Journal of Paleolimnology **32:** 233–246, 2004.

Diatom-based inference models and reconstructions revisited: methods and transformations

Dörte Köster*, Julien M. J. Racca and Reinhard Pienitz
*Paleolimnology–Paleoecology Laboratory, Centre d'études nordiques, Université Laval, Québec, G1K 7P4, Canada; *Author for correspondence (e-mails: doerte.koster.1@ulaval.ca; reinhard.pienitz@cen.ulaval.ca; julien.racca.1@ulaval.ca)*

Received 18 January 2004; accepted in revised form 7 March 2004

Key words: Artificial neural networks, Diatoms, Gaussian logit regression, Inference models, Tolerance-downweighting, Weighted averaging, Weighted averaging partial least squares

Abstract

Different calibration methods and data manipulations are being employed for quantitative paleoenvironmental reconstructions, but are rarely compared using the same data. Here, we compare several diatom-based models [weighted averaging (WA), weighted averaging with tolerance-downweighting (WAT), weighted averaging partial least squares, artificial neural networks (ANN) and Gaussian logit regression (GLR)] in different situations of data manipulation. We tested whether log-transformation of environmental gradients and square-root transformation of species data improved the predictive abilities and the reconstruction capabilities of the different calibration methods and discussed them in regard to species response models along environmental gradients. Using a calibration data set from New England, we showed that all methods adequately modelled the variables pH, alkalinity and total phosphorus (TP), as indicated by similar root mean square errors of prediction. However, WAT had lower performance statistics than simple WA and showed some unusual values in reconstruction, but setting a minimum tolerance for the modern species, such as available in the new computer program C^2 version 1.4, resolved these problems. Validation with the instrumental record from Walden Pond (Massachusetts, USA) showed that WA and WAT reconstructed most closely pH and that GLR reconstructions showed the best agreement with measured alkalinity, whereas ANN and GLR models were superior in reconstructing the secondary gradient variable TP. Log-transformation of environmental gradients improved model performance for alkalinity, but not much for TP. While square-root transformation of species data improved the performance of the ANN models, they did not affect the WA models. Untransformed species data resulted in better accordance of the TP inferences with the instrumental record using WA, indicating that, in some cases, ecological information encoded in the modern and fossil species data might be lost by square-root transformation. Thus it may be useful to consider different species data transformations for different environmental reconstructions. This study showed that the tested methods are equally suitable for the reconstruction of parameters that mainly control the diatom assemblages, but that ANN and GLR may be superior in modelling a secondary gradient variable. For example, ANN and GLR may be advantageous for modelling lake nutrient levels in North America, where TP gradients are relatively short.

Introduction

Quantitative reconstructions of past environments using freshwater and marine sediment records have become increasingly accepted over the last decades (Birks 1998). Inference models based on modern relationships between biota (such as diatoms) and the environment [pH, temperature, total

phosphorus (TP), etc.] are routinely applied to fossil biological data in order to infer quantitative environmental values for periods without adequate instrumental data coverage (Kauppila et al. 2002; Ramstack et al. 2003; Siver et al. 2003). In an attempt to obtain the potentially most reliable reconstructions, it is beneficial to compare reconstructed values based on different methods and to assess critical issues of the methodology employed (e.g., data screening, transformations) (Birks 1998). However, in light of the large number of existing models, such considerations have only rarely been addressed (Korsman and Birks 1996; Hall et al. 1997).

Recently, artificial neural networks (ANNs) have been introduced to paleolimnological research and show promising performance when modelling pH with diatoms (Racca et al. 2001). However, the outputs of ANN models have not yet been comprehensively compared to the outputs of standard approaches [e.g., weighted averaging (WA) regression and calibration (ter Braak and van Dam 1989); weighted averaging partial least squares regression (WA-PLS) (ter Braak and Juggins 1993)] in the application to fossil diatom data, by validation with instrumental data and through the use of other variables than pH. This paper is an attempt to fill this gap by comparing diatom-based reconstructions using common methods [Gaussian logit regression (GLR), WA with classical deshrinking (WA_{class}), WA with inverse deshrinking (WA_{inv}), WA with tolerance-downweighting (WAT), and WA-PLS] with estimates obtained by ANNs and with instrumental records for Walden Pond, Massachusetts.

Data

Training set

The water chemistry and modern surface sediment diatom data used to develop diatom-based inference models originate from the United States Environmental Monitoring and Assessment Program – Surface Waters (data available via internet: http://diatom.acnatsci.org/dpdc). In the northeastern United States (Maine, New Hampshire, Vermont, Massachusetts, Connecticut, New York, Rhode Island and New Jersey), 257 lakes were sampled during July and August 1991–1994. Details concerning sampling procedures and diatom sample processing are given in Dixit et al. (1999). A subset of 82 lakes was selected for model development and environmental reconstructions in lakes from Vermont, New Hampshire, Massachusetts and Connecticut (Figure 1; Köster et al. unpubl. data). The sites from Maine, New York, Rhode Island and New Jersey were excluded *a priori* in order to limit the calibration set to the geographical region where the lakes for paleolimnological studies are located. Model and reconstruction comparisons presented here are based on this smaller data set. The main characteristics of the data set are presented in Table 1 and the relation of the 82 surface diatom assemblages to major environmental variables and lake

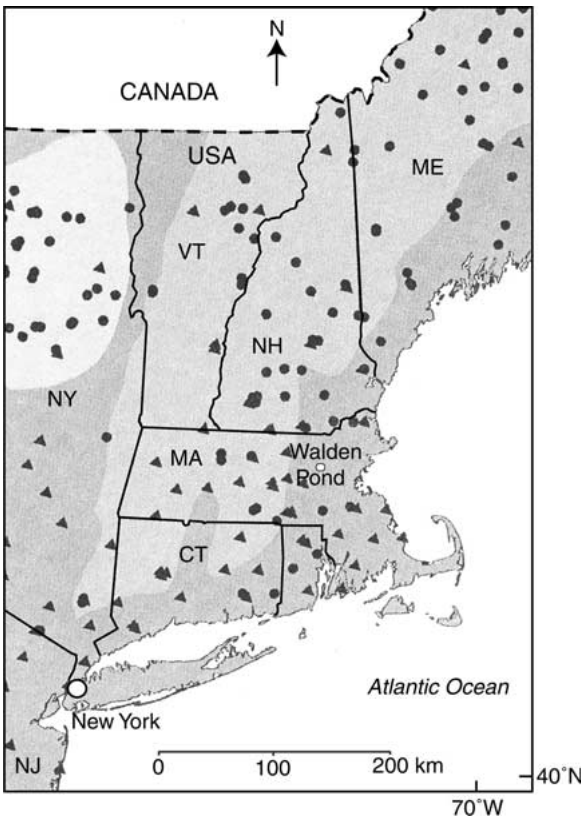

Figure 1. Map of the training set sites in the New England states Vermont (VT), New Hampshire (NH), Massachusetts (MA), and Connecticut (CT) and location of the study site Walden Pond. ME = Maine. NY = New York. NJ = New Jersey. Grey areas: New England Uplands. Dark grey areas: Coastal Lowlands/Plateau. Light grey area = Adirondacks. Modified from Dixit et al. (1999).

