed may be lower or higher than our recommended sizes for Minnesota and the Great Lakes.

One must consider more than a recommended minimum number of sites when developing an indicator model. Even with a probabilistic sampling strategy, environmental gradients in aquatic regional training sets are difficult to accurately characterize based on the sample set selected for diatom analyses. Unfortunately in most cases, little additional environmental data are available to confirm whether the gradient for the region has been adequately characterized. This arouses concerns for smaller diatom-based sample sets. For instance, Reavie and Smol (2001) employed 64 lakes from a region that crosses several geological and environmental boundaries. Despite robust model performance in subsequent applications (e.g., Forrest et al. 2002), it is not surprising that non-analog cases have occurred when the diatom-based model was applied to lakes from the periphery of the training set region (e.g., Ekdahl et al. 2007). Even though Reavie and Smol (2001) recognized that strong relationships existed between phosphorus concentrations and diatom assemblages, the confounding influences of several other variables such as climate and geology need to be carefully considered when applying a model to sites that are uncertainly within the dominion of the calibration set’s environmental range.

As for our iterative subsampling of training sets, it is critical to employ stratified sampling of the environmental gradient in a region. Model training sets containing a skewed representation of environmental conditions are unlikely to perform well at sites (e.g., modern samples or sediment cores) that are not evenly represented in the model, for instance, attempting to

infer condition at a eutrophic site using a model calibrated using mainly oligotrophic sites.

Diatom communities are structured by local and large-scale factors. Some of these factors are measurable, such as nutrients and disturbance, but other factors such as biotic interactions and historical dispersal are less easily quantified. As a substitute for better environmental characterization, it is recommended to perform analyses such as those described herein to determine how well a given training set simulates the ideal diatom-based indicator (i.e., the set containing an infinite number of samples) for the region.

**Acknowledgments** The Minnesota lake dataset has been progressively developed by Steve Heiskary and Mark Tomasek (Minnesota Pollution Control Agency), Dan Engstrom, Mark Edlund, Shawn Schottler and Joy Ramstack (St. Croix Watershed Research Station). Amy Kireta, Gerald Sgro, Norman Andresen and Michael Ferguson supported diatom assessments for GLEI samples. Michael Agbeti supported diatom assessments of the GLNPO phytoplankton samples. There are several people to thank for GLEI project management and field support, including Valerie Brady, Jerry Henneck, John Ameel, Gerald Niemi, John (Jack) Kelly, Russell Kreis and Jeffrey Johansen. This research was supported by grants to E. Reavie from the US Environmental Protection Agency under Cooperative Agreements EPA/R–8286750 (GLEI) and GL-00E23101 (GLNPO). This document has not been subjected to the EPA’s required peer and policy review and therefore does not necessarily reflect the view of the Agency, and no official endorsement should be inferred. This is contribution number 530 of the Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota Duluth.

## References

- Barbiero RP, Tuchman ML (2002) Results from GLNPO’s biological open water surveillance program of the Laurentian Great Lakes 1999. Report to US EPA Great Lakes National Program Office, EPA-905-R-02-001, p 32
- Battarbee R, Jones VJ, Flower RJ, Cameron NG, Bennion H, Carvalho L, Juggins S (2001) Diatoms. In: Smol JP, Birks HJB, Last WM (eds) Tracking environmental change using lake sediments—volume 3: terrestrial, algal, and siliceous indicators. Kluwer Academic Publishers, Dordrecht, pp 155–202
- Bennion H (1994) A diatom-phosphorus transfer function for shallow, eutrophic ponds in southeast England. *Hydrobiologia* 275(276):391–410
- Birks HJB, Line JM (1992) The use of rarefaction analysis for estimating palynological richness from Quaternary pollen-analytical data. *Holocene* 2:1–10
- Birks HJB, Line JM, Juggins S, Stevenson AC, ter Braak CJF (1990) Diatoms and pH reconstructions. *Philos Trans R Soc Lond B Biol Sci* 327:263–278

