PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Diatoms and pH Reconstruction

H. J. B. Birks, J. M. Line, S. Juggins, A. C. Stevenson and C. J. F. Ter Braak

Phil. Trans. R. Soc. Lond. B 1990 **327**, 263-278
doi: 10.1098/rstb.1990.0062

Phil. Trans. R. Soc. Lond. B **327**, 263–278 (1990)
Printed in Great Britain

Diatoms and pH reconstruction

By H. J. B. Birks[1], J. M. Line[2], S. Juggins[3], A. C. Stevenson[4] and C. J. F. ter Braak[5]

[1] *Botanical Institute, University of Bergen, Allégaten* 41, *N*-5007 *Bergen, Norway*
[2] *University of Cambridge Computer Laboratory, Pembroke Street, Cambridge CB*2 3*QG, U.K.*
[3] *Palaeoecology Research Unit, Department of Geography, University College,* 26 *Bedford Way, London WC*1*H* 0*AP, U.K.*
[4] *Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne NE*1 7*RU, U.K.*
[5] *Agricultural Mathematics Group, Wageningen, Box* 100, 6700 *AC Wageningen, The Netherlands, and Research Institute for Nature Management, Box* 46, 3956 *ZR Leersum, The Netherlands*

[Microfiche in pocket]

Palaeolimnological diatom data comprise counts of many species expressed as percentages for each sample. Reconstruction of past lake-water pH from such data involves two steps; (i) regression, where responses of modern diatom abundances to pH are modelled and (ii) calibration where the modelled responses are used to infer pH from diatom assemblages preserved in lake sediments. In view of the highly multivariate nature of diatom data, the strongly nonlinear response of diatoms to pH, and the abundance of zero values in the data, a compromise between ecological realism and computational feasability is essential. The two numerical approaches used are (i) the computationally demanding but formal statistical approach of maximum likelihood (ML) Gaussian logit regression and calibration and (ii) the computationally straightforward but heuristic approach of weighted averaging (WA) regression and calibration.

When the Surface Water Acidification Project (SWAP) modern training set of 178 lakes is reduced by data-screening to 167 lakes, WA gives superior results in terms of lowest root mean squared errors of prediction in cross-validation. Bootstrapping is also used to derive prediction errors, not only for the training set as a whole but also for individual pH reconstructions by WA for stratigraphic samples from Round Loch of Glenhead, southwest Scotland covering the last 10000 years. These reconstructions are evaluated in terms of lack-of-fit to pH and analogue measures and are interpreted in terms of rate of change by using bootstrapping of the reconstructed pH time-series.

Introduction

Diatoms are good ecological indicators of lake-water pH (Battarbee *et al.* 1986). In recent years this feature has been exploited to reconstruct pH from diatom assemblages preserved in lake sediments (Battarbee 1984; Battarbee & Charles 1987) and a variety of numerical procedures have been developed for quantitative inference of pH. Quantitative reconstruction of pH from diatoms is, in practice, a two-step process. First, the responses of modern diatoms to contemporary pH are modelled. This is a regression problem (ter Braak & Looman 1987; ter Braak & Prentice 1988) and involves a modern, training set of diatom assemblages ('response' variables) from surface lake-sediment samples with associated pH data ('predictor' variable).

Second, the modelled responses are used to infer past pH from the composition of fossil diatom assemblages. This is a calibration problem (ter Braak 1987*a*; ter Braak & Prentice 1988).

There are at least five major ecological assumptions in quantitative, palaeoenvironmental reconstructions (Imbrie & Webb 1981).

1. The taxa in the training set are systematically related to the physical environment in which they live.

2. The environmental variable to be reconstructed (in our case pH) is, or is linearly related to, an ecologically important variable in the system of interest.

3. The taxa in the training set are the same as in the fossil data-set and their ecological responses have not changed significantly over the timespan represented by the fossil data. Contemporary patterns of diatom abundance in relation to pH can thus be used to reconstruct pH changes through time.

4. The mathematical methods used in regression and calibration adequately model the biological responses to the environmental variable of interest.

5. Environmental variables other than the one of interest (e.g. pH) have negligible influence, or their joint distribution with the variable of interest in the fossil set is the same as in the training set.

Diatom data contain many taxa (*ca.* 100–300 taxa) and many zero values; values are commonly expressed as percentages of the total valves counted in a sample. They are thus closed, multivariate compositional data and have a constant-sum constraint. Diatom responses to pH are frequently nonlinear. These features confer important statistical properties on the data.

A variety of numerical procedures (reviewed by Battarbee (1984), Charles (1985) and Birks (1987)) have been used to reconstruct pH but none of these are fully satisfactory either theoretically or ecologically (Birks 1987; ter Braak & van Dam 1989). They are nearly all variants of the basic multiple linear regression model and usually involve grouping diatoms into ecological categories (see, for example, Charles (1985); Davis & Anderson (1985); Flower (1986); Charles & Smol (1988)). ter Braak and van Dam (1989) introduced two procedures, maximum likelihood (ML) and weighted averaging (WA) regression and calibration (see also ter Braak (1987*b*)). They are more sound theoretically and perform better than other, more widely used 'ad hoc' pH reconstruction techniques. These two procedures form the basis of this paper; WA regression and calibration are used for pH reconstructions within the Surface Water Acidification Project (SWAP).

The data

(*a*) *The modern data*

The SWAP training set consists of diatom counts of 178 surface samples from lakes in England (five lakes), Norway (51), Scotland (60), Sweden (30) and Wales (32). It includes all taxa (267) that are present in at least two samples with an abundance of 1 % or more in at least one sample and that are identified to species level or below. Abundances are expressed as percentages of the total diatom count (*ca.* 500 valves) for that sample. The sum of the percentages of taxa included range from 65.0 % to 99.1 % (mean = 92.6 %) of the total diatom count per sample.