Table 1. Major characteristics of the diatom and environmental data of the training set. Ordination results are given for the 189-species set (cut-off-criterion: 1 occurrence at 1%).

No. of samples	82
No. of species	
Total	371
One occurrence at 1%	189
Min. 10 occurrences	121
Species DCA	
Lambda	7.2
CCA axis 1	
% variance	8.4
CCA axis 2	
% variance	4.1
pH	
Min.	4.99
Max.	8.6
Mean	7.5
Median	7.6
Length of gradient in DCCA	4.0
% variance in CCA	6.0 ($p = 0.005$)
Alkalinity	
Min.	−9.5
Max.	1858
Mean	399
Median	201.5
Length of gradient in DCCA	4.5
% variance in CCA	6.1 ($p = 0.005$)
TP	
Min.	0.85
Max.	109.5
Mean	16.1
Median	11
Length of gradient in DCCA	2.6
% variance in CCA	3.4 ($p = 0.005$)

DCA = detrended correspondence analysis. CCA axes 1 and 2 = first two axes in a canonical correspondence analysis with 17 environmental variables (see also Figure 2). CCA = CCA constrained to one variable. DCCA = detrended canonical correspondence analysis. % variance = percentage of variance in species data which is explained by this axis or variable.

characteristics are illustrated in the ordination biplot resulting from a canonical correspondence analysis (CCA; Figure 2).

Fossil data and analogs with training set

For reconstruction purposes, we used fossil diatom data from a 140-cm-long surface sediment core of Walden Pond (42°26.3′N, 71°20.4′W), spanning ca. 1600 years (Köster et al. unpubl. data). The ecological interpretation of the fossil diatom assemblages, the sedimentary stable isotope record as well as the instrumental data of Walden Pond indicate a clear, albeit seasonal change in the lake water chemistry to

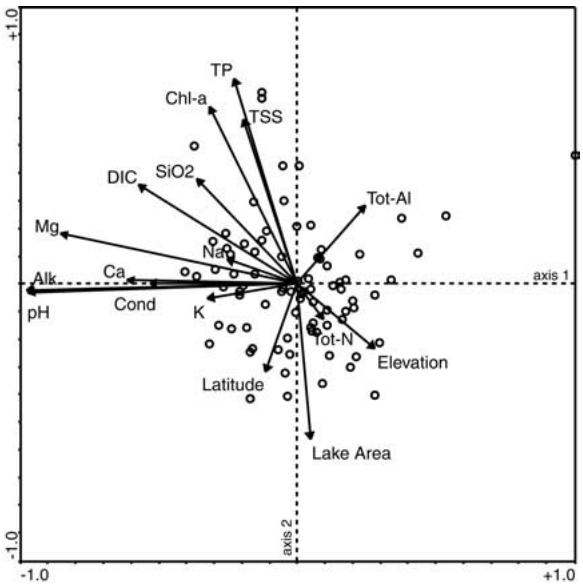

Figure 2. Environmental variables/sample biplot derived from CCA including subfossil diatom data from 82 New England sites and corresponding lake water measurements of TP, turbidity (TSS), chlorophyll a (Chl-a), silica (SiO_2), dissolved inorganic carbon (DIC), magnesium (Mg), sodium (Na), calcium (Ca), alkalinity (alk), pH, conductivity (Cond), potassium (K), latitude, lake area, total nitrogen (Tot-N), elevation and total aluminium (Tot-Al).

higher nutrient concentrations during the 20th century (starting at about 10 cm depth; Köster et al. unpubl. data). This change is evident in the ordination of the fossil percentage data in a principal components analysis (PCA), with inter-sample distance scaling and covariance matrix (Figure 3).

The analogs of the fossil samples with the training set were estimated by means of dissimilarity coefficients using chord distance (Overpeck et al. 1985), where fossil samples inside the 75% confidence interval of the mean minimum dissimilarity coefficient of the training set samples have good analogs, samples outside the 75% and inside the 95% confidence interval have poor analogs, and samples outside the 95% limit have no analogs (Laing et al. 1999). Fit of the fossil samples to the environmental gradient in the training set was estimated by CCA constrained to pH and TP as the single explanatory variables. Fossil samples with a residual distance inside the 90% confidence interval of the residual distances of the modern samples to the first CCA axis have a good fit, and samples outside the 90% limit have poor fit (Birks et al. 1990).

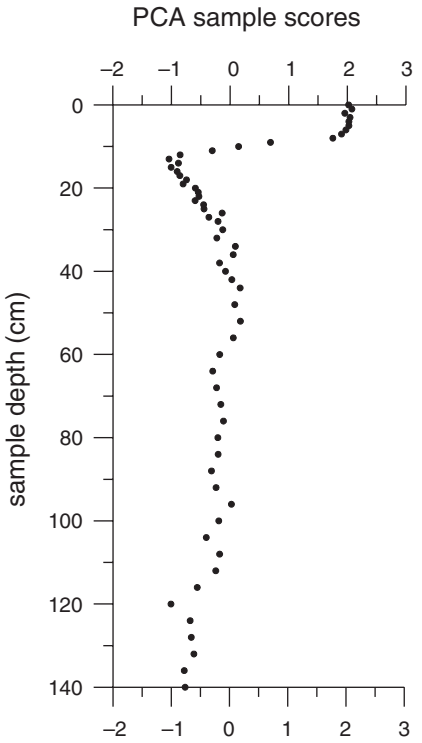

Figure 3. PCA scores of fossil diatom samples from Walden Pond sediments.

Analog and goodness-of-fit analyses of modern and fossil samples distinguished two different groups of core samples (Köster et al. unpubl. data). Samples between 140 and 9 cm had good analogs and fits, indicating that the fossil diatom flora is well represented in the modern training set. With the exception of levels 4, 3 and 0 cm, the samples from 8 to 0 cm had poor analogs, and all levels from 8 to 0 cm had poor fit to both pH and TP. The poor analogs in the upper 8 cm were caused by very high abundances of three species, which are present at lower abundances in the calibration set (e.g., *Asterionella formosa* Hassal, *Tabellaria flocculosa* (Roth) Kütz. str. IIIp *sensu* Koppen, *Fragilaria nanana* Lange-Bertalot, likely synonymous with *Synedra delicatissima* W. Smith in the training set). These species show a significant unimodal response to pH in the training set (Köster et al. unpubl. data), indicating that useful parameters were estimated for the pH model. However, only *S. delicatissima* and *T. flocculosa* str. IIIp show a unimodal response to TP, in contrast to *A. formosa* with no significant response to TP. This may have resulted in less reliable TP-model parameters for *A. formosa*.

Table 2. Instrumental record for pH, alkalinity and TP in Walden Pond, measured on epilimnetic water (Baystate Environmental Consultants 1995; Colman and Friesz 2001).