- Bowen ZH, Freeman MC (1998) Sampling effort and estimates of species richness based on prepositioned area electro-fisher samples. *N Am J Fish Manag* 18:144–153
- Bradshaw EG, Anderson NJ (2001) Validation of a diatom-phosphorus calibration set for Sweden. *Freshw Biol* 46:1035–1048
- Danz NP, Niemi GJ, Regal RR, Hollenhorst T, Johnson LB, Hanowski JM, Axler RP, Ciborowski JHH, Hrabik T, Brady VJ, Kelly JR, Brazner JC, Howe RW, Johnston CA, Host GE (2007) Integrated gradients of anthropogenic stress in the US Great Lakes basin. *Environ Manag* 39:631–647
- Dixit SS, Smol JP (1994) Diatoms as indicators in the Environmental Monitoring and Assessment Program-Surface Waters (EMAP-SW). *Environ Monit Assess* 31:275–307
- Edlund MB, Kingston JC (2004) Expanding sediment diatom reconstruction model to eutrophic southern Minnesota lakes. Final report to Minnesota Pollution Control Agency, p 33
- Ekdahl EJ, Teranes JL, Wittkop CA, Stoermer EF, Reavie ED, Smol JP (2007) Diatom assemblage response to Iroquoian and Euro-Canadian eutrophication of Crawford Lake, Ontario. *Can J Paleolimnol* 37:233–246
- Environment Canada, USEPA (2009) State of the Great Lakes 2009. EPA 905-R-09-031
- Forrest F, Reavie ED, Smol JP (2002) Comparing the trophic impacts of canal construction to other catchment disturbances in four lakes within the Rideau Canal system, Ontario, Canada. *J Limnol* 61:183–197
- Hall RI, Smol JP (1992) A weighted-averaging regression and calibration model for inferring total phosphorus concentration from diatoms in British Columbia (Canada) lakes. *Freshw Biol* 27:417–434
- Hayek LC, Buzas MA (1997) Surveying natural populations. Columbia University Press, New York, p 563
- Heiskary SA, Swain EB (2002) Water quality reconstruction from fossil diatoms: applications for trend assessment, model verification, and development of nutrient criteria for lakes in Minnesota, USA. Minnesota Pollution Control Agency, Environmental Outcomes Division, St. Paul, Minnesota, p 103
- Juggins S (2009) rioja: analysis of Quaternary science data, R package version 0.5-6. <http://cran.r-project.org/package=rioja>
- Kingston JC, Engstrom DR, Norton AR, Peterson MR, Griese NA, Stoermer EF, Andresen NA (2004) Paleolimnological inference of nutrient loading in a eutrophic lake in north-central Minnesota (USA) and periodic occurrence of abnormal *Stephanodiscus niagarae*. In: Poulin M (ed) Proceedings of the XVIIth international diatom symposium. Biopress Ltd, Bristol, pp 187–202
- Kireta AR, Reavie ED, Axler RP, Sgro GV, Kingston JC, Brown TN, Danz NP, Hollenhorst T (2007) Coastal geomorphic variability in the Laurentian Great Lakes: implications for a diatom-based monitoring tool. *J Gt Lakes Res* 33:136–153
- Quinn G, Keough M (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge
- R Development Core Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>
- Ramstack JM, Fritz SC, Engstrom DR, Heiskary SA (2003) The application of a diatom-based transfer function to evaluate regional water-quality trends in Minnesota since 1970. *J Paleolimnol* 29:79–94
- Reavie ED (2007) A diatom-based water quality index for Great Lakes coastlines. *J Gt Lakes Res* 33:86–92
- Reavie ED, Baratono NG (2007) Multi-core investigation of a lotic bay of Lake of the Woods (Minnesota, USA) impacted by cultural development. *J Paleolimnol* 38:137–156
- Reavie ED, Smol JP (2001) Diatom-environmental relationships in 64 alkaline southeastern Ontario (Canada) lakes: a diatom-based model for water quality reconstructions. *J Paleolimnol* 25:25–42
- Reavie ED, Kingston JC, Edlund MD, Peterson M (2005) Sediment diatom reconstruction model for Minnesota lakes. Report to Itasca Soil and Water Conservation District
- Reavie ED, Axler RP, Sgro GV, Danz NP, Kingston JC, Kireta AR, Brown TN, Hollenhorst TP, Ferguson MJ (2006) Diatom-based weighted-averaging transfer functions for Great Lakes coastal water quality: relationships to watershed characteristics. *J Gt Lakes Res* 32:321–347
- Reavie ED, Sgro GV, Danz NP, Axler RP, Kireta AR, Kingston JC, Hollenhorst TP (2008) Comparison of simple and multimetric diatom-based indices for Great Lakes coastline disturbance. *J Phycol* 44:787–802
- Sgro GV, Reavie ED, Kingston JC, Kireta AR, Ferguson MJ, Danz NP, Johansen JR (2007) A diatom quality index from a diatom-based total phosphorus inference model. *Environ Bioindic* 2:15–34
- ter Braak CJF, van Dam H (1989) Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* 178:209–223
- Tibby J (2004) Development of a diatom-based model for inferring total phosphorus in southeastern Australian water storages. *J Paleolimnol* 31:23–36
- USEPA (2010) Sampling and analytical procedures for GLNPO's open lake water quality survey of the Great Lakes. United States Environmental Protection Agency, Great Lakes National Program Office. Chicago, Illinois. EPA 905-R-05-001. <http://www.epa.gov/glnpo/monitoring/sop/>. Accessed 7 October 2010
- Weilhoefer CL, Pan Y (2006) Diatom-based bioassessment in wetlands: how many samples do we need to adequately characterize the diatom assemblage in a wetland? *Wetlands* 26:793–802
- Wilson SE, Cumming BF, Smol JP (1996) Assessing the reliability of salinity inference models from diatom assemblages: an examination of a 219-lake data set from western North America. *Can J Fish Aquat Sci* 53:1580–1594
- Wood SN (2006) Generalized additive models. An introduction with R. Chapman & Hall, Boca Raton, p 391