The pH data for each lake are based on the arithmetic mean of $[H^+]$ (Barth 1975; cf. Middleton & Rovers 1976; Charles 1985), after initial data screening (Munro *et al.*, this

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

symposium). Many lakes have pH data based on three or more readings (131 lakes), though some only have one (32) or two (15) readings. The pH range is 4.33–7.25 ($\bar{x} = 5.59$, median = 5.51, standard deviation (s.d.) = 0.77). Further details of the data and their taxonomic consistency are given by Munro *et al.* (this symposium).

(*b*) *The fossil data*

We use the data of Jones *et al.* (1989) from Round Loch of Glenhead (RLGH), Galloway, southwest Scotland for reconstruction purposes. The 101 samples from 0.3 to 256.5 cm in core RLGH3 cover the last 10000 years. They contain some taxa absent in the training set and vice versa. Only taxa present in both sets are included. These represent 72.8–98.7% (mean = 88.2%) of the total diatom count per sample in the core.

Theory and methods

(*a*) *Notation*

We use the following notation throughout; x is the environmental variable to be reconstructed, in our case pH; x_i is the value of x in sample (in our case lake) i; y_{ik} is the abundance of taxon k in sample i ($y_{ik} \geqslant 0$) ($i = 1, \ldots n$ lakes and $k = 1, \ldots m$ diatom taxa); $\hat{x}_i$ is the estimated or inferred value of x for sample i.

(*b*) *Maximum likelihood regression and calibration*

The basic idea (ter Braak & van Dam 1989; ter Braak 1987*a*, *b*; ter Braak & Prentice 1988) is that the relation between the abundance of a diatom taxon and pH can be modelled by an ecological response curve consisting of systematic and random (error) components. Such a curve is fitted to the training set by nonlinear regression. The response curves and their assumed error structure form a statistical model of diatom composition in relation to pH. The curves for all taxa determine jointly what diatom composition is expected at a given pH. This model of responses and their error structure is used to calculate the probability that a particular pH would occur with a given assemblage over the range of possible pH values. The pH that gives the highest probability is the ML estimate.

There are many types of ecological response curves. A compromise is necessary between ecological realism and simplicity (ter Braak & van Dam 1989); the Gaussian unimodal response model with symmetric unimodal curves is a suitable compromise (ter Braak 1987*b*).

The Gaussian logit model is usually applied to presence–absence data (see, for example, ter Braak & Looman (1986)). However, it can be used, as here, as a quasi-likelihood model for proportions and as an approximation to the more complex multinomial logit model (ter Braak & van Dam 1989). The multinomial model can be difficult to fit and its parameters difficult to interpret because of indeterminacies (ter Braak 1988).

Following ter Braak and van Dam (1989), we fitted a Gaussian logit model to all 229 taxa that occurred in six or more lakes in the training set by logit regression (with binomial error structure). (Oksanen *et al.* (1988) fitted the related Gaussian model with Poisson error structure.) From the Gaussian logit regression coefficients, the optimum ($\hat{u}_k$), tolerance ($\hat{t}_k$) and height of the peak ($\hat{c}_k$) of the fitted Gaussian response curve were calculated (ter Braak & Looman 1986, 1987), along with the approximate 95% confidence intervals for the estimated u_k and the standard error of the estimated t_k (ter Braak & Looman 1986, 1987).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

For each taxon, the significance ($\alpha = 0.05$) of the Gaussian logit model was tested against the simpler linear-logit (sigmoidal) model by a residual deviance test. The significance ($\alpha = 0.05$) of the Gaussian logit regression coefficient b_2 against the null hypothesis ($b_2 \geqslant 0$) was also assessed by a one-sided t-test (ter Braak & Looman 1986, 1987). If the null hypothesis was rejected in favour of $b_2 < 0$, the taxon's optimum was considered significant. If either the Gaussian unimodal model or the optimum were not significant, the linear logit model and its regression coefficient b_1, were tested against the null model that the taxon showed no relation to pH, by using deviance and two-sided t-tests.

For taxa with estimated optima clearly outside the range of sampled pH values, and with a significant linear logit model, the optimum was assumed to be the lowest pH sampled for decreasing linear logit curves and the highest pH sampled for increasing linear logit curves. Some taxa had fitted curves with a minimum ($b_2 > 0$) instead of a maximum. For these taxa a linear logit model was fitted and, for WA calibration (table 3), the optima were taken to be the lowest or highest pH in the training set for decreasing curves and increasing curves, respectively. The tolerances are defined only for taxa with unimodal response curves.

The response curves and their assumed error structure were then used in ML calibration to find the pH with the highest probability of producing the observed diatom assemblages for each sample in both training and fossil data sets. This was done by using an iterative Gauss–Newton numerical optimization procedure with Gallant's (1975) chopping rule for step-shortening. Estimates of WA were used as initial estimates. Some samples failed to converge, however, with this procedure.

Regression and calibration of ML are computer-intensive and liable to find local maxima rather than the overall maximum, especially when the taxon tolerances are very unequal. An alternative approach that is both simpler and computationally easier and has essentially the same aims as ML is WA regression and calibration (ter Braak & van Dam 1989; ter Braak & Prentice 1988).

(c) *Weighted averaging regression and calibration*

The idea behind WA (ter Braak 1987*b*) is that in a lake with a certain pH range, diatoms with their pH optima close to the lake's pH will tend to be the most abundant taxa present. A simple and ecologically reasonable estimate of a taxon's pH optimum is thus the average of all the pH values for lakes in which the taxon occurs, weighted by the taxon's relative abundance (WA regression). Conversely, an estimate of the lake's pH is the weighted average of the pH optima of all the taxa present (WA calibration). Taxa with a narrow pH tolerance or amplitude can, if required, be given greater weight in WA than taxa with a wide pH tolerance.