Year	1986	1989	1994	1997	1999
pH					
Median	n.d.	n.d.	n.d.	7.9	7.6
Average	n.d.	n.d.	n.d.	7.8	7.6
Min.	n.d.	n.d.	n.d.	6.4	6.4
Max.	n.d.	n.d.	n.d.	9.4	9.4
Number of samples	n.d.	n.d.	n.d.	129	111
Total phosphorus (μg l^{-1})					
Median	n.d.	40	10	8.4	6.5
Average	n.d.	57.5	<16.7	8.6	6.7
Min.	n.d.	10	<10	2.5	4.5
Max.	n.d.	140	60	19.5	8.7
Number of samples	n.d.	8	12	38	7
Alkalinity (μq l^{-1})					
Median	259	220	152	n.d.	n.d.
Average	269	225	155	n.d.	n.d.
Min.	214	220	134	n.d.	n.d.
Max.	340	240	182	n.d.	n.d.
Number of samples	6	8	12	n.d.	n.d.

Values preceded by a less than sign (<) indicate measurements below detection limit. n.d. = no data.

In summary, the analog analyses indicate that the reconstructions from 140 to 9 cm are reliable for pH and TP, and that the reconstructions from 8 to 0 cm are reliable for pH and probably not precise for TP. Analog analyses with alkalinity were not carried out, but as alkalinity is closely correlated with pH, it is likely to behave similarly.

Instrumental record of Walden Pond

Limnological surveys of the study site (Baystate Environmental Consultants 1995; Colman and Friesz 2001) provide instrumental data for validation of the diatom-inferred values (Table 2). As the records are not continuous and were established by different investigators, some details regarding the variables of interest are detailed below.

The pH of Walden Pond can change significantly during the course of 1 year (Table 2), particularly in the euphotic zone, where planktonic algal growth takes place and photosynthetic CO_2 depletion leads to pH values up to 9. The arithmetic mean pH of 7.8 for the years 1997 and 1999 is based on 129 and 111 measurements, respectively, dating from all months and integrating several depths of the epilimnion (0–15 m, in 1 m-steps). pH measurements

before 1990 were taken after transport of the water to the laboratory (Arthur Johnson, Massachusetts Department of Environmental Protection, pers. comm.), which may affect significantly the pH of the samples. Therefore we use only the pH data based on standard, *in situ* methods for validation of the diatom-inferred pH values.

Total phosphorus concentrations were below the detection limit of 10 $\mu g\, l^{-1}$ during much of the year, but peaked in summer when intensive recreational use caused significant external nutrient loading to the lake (Baystate Environmental Consultants 1995). Therefore, mean annual data for TP camouflage the nutrient enrichment over the summer months that strongly affects the diatom assemblages (Köster et al. unpubl. data).

In contrast to pH and TP, alkalinity was more stable throughout the year and thus easier to compare with diatom-inferred values.

Methods

Detrended correspondence analysis (DCA) on the raw species data indicated high variation in diatom assemblages with a total variance of 7.2 (Table 1). Therefore, methods assuming unimodal species responses to the environmental gradients, such as CCA, can be applied (Birks 1995). As the distributions of alkalinity and TP data were skewed, they were normalized by $\log_{10}$-transformation. Models based on non-transformed environmental data were developed for comparative purposes.

Diatom inference models were developed for the variables pH, alkalinity (alk) and TP, as they explained most of the variation in the surface diatom data, as indicated by CCA (Figure 2, Table 1). DCCA (detrended canonical correspondence analysis) with each individual variable as a predictor and detrending by segments and non-linear scaling indicated gradient lengths of larger than 2 standard deviations (SD) for all three variables (pH: 4.0, alk: 4.5, TP: 2.6), indicating that methods dealing with unimodal species responses are appropriate for model development. Although alkalinity and pH were highly correlated, we included models and reconstructions for both parameters. The performances of the models were compared by means of the determination coefficient (r^2), root mean square error of prediction (RMSEP), the mean and maximum bias as well as by the degree of coherence between the diatom-inferred values and the instrumental data. Bootstrapped performance statistics are presented, but comparisons between different methods are based on leave-one-out cross-validation (jackknifing), because it was the only available cross-validation method for ANN. Mean and maximum bias was calculated using the new approach presented by Racca and Prairie (2004).

The ordination techniques DCA, DCCA, CCA, and PCA were performed using the computer program CANOCO for Windows version 4.0. The computer program C^2 version 1.4 (Juggins 2003) was used to develop diatom-based inference models, to assess their performance, to reconstruct environmental variables and to calculate the sample-specific errors using bootstrapping. Aside from the common procedure for WAT, a new algorithm was used where small tolerances were replaced by a fraction (0.1) of the gradient. This strategy was developed in order to avoid the attribution of very high weight to rare species with small tolerances (Steve Juggins, pers. comm.), which otherwise may lead to low performance estimates (Table 3) and erroneous reconstructions (Figure 4c).

The models based on ANN were implemented using Yet Another Neural Network Simulator (Boné et al. 1998). Principles underlying this method are described in detail by Racca et al. (2001). For modelling alkalinity with ANN, one extreme site was removed, because the difference between the measured and the predicted value (residual) of this site was three times greater than the mean residual of all other sites.

The distribution of species over the environmental gradients was determined by testing a hierarchical set of response models (Huisman et al. 1993), implemented in the program HOF (Oksanen and Minchin 2002). For each species, HOF is returning the simplest of five possible models skewed (= asymmetric) unimodal, symmetric unimodal, sigmoidal (increasing or decreasing), plateau, no trend), which does not result in a statistically significant rise in the residual deviance. All species with at least 10 occurrences were selected for this analysis, which reduced the number of species included from 189 to 121. For percentage species data, a binomial error distribution was used in the program, whereas for square-root-transformed data, the Poisson distribution was chosen.

Table 3. Performance statistics for different models.