The WA estimate of a taxon's optimum (equivalent to abundance weighted mean (see, for example, Charles (1985); Charles & Smol (1988)), centroid), $\hat{u}_k$, is:

$$\hat{u}_k = \sum_{i=1}^{n} y_{ik} x_i \Big/ \sum_{i=1}^{n} y_{ik},$$

and a taxon's tolerance, $\hat{t}_k$, or weighted standard deviation is:

$$\hat{t}_k = \left[\sum_{i=1}^{n} y_{ik} (x_i - \hat{u}_k)^2 \Big/ \sum_{i=1}^{n} y_{ik} \right]^{\frac{1}{2}}.$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The estimated optima can be used to infer a lake's pH from its diatom assemblage (WA calibration) by:

$$\hat{x}_i = \sum_{k=1}^{m} y_{ik}\,\hat{u}_k \Big/ \sum_{k=1}^{m} y_{ik},$$

whereas a tolerance-weighted estimate would be:

$$\hat{x}_i = \left(\sum_{k=1}^{m} y_{ik}\,\hat{u}_k/\hat{t}_k^2\right)\Big/\left(\sum_{k=1}^{m} y_{ik}/\hat{t}_k^2\right).$$

The theory of WA and the conditions under which WA approximates ML are fully discussed by ter Braak (1985, 1987*b*), ter Braak & Looman (1986), ter Braak & Barendregt (1986) and ter Braak & Prentice (1988).

In WA reconstructions, averages are taken twice, once in WA regression and once in WA calibration. This results in shrinkage of the range of inferred pH values. To correct for this, a simple linear deshrinking was done by regressing the initial inferred values $\hat{x}_i$ for the training set on the observed values, x_i, by using the linear regression model, so-called 'classical regression':

$$\text{initial } \hat{x}_i = a + bx_i + \epsilon_i,$$

and

$$\text{final } \hat{x}_i = (\text{initial } \hat{x}_i - a)/b,$$

where a is the intercept and b is the slope of the linear regression (ter Braak 1988). ter Braak & van Dam (1989) discuss the importance of deshrinking. They used 'inverse' regression (where x_i is regressed on initial $\hat{x}_i$ values) to 'deshrink', because this minimizes the root mean squared error in the training set. Classical regression deshrinks more than inverse regression (see, for example, Lwin & Maritz (1982)); it takes inferred values further away from the mean. In our case, the mean lies in the pH interval where a lake's pH is very variable, and acidification studies require the reconstructions to be most precise at the lower end of the pH range in the training set. For that, classical regression is preferable (Martinelle 1970; Lwin & Maritz 1982). This deshrinking regression was also done for ML calibration to ensure comparability between results from the two approaches.

(*d*) *Summary statistics*

The root mean square of the error (RMSE) $(x_i - \hat{x}_i)$ was calculated for the training set for comparison of the predictive abilities of ML, WA and weighted averaging with tolerance downweighting (WA(tol)). Wallach & Goffinet (1989) discuss the value of RMSE as a means of evaluating how well a model can be expected to function as a predictive tool. The correlation (r) between x_i and $\hat{x}_i$ was also calculated. As RMSE is invariably under-estimated when based solely on the training set (ter Braak & van Dam 1989; Oksanen *et al.* 1988), split-sampling or cross-validation (Stone 1974; Snee 1977; Picard & Cook 1984) was used to derive a reliable estimate of prediction error and hence to evaluate the predictive abilities of the different methods. This involves randomly splitting the modern data into a training set and a test set, and ensuring that only taxa fulfilling the criteria of two or more occurrences and 1% or more in any one sample in the new training set are included.

(*e*) *Error estimation*

Bootstrapping, a computer-intensive resampling procedure (Efron 1982; Efron & Gong 1983; Diaconis & Efron 1983; Wallach & Goffinet 1989) was used to derive RMSE of prediction

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

for individual pH reconstructions. It also provides another estimate of the overall RMSE of prediction for the training set. The underlying theory is described in Appendix 1.

The idea is that in each of many bootstrap cycles (in our case 1000) a subset of training samples is selected randomly from the original training set to form a bootstrap training set of the same size as the actual training set. This mimics sampling variation in the training set. Sampling is with replacement, so that samples may be selected more than once. Typically some samples will not be selected, and these form a bootstrap test set. In each cycle, WA regression and calibration are used with the bootstrap training set to infer pH for the modern samples in the bootstrap test set. This parallels the use of test sets in cross-validation and provides another estimate of the RMSE of prediction (see Appendix 1). This estimate is less prone to bias (Efron 1982, 1983) because the bootstrap uses the full size of the training set rather than the smaller size used in cross-validation.

In each cycle WA calibration is also used to infer pH for the fossil samples. For each fossil sample, the standard deviation of inferred pH for all bootstrap cycles is calculated. A naive user of the bootstrap might think that this standard deviation, e.g. s_{i1}, is an estimate of prediction error. It is not, because s_{i1} would approach zero if the size of the training set steadily increases. In fact, s_{i1} is that part of the prediction error that is due to the estimation error in the taxon parameters in the calibration formulae $(\hat{u}_k, \hat{t}_k)$. The other, often larger, part of the prediction error is due to variation in taxon abundances at a given value of pH. (In the bootstrap cycles, the taxon abundances of a fossil sample were kept fixed so s_{i1} cannot catch this variation). The latter part, e.g. s_2, is estimated from the training set by the root mean square (across all training samples) of the difference between observed pH and the mean bootstrap pH in all bootstrap cycles when that sample is in the bootstrap test set. Whereas the first part of the error varies from fossil sample to fossil sample, the second part is constant. The estimated RMSE of prediction is the square root of the sum of squares of the two components (see Appendix 1). This procedure can also be applied to each training sample.