Variable	Method	Sites	Environmental data transformation	Species data transformation	r^2	RMSE	r^2_{jack}	$RMSEP_{jack}$	r^2_{boot}	$RMSEP_{boot}$	RMSEP back-trans	Mean $bias_{jack}$	Max. $bias_{jack}$
pH	WA_{class}	82	–	Square-root	0.85	0.28	0.81	0.30	0.80	0.31	–	−0.01	0.69
	WA_{inv}	82	–	Square-root	0.85	0.26	0.81	0.30	0.80	0.32	–	−0.01	0.89
	WAT_{inv}	82	–	Square-root	0.81	0.29	0.65	0.40	0.75	0.43	–	0	1.29
	WAT^*_{inv}	82	–	Square-root	0.87	0.24	0.83	0.28	0.83	0.32	–	0	0.85
	WA_{inv}	82	–	–	0.87	0.24	0.82	0.29	0.81	0.33	–	−0.01	0.57
	WA-PLS 2	82	–	Square-root	0.94	0.16	0.83	0.28	0.82	0.31	–	0.01	0.94
	GLR	82	–	–	0.88	0.24	0.79	0.31	0.74	0.38	–	−0.01	0.76
	ANN	82	–	–	0.93	0.18	0.77	0.34	–	–	–	−0.06	−0.65
	ANN	82	–	Square-root	0.99	0.08	0.86	0.25	–	–	–	−0.01	0.86
Alk	WA_{class}	82	$\log_{10}(x+10)$	Square-root	0.86	0.23	0.80	0.27	0.80	0.28	1.88	−0.01	1.08
	WA_{inv}	82	$\log_{10}(x+10)$	Square-root	0.86	0.22	0.80	0.27	0.79	0.28	1.92	−0.01	1.30
	WAT_{inv}	82	$\log_{10}(x+10)$	Square-root	0.75	0.29	0.47	0.45	0.75	0.38	2.41	0.04	1.66
	WAT^*_{inv}	82	$\log_{10}(x+10)$	Square-root	0.87	0.21	0.82	0.25	0.82	0.28	1.90	0	0.74
	WA-PLS 2	82	$\log_{10}(x+10)$	Square-root	0.95	0.13	0.80	0.26	0.81	0.28	1.89	0.01	1.45
	WA_{inv}	82	$\log_{10}(x+10)$	–	0.88	0.20	0.81	0.26	0.81	0.29	1.95	0	1.08
	WA_{inv}	82	–	–	0.76	240	0.63	301	0.62	327	–	127	−533
	WA_{inv}	82	–	Square-root	0.82	211	0.73	256	0.72	282	–	129	−487
	GLR	82	$\log_{10}(x+10)$	–	0.90	0.19	0.75	0.29	0.71	0.34	0.00	−0.01	1.43
	ANN	81	$\log_{10}(x+10)$	–	0.94	0.15	0.78	0.28	–	–	1.91	−0.05	0.82
	ANN	81	$\log_{10}(x+10)$	Square-root	0.96	0.11	0.85	0.23	–	–	1.69	−0.04	1.38
TP	WA_{class}	82	$\log_{10}$	Square-root	0.70	0.22	0.45	0.26	0.43	0.27	1.86	−0.01	0.47
	WA_{inv}	82	$\log_{10}$	Square-root	0.70	0.18	0.44	0.25	0.42	0.27	1.84	−0.01	0.56
	WAT_{inv}	82	$\log_{10}$	Square-root	0.43	0.25	0.11	0.34	0.34	0.31	2.04	0.02	−0.75
	WAT^*_{inv}	82	$\log_{10}$	Square-root	0.70	0.18	0.42	0.25	0.40	0.27	1.88	0	0.45
	WA-PLS 2	82	$\log_{10}$	Square-root	0.80	0.15	0.50	0.23	0.48	0.25	1.79	−0.01	0.53
	WA_{inv}	82	$\log_{10}$	–	0.47	0.20	0.37	0.26	0.35	0.28	1.89	−0.02	−0.54
	WA_{inv}	82	–	Square-root	0.81	8	0.41	14	0.39	15	–	−0.84	−54
	WA_{inv}	82	–	–	0.77	9	0.30	16	0.29	16	–	−1.41	−64
	GLR	82	$\log_{10}$	–	0.66	0.23	0.35	0.30	0.34	0.30	2.00	−0.05	0.39
	ANN	82	$\log_{10}$	–	0.66	0.20	0.33	0.27	–	–	1.86	0	0.56
	ANN	82	$\log_{10}$	Square-root	0.76	0.16	0.39	0.26	–	–	1.81	−0.01	−0.56

Alk = alkalinity. TP = total phosphorus. WA_{class} = Weighted averaging with classical deshrinking. WAT_{inv} = WA with inverse deshrinking and with tolerance-downweighting. WAT^*_{inv} = WAT_{inv} computed with small tolerances replaced by a fraction of the gradient (0.1), see Method section for details. WA_{inv} = WA with inverse deshrinking. WA-PLS 2 = WA partial least squares, two-component model. ANN = artificial neural networks. GLR = Gaussian logit regression (maximum likelihood). r^2 = apparent coefficient of determination of the regression of the predicted on the observed values. RMSE = apparent root mean square error of prediction. r^2_{jack} = jackknifed r^2. r^2_{boot} = bootstrapped r^2. $RMSEP_{jack}$ = jackknifed RMSE. $RMSEP_{boot}$ = bootstrapped RMSE. RMSEP back-trans = back-transformed RMSEP boot. Mean bias = average of residuals. Max. bias = maximum bias. Grey-shaded lines indicate the models with the best jackknifed performance for each variable.

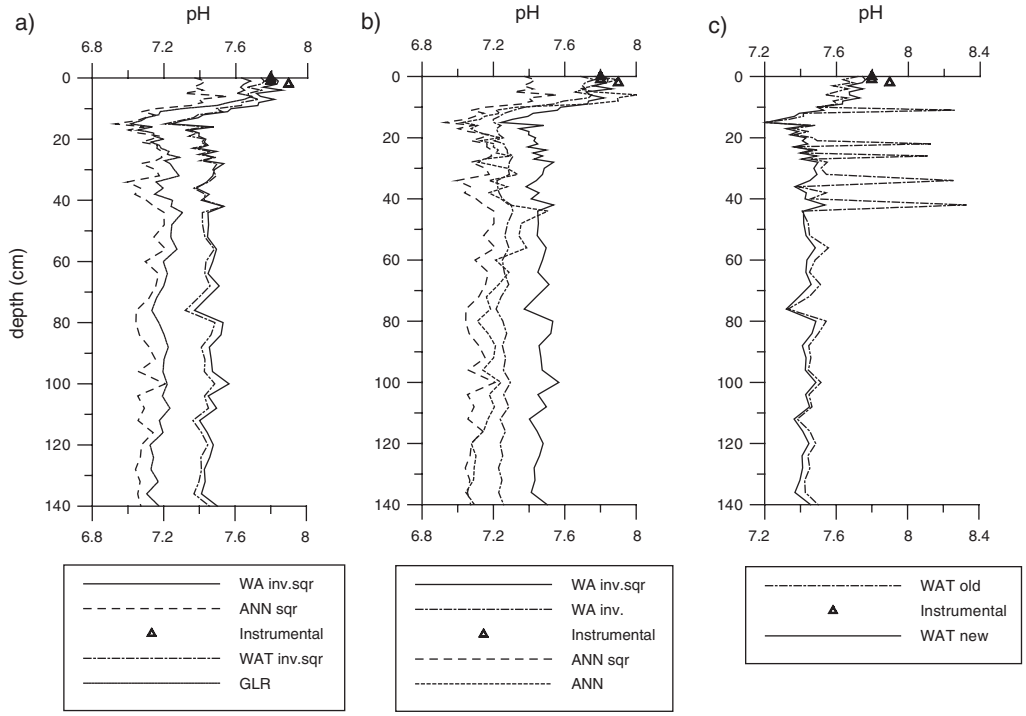

Figure 4. Comparison of diatom-inferred pH. For error values see text. (a) inferred pH produced by different methods. (b) inferred pH using different species transformations for WA_{inv} and ANN. (c) inferred pH using WAT. WAT old = WAT using the common algorithm. WAT new = WAT with minimum tolerance set to 0.1 of the gradient. WA_{inv} = weighted averaging with inverse deshrinking. WAT_{inv} = weighted averaging with tolerance-downweighting and inverse deshrinking. GLR = Gaussian logit regression (= maximum likelihood). ANN = artificial neural networks. Instrumental = measured pH. sqr = square-root-transformed species data.

Results and discussion

Performance of different models

In general, the predictive abilities of the models for pH and alkalinity as reflected by r^2_{jack} were better than those of the TP models (Table 3), a common feature of training sets including a long pH and a shorter TP gradient (Hall and Smol 1996; Dixit et al. 1999; Siver 1999). For pH and alkalinity, ANN resulted in the lowest RMSEP (0.23–0.25), followed by WA-PLS 2 and WAT with RMSEPs of 0.26–0.28 and WA (0.27–0.30). For modelling TP, WA-PLS 2 had slightly lower RMSEP (0.23) than WA_{class}, WA_{inv}, and ANN, which performed equally well with RMSEPs between 0.25 and 0.26. GLR had lower jackknifed performance, but better apparent performance than the WA methods for pH and alkalinity and lower apparent and bootstrapped performance for TP.

In contrast to WAT using a lower limit for species tolerances, such as applied here, WAT without this option, such as in the former version of C^2 and the programs WACALIB and CALIBRATE, performed less well by ca. 0.1–0.2 in RMSEP (Table 3). When rare taxa were excluded (Köster et al. unpubl. data), the performance of tolerance-downweighted WA equalled that of the other WA methods. This confirms results of previous studies (Birks 1994; Wilson et al. 1996), which showed that the performance of WAT decreases with inclusion of rare species. This may be explained by the difficulty of estimating a relatively realistic tolerance value for species with a low number of occurrences. In our training set, e.g., a tolerance value of 0.1 and an optimum of 8.5 was calculated for *Fragilaria construens* var. *venter*, which does not make sense ecologically and led to low model performance (Table 3) and unusual reconstructions (Figure 4c). The definition of a minimum tolerance for species in the training set, such as implemented in the

Table 4. Number and percentage of species with different response models for pH, alkalinity, and TP. Note that for the analyses only 121 species (those with at least 10 occurrences) out of 189 species were used.

Response	% pH	% alk	% alk log	sq alk	sq alk log	% TP	% TP log	sq TP	sq TP log
Number									
1	22	19	23	24	22	67	62	81	80
2	30	59	25	57	38	13	26	19	21
3	1	0	1	0	0	1	1	0	0
4	48	30	56	36	53	32	25	22	18
5	20	13	16	4	8	8	7	0	2
Percent									
1	18	16	19	20	18	55	51	67	66
2	25	49	21	47	31	11	21	16	17
3	1	0	1	0	0	1	1	0	0
4	40	25	46	30	44	26	21	18	15
5	17	11	13	3	7	7	6	0	2

alk = alkalinity. TP = total phosphorus. sq = species data square-root transformed. % = species percentage data. log = environmental data log-transformed. Species response code [for details see Huisman et al. (1993)]: 1 = no trend. 2 = increasing or decreasing trend. 3 = increasing or decreasing trend bounded below the maximum attainable response (= plateau). 4 = symmetrical unimodal response curve. 5 = skewed unimodal response curve.

recent version 1.4 of the computer program C^2 (Juggins 2003), resolved this problem. Alternatively, tolerance estimates for species that do not display unimodal distributions over the gradient may be incorrect. Such taxa were well represented in our models (44% and 67% for pH and TP, respectively; Table 4), percentages that are comparable to those reported for other diatom training sets (Lotter et al. 1998; Cameron et al. 1999). WAT has often been discarded in model choice for transfer functions because of lower performance (e.g., Birks et al. 1990; Hall and Smol 1992; Siver 1999). However, our results indicate that WAT may be a good alternative to WA for future inference model developments, if rare species are deleted or minimum tolerances are defined.

Transformation of species data

Square-root transformation of the species data did not considerably affect the performance of any WA model with decreases by 0.01 in RMSEP (Table 3). All WA models behaved similar in this regard for all variables, therefore only the results for WA_{inv} are presented as an example. For ANN models, it resulted in lower RMSEP by 0.05 and 0.09 for alkalinity and pH, respectively, but only by 0.01 for TP (Table 3). Therefore, we can conclude for our training set data that species transformation did not affect the performance of WA models, but improved the performance of some ANN models.

Transformation of environmental data

Powerful ANN models for alkalinity and TP were obtained when using log-transformed environmental data. As RMSEP and bias are given in untransformed units, comparison with log models is difficult. However, log-transformation resulted in higher r^2_{jack} for all WA models, indicating much better performance for alkalinity (increase in r^2_{jack} by 0.07–0.18) and fairly better performance for TP (increase of r^2_{jack} by 0.03–0.07). When alkalinity data were log-transformed, the number of unimodal species responses to this variable increased by 29 and 21 for untransformed and square-root-transformed species data, respectively. This indicates that log-transformation normalized the skewed raw alkalinity data, thereby helping to relate the presumed unimodal species distributions to alkalinity and thus improving model performance. After log-transformation of TP, however, the number of significant unimodal responses declined from 40 to 32 for untransformed species data and from 22 to 20 for square-root-transformed data (Table 4). In this case, log-transformation did not help to relate species distributions to the phosphorus data and therefore did not improve greatly the performance of the TP models. As the diatoms in our data set responded primarily to pH/alkalinity and less to the secondary TP gradient, they were perhaps generally less likely to show significant responses to TP, whatever the data manipulation might have been.

Comparison of reconstructions produced by different models and validation with the instrumental record

The paleoenvironmental reconstructions for Walden Pond produced by different methods showed generally the same pattern for pH and alkalinity, and to some extent also for TP. Relatively stable values were obtained for the samples from 140 to 12 cm and a more or less pronounced

increase is evident between 11 cm depth and the surface (0 cm) (Figures 4–6). This change reflects the recent, major shift in fossil diatom assemblages as demonstrated by the PCA sample scores (Figure 3). Nonetheless, in some cases the absolute diatom-inferred values differed between methods.

pH

All WA models and WA-PLS 2 produced very similar pH reconstructions, therefore we only present WA_{inv} and WAT_{inv} for model comparison (Figure 4a). From 140 to ca. 10 cm, GLR and ANN reconstructed lower values than the WA models by ca. 0.4 units. From 8 to 0 cm, GLR values approximated the WA reconstructions, whereas ANN values remained lower by ca. 0.4 units. The methods that estimated most closely the measured data were WA_{inv}, WAT_{inv} and GLR, despite the lower performance statistics of the GLR model.