(*f*) *Computing*

Gaussian logit and linear logit regressions were done with GLIM (Payne 1986); WA regression and calibration, ML calibration, and associated statistics were implemented by WACALIB 2.1, a special purpose FORTRAN 77 program written by JML. It includes some subroutines from CANOCO (ter Braak 1988) written by CJF ter B. and M. O. Hill and an ML optimization subroutine written by CJF ter B. Detrended correspondence and canonical correspondence analyses were done by means of CANOCO 3.0 (C. J. F. ter Braak (unpublished)). Analogue analysis was implemented by the FORTRAN 77 program ANALOG 1.2 written by JML. Bootstrapping of WA estimates was done by using WACALIB 2.4, an update of WACALIB 2.1 incorporating bootstrapping subroutines written by CJF ter B. Bootstrapping of stratigraphic changes was done with THERRAD (Kitchell *et al.* 1987). All computations were done on an IBM PC/AT or compatible machines.

Results

(*a*) *Comparison of root mean squared error of prediction for different numerical procedures*

When ML, WA and WA(tol) were applied to the full training set, the RMSE (table 1) suggest that there is little to be gained in terms of RMSE by using ML. Moreover 14 of the 178 training samples (pH range 4.79–6.75) failed to converge with ML, i.e. for these samples our numerical

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

procedure was unable to find a single most likely pH value. Cross-validation with a training set of 130 and a test set of 48 lakes produced RMSE for the test set in the order WA < ML < WA(tol).

In a large, heterogeneous data set such as the 178-lake training set, it is possible that some samples are 'rogues' or atypical observations, for example with unusual diatom assemblages weakly related to pH, with poor or unreliable pH data, or with environmental variables other than pH having a major influence on the diatom composition. A data-screening exercise was performed to detect potential 'rogues'.

TABLE 1. ROOT MEAN SQUARED ERROR OF PREDICTION (RMSE) FOR WEIGHTED AVERAGING REGRESSION AND CALIBRATION WITH (WA(tol)) AND WITHOUT (WA) TOLERANCE-DOWNWEIGHTING AND MAXIMUM LIKELIHOOD (ML) REGRESSION AND CALIBRATION AND THE CORRELATION (r) BETWEEN OBSERVED AND PREDICTED pH BY USING DIFFERENT TRAINING SETS

training set		WA	WA(tol)	ML
178 lakes × 267 taxa (229 in ML)	RMSE	0.343	0.324	0.341[a]
	r	0.913	0.921	0.911
130 lakes × 240 taxa	RMSE	0.353	0.338	0.361[b]
	r	0.907	0.914	0.905
Test set 48 lakes × 240 taxa	RMSE	0.331	0.404	0.356
	r	0.915	0.849	0.894
172 lakes × 267 taxa (6 'rogues' deleted)	RMSE	0.323	0.301	—
	r	0.921	0.931	—
167 lakes × 267 taxa (225 taxa in ML) (11 'rogues' deleted)	RMSE	0.297	0.278	0.317[c]
	r	0.933	0.941	0.921

[a] Fourteen lakes failed to converge.
[b] Eleven lakes failed to converge.
[c] Ten lakes failed to converge.

In the first screening, a lake was deleted if it (i) formed an extreme outlier on any of the first four axes of a detrended correspondence analysis (Hill 1979) of the diatom data in the full training set; (ii) had a large (extreme 5%) residual distance to the pH-axis in a canonical correspondence analysis (ter Braak 1989) of the diatom data with pH as the only environmental variable; (iii) had a high (> 0.75 pH units) difference between observed and inferred pH in both WA and WA(tol) reconstructions. Six lakes fulfilled all three criteria. The RMSE for WA and WA(tol) for the training set with these six deleted showed some improvement (table 1). In a second screening of the reduced training set (172 lakes), a further five lakes now fulfilled two or three of the above criteria for deletion. When these five were also deleted, the RMSE for WA (0.297), WA(tol) (0.278), and ML (0.317) all showed further improvement (table 1). No lakes appeared to be obvious rogues within the reduced training set of 167 when screened again.

For pH reconstructions at RLGH and all other SWAP sites, this reduced training set of 167 lakes is used. This set has a pH range of 4.33–7.25 ($\bar{x} = 5.56$, median = 5.27, s.d. = 0.77) and 262 taxa.

A series of calibration experiments was done by using this training set to compare different regression and calibration procedures (table 2). Besides WA regression and calibration with and