Different species transformations caused some differences between the model outputs. Square-root transformation led to higher values by 0.3 units for WA_{inv} inferences (Figure 4b), as well as to lower values by 0.4 units for ANN reconstructions in the samples from 11 to 0 cm (Figure 4b), compared to reconstructions using percentage species data. However, the differences remained well within the overlapping bootstrapped error ranges of the methods (0.32 and 0.33 for WA_{inv}; 0.25 and 0.34 for ANN) indicating that the pH reconstructions are not significantly affected by different species data transformations.

The differences between WAT-based reconstructions without defining a minimum tolerance (common method) and with tolerance limit (new method) are large in five samples, with deviations around 0.8 pH units (Figure 4c). As all other methods reconstructed similar values to those generated by the new WAT method, without yielding abrupt pH changes in the past, the old method appears to produce unrealistic deviations in reconstructions. This problem is likely related to the unreliable tolerance estimates for rare species in the training set and can be avoided by defining a minimum tolerance, as discussed above (Steve Juggins, pers. comm.).

Alkalinity

Back-transformed diatom-inferred alkalinity values were comparable for all WA models (Figure 5a). ANN and GLR reconstructed lower values by ca. 100 μeq l^{-1} from 140 to 9 cm. From 8 to 0 cm, GLR approximates the WA reconstructed values, whereas ANN values remained lower than the WA values by ca. 200 μeq l^{-1} (Figure 5a). The difference seemed to be larger in recent sediments, with a maximum divergence in diatom-inferred alkalinity between the ANN and the WA_{inv} model of ca. 250 μeq l^{-1}, but the log-alkalinity values showed that the differences were equally large throughout the whole sediment sequence (Figure 5b). The exponential function underlying the back-transformation of log values boosted the reconstructed values (and the associated prediction errors) in the upper levels and made them appear artificially high. Actually, the differences between the ANN and WA_{inv} reconstructions were within the overlapping errors of both inferences, which intersect by about 50–100 μeq l^{-1} (data not shown).

Each of the methods inferred closely one of the different measured alkalinity values, but over- or under-estimated the other values, with one value outside the error limit of ANN and WA reconstructions, respectively. The coincidence of these deviations with poor analogs for the levels 8–0 cm suggests that the no-analog situation plays a role in the observed pattern. When species, which are rare in the training set (here, *A. formosa* and *F. nanana*), become a dominant part of the fossil assemblage, unusual reconstruction values can be the result. Classical methods are known to extrapolate better at the end of gradients (ter Braak 1995), a situation that may be present in the recent sediments of Walden Pond. However, in our example WA_{class} (data not shown) and WA_{inv} resulted in the same reconstructions, indicating that both methods dealt likewise with the no-analog situation. The only method that approximated each of the values sufficiently was GLR, despite its lower performance. In our case, GLR appears to provide a "mean" reconstruction between WA and ANN, which simulates a consensus reconstruction developed by combining results of different procedures, such as recommended by Birks (1995). The recent declining tendency in alkalinity was not evident in the reconstructions, perhaps indicating a delayed response of the diatoms to the change.

Different data transformations resulted in different alkalinity reconstructions from 140 to 10 cm,

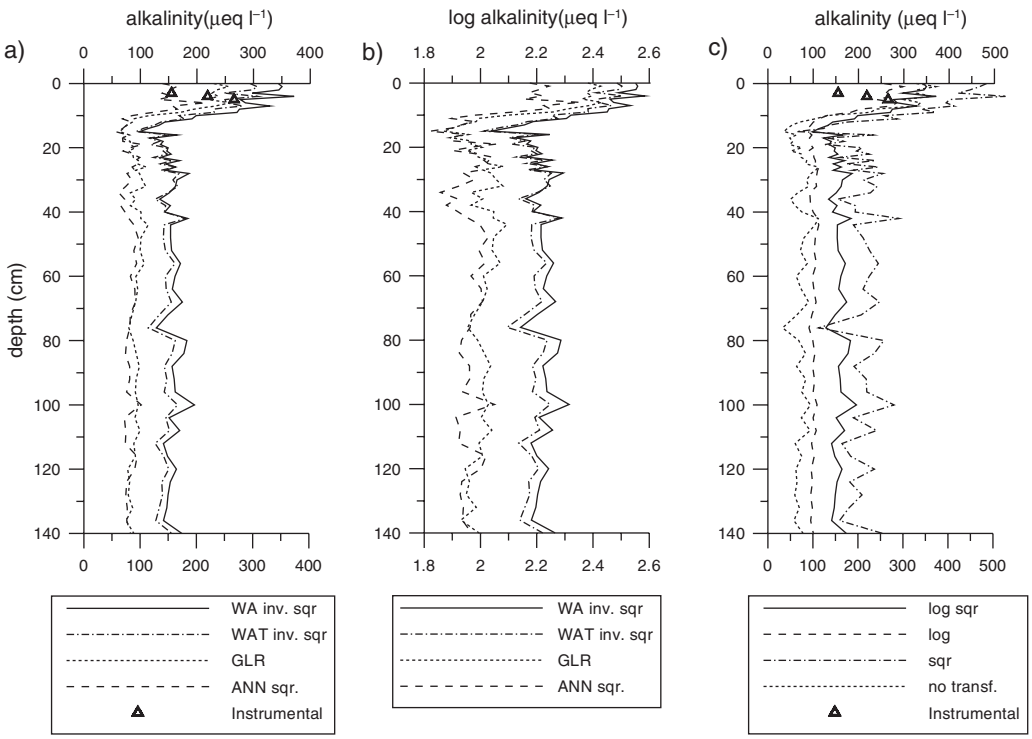

Figure 5. Comparison of diatom-inferred alkalinity. For error values see text. (a) inferred alkalinity produced by different methods. (b) inferred log-alkalinity. (c) inferred alkalinity using different data transformations. WA_{inv} = weighted averaging with inverse deshrinking. WAT_{inv} = weighted averaging with tolerance-downweighting and inverse deshrinking. GLR = Gaussian logit regression (= maximum likelihood). ANN = artificial neural networks. Instrumental = measured alkalinity. log sqr = log alkalinity, square-root transformed species data. log = log-alkalinity, untransformed species data. sqr = alkalinity, square-root-transformed species data. no transf. = no data transformation.

with the largest differences between models with and without square-root transformation (Figure 5c). In the upper 8 cm, the reconstructions of the models converge, except of the model using untransformed alkalinity data and square-root transformation of species data, which exceeded by 100–170 μeq l^{-1} the other inferences. The WA alkalinity models without species transformation produced similar reconstructions to those of ANN and GLR displayed in Figure 5a, suggesting that WA models without species transformation approximated better the consensus reconstruction, as discussed above. Noticeably, the square-root transformation of species abundances, which normally should stabilize variances, resulted in higher intra-series variability, as indicated by higher SDs in the reconstructed data (33 vs. 6 μeq l^{-1} for the reconstructions from 140 to 10 cm for WA_{inv} sqr and WA_{inv}, respectively, note that this is also the case for pH and TP reconstructions). This may be a disadvantage when quantitative reconstructions are to be correlated with other independent proxy estimates.