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

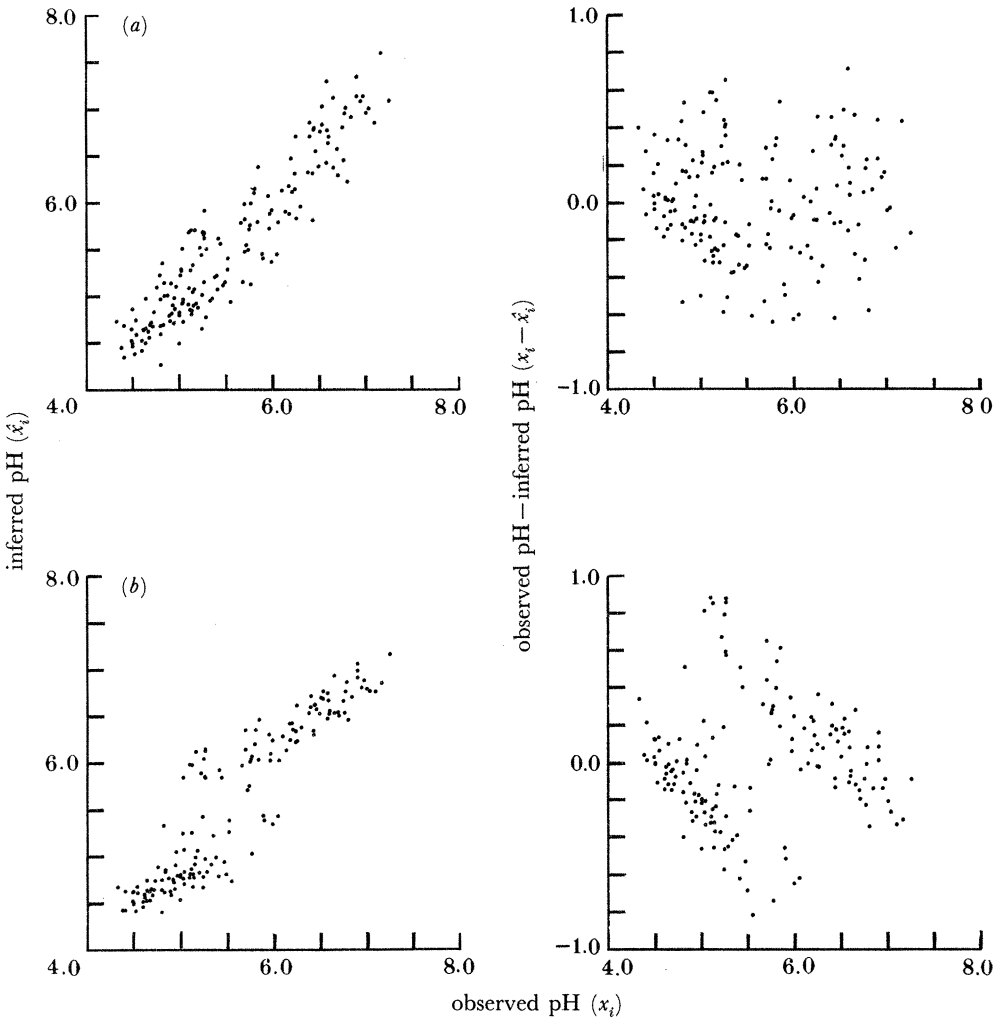

FIGURE 1. Plots of inferred pH ($\hat{x}_i$) against observed pH (x_i) (left hand plots) and of the differences ($x_i - \hat{x}_i$) against observed pH (right-hand plots) for pH inferences derived from (*a*) weighted averaging and (*b*) maximum likelihood regression and calibration.

TABLE 2. ROOT MEAN SQUARED ERROR OF PREDICTION (RMSE) AND CORRELATION (r) BETWEEN OBSERVED AND INFERRED pH WHEN DIFFERENT REGRESSION AND CALIBRATION PROCEDURES ARE USED WITH THE 167 LAKE TRAINING SET

regression procedure	calibration procedure	number of taxa	RMSE	r
WA	WA	274	0.297	0.993
WA	WA(tol)	274	0.278	0.941
ML	WA	224	0.294	0.934
ML	WA(tol)	224	0.282	0.939
ML	WA	168[a]	0.290	0.936
ML	WA(tol)	168[a]	0.292	0.935
ML	ML	225	0.317	0.921

[a] All taxa with statistically significant relation to pH (see table 4).

without tolerance downweighting and ML regression and calibration, 'hybrid' procedures were used with taxon optima and tolerances estimated by ML regression but by using WA (with and without tolerance downweighting) calibration (see, for example, Juggins (1988); Oksanen *et al.* (1988)) and all taxa and only those with a significant relation to pH. The overall conclusion is that ML regression and calibration perform the worst (cf. ter Braak & van Dam 1989), but that all the other procedures perform about equally well, as judged by RMSE. In terms of RMSE there seems little advantage in using ML regression compared with WA regression. Plots of inferred pH ($\hat{x}_i$) against observed pH (x_i) and of the differences ($x_i - \hat{x}_i$) against x_i for WA and ML are shown in figure 1. In WA there is no systematic bias (cf. Oksanen *et al.* 1988) except for a tendency for the differences to be greatest in the pH range 5.3–6.6, and lowest at the extremes. These high differences all occur in poorly buffered lakes where small seasonal changes in alkalinity can cause large variations in pH. In the absence of multiple pH measurements, estimation of a meaningful mean pH is impossible (Flower 1986). In ML two clusters emerge and pH estimates for ten lakes failed to converge, all in the pH range 4.79–5.89.

TABLE 3. ROOT MEAN SQUARED ERROR OF PREDICTION FOR TEN CROSS-VALIDATION EXPERIMENTS USING FOUR RANDOMLY SELECTED TEST SETS OF 50 LAKES, FOUR RANDOMLY SELECTED TEST SETS OF 40 LAKES, ONE RANDOMLY SELECTED TEST SET OF 67 LAKES AND ONE RANDOMLY SELECTED TEST SET OF 47 LAKES

experiment	number of taxa	WA	WA(tol)	WA	WA(tol)
		training set 117 lakes		test set 50 lakes	
1	231	0.310	0.282	0.287	0.339
2	226	0.292	0.276	0.310	0.368
3	228	0.307	0.294	0.309	0.383
4	231	0.313	0.291	0.269	0.305
mean		0.306	0.286	0.294	0.349
		training set 127 lakes		test set 40 lakes	
5	230	0.285	0.271	0.326	0.287
6	229	0.284	0.267	0.313	0.314
7	240	0.298	0.276	0.307	0.303
8	240	0.306	0.281	0.274	0.318
mean		0.299	0.280	0.299	0.327
		training set 100 lakes		test set 67 lakes	
9	248	0.257	0.240	0.338	0.541
		training set 120 lakes		test set 47 lakes	
10	258	0.283	0.267	0.300	0.288

In light of the RMSE results (table 2) and plots of x_i against $\hat{x}_i$ and of ($x_i - \hat{x}_i$) against x_i (not presented here) for these calibration experiments, WA regression and calibration with and without tolerance downweighting were selected for pH reconstructions at RLGH and all other SWAP lakes because of their low RMSE, their computational ease, and their robustness.

Ten cross-validation experiments were done to derive more reliable RMSE for WA and WA(tol), by using four randomly selected test sets of 50 lakes, four test sets of 40 lakes, a single test set of 67 lakes and one test set of 47 lakes (table 3). The RMSE for test sets in these experiments were 0.287–0.338 (mean = 0.308) for WA and 0.287–0.541 (mean = 0.376) for WA(tol). This indicates that WA has a lower prediction error in test sets than WA(tol). This error is, as ter Braak & van Dam (1989) emphasize, the 'appropriate benchmark to compare methods' because all errors are considered (see also Oksanen *et al.* (1988)).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

(*b*) *Comparison of the* pH *optima and tolerance estimates*

The basis for these calibration experiments is the estimates of the taxon parameters. The WA and ML estimated optima and tolerances for all taxa in the 167-lake training set are listed in table 4 (see microfiche), along with each taxon's maximum value, number of occurrences, shape of response curve, 95% confidence intervals of the ML-estimated optimum, standard error of the ML-estimated tolerance, curve height, literature pH category, and statistical relation to pH. Ecological discussion of these estimates will be presented in a subsequent SWAP publication. Of the 225 taxa (in six or more lakes) used for ML regression, 88 have unimodal curves with maxima, 78 have sigmoidal curves, 53 show no pattern, five have unimodal curves with minima and one failed to converge. A significant unimodal (88) or sigmoidal (78) response to pH is seen in 166 taxa. Of the 58 taxa with non-significant pH relations, only one has a fitted curve peak ($\hat{c}_k$) > 1% and only 15 have maximum values in the training set > 2.5% (range = 1.03–4.92%). As 74% of the taxa have a significant relation with pH, this provides strong confirmation for the assumption of Davis & Smol (1986) 'that there is a good statistical relation between pH and relative abundance of diatoms.' In the sampled pH range (4.33–7.25) WA estimates of optima are close to the ML estimates, although WA consistently but very slightly underestimates optima compared to ML. This slight bias is, in part, due to the over-representation of acid lakes in the training set, because WA estimates are sensitive to the distribution of x_i (ter Braak & Looman 1986, 1987). Major differences occur, however, at the extreme ends. At low pH, WA overestimates the optimum, whereas at high pH it underestimates it, because of truncation of the taxon response curves at the edges of the pH gradient. As a result of this inevitable truncation, WA compresses estimates of optima towards the centroid of the sampled pH gradient (Oksanen *et al.* 1988).

Estimates of tolerances by WA are almost all underestimated compared with ML, with a range of WA tolerances from *ca.* 0.2 to 0.75 pH units and an ML range of 0.2–1.8 pH units. For taxa with ML tolerances less than 0.75 pH units, WA provides reliable estimates, but at higher values WA systematically underestimates the tolerance (Oksanen *et al.* 1988). Many of the taxa with no significant relation to pH (table 4) have ML estimated tolerances of more than 1 pH unit, suggesting a range of occurrence over at least 4 pH units.

(*c*) *Reconstructions of* pH *for Round Loch of Glenhead*

The WA, WA(tol), and ML pH reconstructions for RLGH are shown in figure 2. The two WA reconstructions are closely parallel; 34 samples failed to converge in ML calibration.

Evaluation of reconstructed pH values

All quantitative palaeoenvironmental reconstruction procedures produce a result, but there is no simple means of evaluating how reliable it is. In addition to overall performance measures like RMSE calculated once from test sets, it is desirable to assess the reliability of individual reconstructed values. Two means towards this end are (i) measures of lack-of-fit of diatom assemblages to pH and (ii) estimated mean squared errors for each inferred pH.

(*a*) *Lack-of-fit to* pH *and analogue measures*

In the screening of the training data we used the squared residual distance of the modern samples to the pH axis in a canonical correspondence analysis of the diatom data as a criterion