Total phosphorus

From 140 to 9 cm, all methods produced a similar, stable trend, but different values between ca. 3 μg l^{-1} for GLR and ca. 7 μg l^{-1} for WA and WAT (Figure 6a). From 8 to 0 cm, ANN and GLR-inferred values increased by around 4 μg l^{-1}, whereas WA_{inv} and WAT-inferred values continued to fluctuate in the same range as in the other samples, with a slight trend to lower values. The weighted averaging methods resulted in more inter-sample variability than ANN and GLR throughout the core (Figure 6a).

When different data transformations were applied in WA_{inv}, the reconstructed values differed

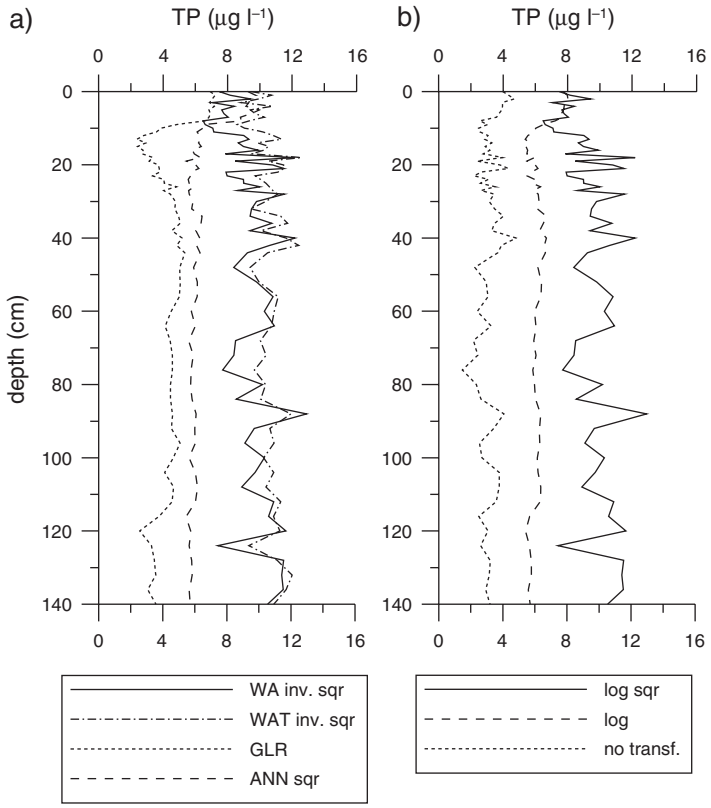

Figure 6. Comparison of diatom-inferred TP. For error values see text. (a) inferred total phosphorus produced by different methods. (b) inferred TP using different data transformations. WA_{inv} = weighted averaging with inverse deshrinking. WAT_{inv} = weighted averaging with tolerance-downweighting and inverse deshrinking. GLR = Gaussian logit regression (= maximum likelihood). ANN = artificial neural networks. Instrumental = measured TP. log sqr = log TP, square-root-transformed species data. log = log TP, untransformed species data. no transf. = no data transformation.

by 3 μg l^{-1} (Figure 6b). Most reconstructions showed no trend, except of the log TP model without species transformation, which increased in the upper levels by 2 μg l^{-1}. The model with untransformed TP data and square-root-transformed species data resulted in negative values and is therefore not presented.

Almost all reconstructed TP values were smaller than 10 μg l^{-1}, corresponding well to the measured mean annual TP of 10 μg l^{-1} and less. The WA_{inv} model using untransformed species data, GLR and the ANN model tracked best the important assemblage change and higher seasonal TP maxima (Table 2) by inferring increasing TP values in the upper levels, although they had similar or lower statistical performance as the other WA models (Table 3). Similar results were obtained in some Ontario lakes, where the best accordance of diatom-inferred TP with the instrumental record was achieved using untransformed species data in WA (Hall et al. 1997). In our case, this result is likely due to the fact that few species dominate the recent samples. Their high abundances helped detect a signal that was otherwise down-weighted by transformation. Another explanation may be found in the modern species response to TP. The untransformed TP and species data resulted in 40 significant skewed and symmetric unimodal distributions compared to about the half of that when square-root-transformed diatom data were related to TP (22) and log TP (20) (Table 4). It appears that the square-root transformation of species data flattened unimodal species responses to TP to sigmoidal or non-significant ones, thereby removing information that may be useful for environmental inferences performed with this calibration set. However, square-root transformation may have no effect at all in reconstructions using more

diverse fossil assemblages with low abundances displayed by many species.

Reasons for the better correspondence between the ANN reconstructed TP values and the instrumental record may be that this method copes better than WA with poor-analog situations and/or high percentage of non-unimodal species responses. The latter is a known advantage of the ANN models (Racca et al. 2001). Therefore, this method may prove advantageous for modelling lake nutrient status in regions where the development of TP models using weighted averaging is challenging due to relatively short TP gradients, such as in North America.

As the TP models had lower performance in comparison to the pH and alkalinity models, the TP reconstructions may be considered less reliable. Nonetheless, the comparison with the instrumental record showed that our models were suitable for the reconstruction of past TP concentrations. Also, reconstructions of TP for lake Saint-Augustin in Québec using our TP model without species transformation showed good agreement with measured TP (K. Roberge, unpubl. data).

Conclusions

The comparison of different methods in diatom-based reconstruction and validation by the instrumental record showed that weighted averaging with classical and inverse deshrinking as well as with and without tolerance-downweighting, GLR and ANN all provided reliable inference models and reconstructions for pH and alkalinity. In our study from Walden Pond, this was even the case under poor-analog conditions. However, using the common algorithm for WAT resulted in lower performance statistics than simple WA and in unusual reconstruction values, because rare species with small tolerances were highly weighted. Defining a minimum tolerance for the modern species, such as available in the new computer program C^2 version 1.4, resolved these problems. While WAT was often discarded for transfer function development in previous studies because of low performance, our results indicate that it may be an equally useful tool for paleoecological studies as simple WA.

Not all WA models for TP did track a nutrient enrichment which was evident in the species assemblages and which was inferred by the statistically equally well performing ANN and GLR models. These results suggest that the tested methods are equally suitable for the reconstruction of parameters that mainly control the diatom assemblages, but that ANN and GLR may be superior in modelling a secondary gradient variable. For example, ANN and GLR may be advantageous for modelling lake nutrient levels in North America, where TP gradients are relatively short.

Logarithmic transformation of skewed environmental data improved much the model performance of alkalinity, but only slightly the TP models. It appears that the primary response of diatom species to the main gradient inhibits the sensitivity of model performance to data manipulations regarding the secondary gradient.