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

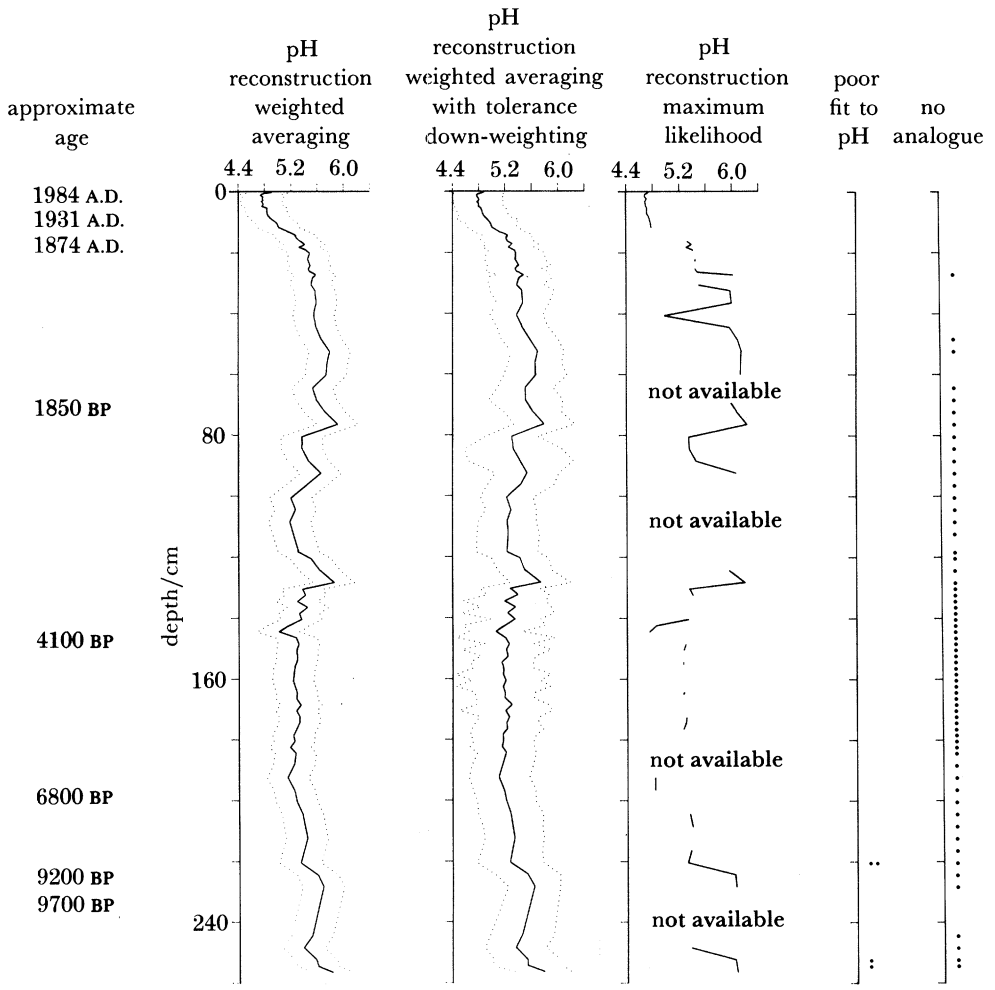

FIGURE 2. Reconstructed pH values for Round Loch of Glenhead plotted against depth (solid lines). The reconstructions are based on weighted averaging, tolerance-downweighted weighted averaging and maximum likelihood procedures. The RMSE of prediction for the weighted averaging estimates are shown as dotted lines. Samples with poor (•) or very poor (••) fit and those lacking close modern analogues in the SWAP training-set (•) are indicated.

of lack-of-fit to pH. Samples with a high residual distance from the pH axis have a poor fit to pH. Fossil, so-called passive, samples can also be positioned on this axis by means of transition formulae (ter Braak 1988). Any fossil sample whose residual distance is equal to or larger than the residual distance of the extreme 5% of the training set is considered to have a 'very poor' fit to pH, and those with values equal to or larger than the extreme 10% are deemed to have a 'poor' fit. One sample at RLGH (220.5 cm) has a very poor fit, and two (252.5, 254.5 cm) have poor fits (figure 2).

A reconstructed pH value is likely to be more reliable if the fossil sample in question has close modern analogues within the training set. Every fossil sample at RLGH was compared with all samples in the training set by using squared X^2 distance as a dissimilarity measure (Prentice 1980). This is:

$$d_{ij}^2 = \sum_{k=1}^{m} [(y_{ik} - y_{jk})^2 / (y_{ik} + y_{jk})],$$

where y_{ik} is the proportion of diatom taxon k in sample i and d_{ij} is the X^2 distance between samples i and j.

Any fossil sample that has a minimum $d_{ij} > 0.45$ appears to lack a *close* modern analogue in the training set (H. J. B. Birks, unpublished results). All RLGH samples below 48.5 cm fall in this category, as does the 27.3 cm sample (figure 2). This lack of modern analogues results, in part, from the abundance of *Melosira arentii* at RLGH, a taxon absent from the training set, and, in part, from the rarity of analogous, pristine, naturally acid but not acidified lakes in the SWAP study areas that could be sampled for the training set.

(*b*) *Estimated root mean squared error for* pH *reconstructions*

The RMSE of prediction for the training set, estimated by bootstrapping, are as follows:

	WA	WA(tol)
RMSE s_{i1}	0.072	0.305
s_2	0.312	0.371
RMSE prediction	0.320	0.480

The first error component is small for WA but large for WA(tol). The training set is thus clearly adequate to yield reliable estimates of the taxon parameters in the WA calibration formula ($\hat{u}_k$), but is probably not large enough for reliable estimation of the tolerances ($\hat{t}_k$) used in WA(tol).

The RMSE of prediction for the RLGH samples are plotted in figure 2 for WA and WA(tol). The RMSE for individual RLGH samples varies from 0.314 to 0.322 (WA) and 0.374 to 0.798 for WA(tol) and, for individual training samples, 0.314 to 0.376 for WA and 0.373 to 0.915 for WA(tol). These RMSE estimates indicate that there is no advantage in using tolerance-downweighting in WA calibration with this training set.

Discussion

(*a*) *Procedures for* pH *reconstruction*

Several conclusions emerge from our analyses (see also ter Braak & van Dam (1989)). Regression and calibration by WA can now replace earlier 'ad hoc' methods for pH reconstruction; WA is ecologically more realistic, statistically more robust, and numerically more accurate than other methods. Although some 'ad hoc' methods may produce lower apparent RMSE, these error estimates are not based on rigorous error-estimation procedures such as cross-validation or bootstrapping, but on regression statistics derived solely for training sets. As ter Braak & van Dam (1989) emphasize, RMSE based on training sets alone give an over-optimistic idea of prediction and performance error. In all our cross-validation experiments, RMSE for test sets is larger than for training sets (see also Juggins (1988); Oksanen *et al.* (1988)). In the SWAP training set the apparent RMSE for WA is 0.297 whereas the more realistic RMSE are 0.308 (cross-validation) and 0.320 (bootstrapping).

Although WA with tolerance-downweighting gives a lower apparent RMSE (0.278) than WA, tolerance-weighting does not improve RMSE in cross-validation (0.376) or bootstrapping (0.480). We therefore recommend simple WA regression and calibration with classical regression deshrinking as the easiest and most reliable pH reconstruction procedure currently available.

The overall RMSE is 0.320 for WA of the 167-lake training set. Standard errors of prediction for individual training samples and fossil samples at RLGH vary from 0.314 to 0.376. As lake

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

pH has inherent seasonal and annual variation and as there are errors in measuring pH, particularly at low ionic-strength water, further reductions in prediction errors seem unlikely. Compared with other pH training sets (reviewed by ter Braak & van Dam (1989)) this RMSE for a large ($n = 167$) training set is lower than is usually found. This is probably because the diatom assemblages are all from surficial sediments collected at or near the deepest part of the lake, many of the lakes have multiple pH determinations, the diatom taxonomy has been carefully harmonized as a result of SWAP taxonomic workshops (Munro *et al.*, this symposium) and the training set was screened by using three different numerical techniques; 11 'rogue' lakes were detected and deleted.