Square-root transformation of species data did not improve the performance or the paleoecological inferences of the WA models, but was advantageous for the ANN models. Untransformed species data resulted in better accordance of the TP inferences with the instrumental record using WA, indicating that, in some cases, ecological information encoded in the modern and fossil species data might be lost by square-root transformation. In contrast to our expectations, square-root transformation of species data did not stabilize variances, but created more noisy reconstructions than models without square-root transformation. Thus it may be useful to consider different species data transformations for different environmental reconstructions.

Obviously, these conclusions cannot be generalized as they are only based on tests using one modern and fossil data set. Future work on other fossil sequences from the same region will assess if diatom models using ANNs are more widely applicable.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada through an operating grant to R. Pienitz, and by a Marie-Curie-Fellowship awarded to D. Köster. Data for model development were provided by S. Dixit. We would like to thank H.J.B. Birks for comments and encouragement in an early stage of this study. We thank Steve Juggins for data discussion and for implementing the new option for WAT in the computer program C^2. Comments

and criticisms by an anonymous reviewer and H.J.B. Birks helped to significantly improve this manuscript.

References

Baystate Environmental Consultants I, 1995. Study of trophic level conditions of Walden Pond, Concord, Massachusetts. Commonwealth of Massachusetts, Department of Environmental Management, Division of Resource Conservation. East Longmeadow, 111 pp.

Birks H.J.B. 1994. The importance of pollen and diatom taxonomic precision in quantitative palaeoenvironmental reconstructions. Rev. Palaeobot. Palynol. 83: 107–117.

Birks H.J.B. 1995. Quantitative paleoenvironmental reconstructions. In: Maddy D. and Brew J.S. (eds), Statistical Modelling of Quaternary Science Data. Quaternary Research Association, Cambridge, pp. 161–254.

Birks H.J.B. 1998. Numerical tools in paleolimnology – progress, potentialities, and problems. J. Paleolim. 20: 307–332.

Birks H.J.B., Line J.M., Juggins S., Stevenson A.C. and ter Braak C.J.F. 1990. Diatoms and pH reconstruction. Philos. T. Roy. Soc. B 327: 263–278.

Boné R., Cruciano M. and Asselin de Beauville J.-P. 1998. Yet Another Neural Network Simulator. Proceedings of the conference Neural Networks and their Applications (NEURAP '98), Marseilles, France, 421–424.

Cameron N.G., Birks H.J.B., Jones V.J., Berge F., Catalan J., Flower R.J., Garcia J., Kawecka B., Koinig K.A., Marchetto A., Sanchez-Castillo P., Schmidt R., Sisko M., Solovieva N., Stefkova E. and Toro M. 1999. Surface-sediment and epilithic diatom pH calibration sets for remote European mountain lakes (AL : PE Project) and their comparison with the Surface Waters Acidification Programme (SWAP) calibration set. J. Paleolim. 22: 291–317.

Colman J.A. and Friesz P.J. 2001. Geohydrology and Limnology of Walden Pond, Concord, Massachusetts. U.S. Geological Survey. Northborough, Massachusetts, Water-Resources Investigations Report, 61 pp.

Dixit S.S., Smol J.P., Charles D.F., Hughes R.M., Paulsen S.G. and Collins G.B. 1999. Assessing water quality changes in the lakes of the northeastern United States using sediment diatoms. Can. J. Fish. Aquat. Sci. 56: 131–152.

Hall R.I. and Smol J.P. 1992. A weighted-averaging regression and calibration model for inferring total phosphorus concentration from diatoms in British Columbia. Freshwat. Biol. 27: 417–437.

Hall R.I. and Smol J.P. 1996. Paleolimnological assessment of long-term water-quality changes in south-central Ontario lakes affected by cottage development and acidification. Can. J. Fish. Aquat. Sci. 53: 1–17.

Hall R.I., Leavitt P.R., Smol J.P. and Zirnhelt N. 1997. Comparison of diatoms, fossil pigments and historical records as measures of lake eutrophication. Freshwat. Biol. 38: 401–417.

Huisman J., Olff H. and Fresco L.F.M. 1993. A hierarchical set of models for species response analysis. J. Vegetation Sci. 4: 37–46.

Juggins S. 2003. C^2 User Guide. Software for Ecological and Paleoecological Data Analysis and Visualisation. University of Newcastle, Newcastle upon Tyne, UK, 69 pp.

Kauppila T., Moisio T. and Salonen V.P. 2002. A diatom-based inference model for autumn epilimnetic total phosphorus concentration and its application to a presently eutrophic boreal lake. J. Paleolim. 27: 261–273.

Korsman T. and Birks H.J.B. 1996. Diatom-based water chemistry reconstructions from northern Sweden: a comparison of reconstruction techniques. J. Paleolim. 15: 65–77.

Laing T.E., Rühland K. and Smol J.P. 1999. Past environmental and climatic changes related to tree-line shifts inferred from fossil diatoms from a lake near the Lena River Delta, Siberia. Holocene 9: 547–557.

Lotter A.F., Birks H.J.B., Hofmann W. and Marchetto A. 1998. Modern diatom, cladocera, chironomid, and chrysophyte cyst assemblages as quantitative indicators for the reconstruction of past environmental conditions in the Alps. II. Nutrients. J. Paleolim. 19: 443–463.

Oksanen J. and Minchin P.R. 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecol. Model. 157: 119–129.

Overpeck J.T., Webb T. and Prentice I.C. 1985. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of Modern Analogs. Quat. Res. 23: 87–108.

Racca J.M.J. and Prairie Y.T. 2004. Apparent and real bias in numerical transfer functions in paleolimnology. J. Paleolim. 31: 117–124.

Racca J.M.J., Philibert A., Racca R. and Prairie Y. 2001. A comparison between diatom-based inference models using Artificial Neural Networks (ANN), Weighted Averaging (WA) and Weighted Averaging Partial Least Squares (WA-PLS). J. Paleolim. 26: 411–422.

Ramstack J.M., Fritz S.C., Engstrom D.R. and Heiskary S.A. 2003. The application of a diatom-based transfer function to evaluate regional water-quality trends in Minnesota since 1970. J. Paleolim. 29: 79–94.

Siver P.A. 1999. Development of paleolimnologic inference models for pH, total nitrogen and specific conductivity based on planktonic diatoms. J. Paleolim. 21: 45–59.

Siver P.A., Ricard R., Goodwin R. and Giblin A.E. 2003. Estimating historical in-lake alkalinity generation from sulfate reduction and its relationship to lake chemistry as inferred from algal microfossils. J. Paleolim. 29: 179–197.

ter Braak C.J.F. 1995. Nonlinear methods for multivariate statistical calibration and their use in palaeoecology – a comparison of inverse (k-nearest neighbors, partial least-squares and weighted averaging partial least-squares) and classical approaches. Chemometrics Intell. Lab. Syst. 28: 165–180.

ter Braak C.J.F. and van Dam H. 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia 178: 209–223.

ter Braak C.J.F. and Juggins S. 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia 269/270: 485–502.

Wilson S.E., Cumming B.F. and 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. Can. J. Fish. Aquat. Sci. 53: 1580–1594.