It is surprising that the theoretically more rigorous approach of ML regression and calibration did not perform as well in terms of RMSE as the simpler, approximating approach of WA. This is probably because ML uses more of the data, especially the absences (which WA ignores) and the precise percentage values. The problem of no-analogues and of assemblages containing diatoms of contrasting affinities are therefore more serious in ML than in WA. Although the RLGH samples had about the same amount of lack-of-fit as the modern samples, the lack-of-fit was of a different kind than in the modern samples, as the no-analogue measure showed. It is therefore unsafe to rely on the quantitative aspects of an assemblage as heavily as ML does. By its very nature, ML is more susceptible to 'rogues' than WA. Although we were able to eliminate 11 rogues from the original training set, the computational demands of recomputing taxon optima and tolerances for different screened versions of the training set prevented further screening of the data for samples that are possible 'rogues' in ML regression and calibration. The training set is thus critically screened only for WA.

Regression and calibration by WA have several advantages over more widely used pH inference techniques such as Index B (Renberg & Hellberg 1982) and multiple regression of pH categories (see, for example, Flower (1986); Charles (1985); Charles & Smol (1988)). Besides a lower RMSE (ter Braak & van Dam 1989), the main advantage of WA is that there is no need to assign diatom taxa to pH-preference categories. As Battarbee (1984) discusses and Holmes *et al.* (1989) clearly demonstrate, there are many problems in categorizing taxa, and the particular decision as to which category a taxon is assigned can markedly influence pH inferences by using Index B or multiple regression; WA is free of such problems. It uses the available data on the abundances of individual taxa in relation to pH in the training set. Moreover, because of the simple calculations involved in WA, it is possible to use bootstrapping to derive standard errors of predictions. Bootstrapping is, in theory, possible for ML, but, in practice, is computationally prohibitive. Individual standard errors of prediction for pH reconstructions from fossil assemblages can be valuable in avoiding misinterpretation of inferred pH values.

(*b*) *Reconstructions of* pH *at Round Loch of Glenhead*

The reconstructed pH history at RLGH (figure 2) shows little change from the late-glacial about 10000 years ago to about 4100 years before present (BP) (142 cm). Reconstructed pH varied from 5.4 to 5.8 in the late-glacial and earliest Holocene, but by 9200 BP (224 cm) it stabilized to about 5.2–5.4. Between 4100 and 1850 BP (72 cm) there were short-lived fluctuations, probably associated with inwashing of material from the catchment (Jones *et al.* 1989). Lake acidity changed little (5.3–5.7) until about 1870 A.D. when, between 17.3 cm (1874) and 7.3 cm (1931), pH dropped by over 0.5 units. Reconstructed pH values are never

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

below 5.0 until about 1900 (11 cm). Jones *et al.* (1989) conclude that this marked change in lake acidity resulted from an increase in deposition of strong acids from the atmosphere.

The null hypothesis that the rate of pH change per unit depth between 1874 and 1931 (17.3–7.3 cm) is no different from the rates of pH change in pre acid-deposition times (17.3–256–5 cm) was tested by using bootstrapping of the reconstructed pH time-series to generate empirical probability distributions of pH change with depth. The pH time-series was resampled randomly and with replacement 1000 times to create temporally-ordered data sequences of the same thickness as the interval of interest by using the time-duration or elapsed-time test of Kitchell *et al.* (1987). As the time-series contains unequal depth intervals between pH estimates, it is not possible for each bootstrapped time-series to contain exactly 10 cm. Instead samples are added to the time-series until the depth interval equals or exceeds the specific depth interval being tested. The observed rate of pH change at the time of increased acidic deposition is significantly different ($\alpha = 0.021$) from expectation. The null hypothesis is thus rejected, suggesting that the most rapid pH change per unit depth over the last 10000 years occurred between 1874 and 1931 at RLGH, the very time of increased acid deposition.

This research has been supported in part by SWAP, NAVF, and IBM (Norway). We are grateful to all SWAP diatomists for providing the training set, to Martin Munro for data-base management; to Rick Battarbee, John Boyle, Roger Flower and Viv Jones for valuable discussions; to Hilary Birks and John Kingston for commenting on earlier versions of the manuscript; to Sylvia Peglar and Siv Haugen for technical assistance; to Jennifer Kitchell and Norman MacLeod for providing THERRAD and to H. van der Voet for useful discussions about calibration and bootstrapping.

References

Barth, E. F. 1975 Average pH. *J. Wat. Pollut. Control Fed.* **47**, 2191–2192.

Battarbee, R. W. 1984 Diatom analysis and the acidification of lakes. *Phil. Trans. R. Soc. Lond.* B **305**, 451–477.

Battarbee, R. W. & Charles, D. F. 1987 The use of diatom assemblages in lake sediments as a means of assessing the timing, trends, and causes of lake acidification. *Prog. phys. Geog.* **11**, 552–580.

Battarbee, R. W., Smol, J. P. & Merliäinen, J. 1986 Diatoms as indicators of pH: an historical review. In *Diatoms and lake acidity* (ed. J. P. Smol, R. W. Battarbee, R. B. Davis & J. Meriläinen), pp. 5–14. Dordrecht: Dr W. Junk.

Birks, H. J. B. 1987 Methods for pH-calibration and reconstruction from palaeolimnological data: procedures, problems, potential techniques. Proceedings of the Surface Water Acidification Project (SWAP) mid-term review conference, Bergen 22–26 June 1987, pp. 370–380. London: SWAP.

Charles, D. F. 1985 Relationships between surface sediment diatom assemblages and lake water characteristics in Adirondack lakes. *Ecology* **66**, 994–1011.

Charles, D. F. & Smol, J. P. 1988 New methods for using diatoms and chrysophytes to infer past pH of low-alkalinity lakes. *Limnol. Oceanogr.* **33**, 1451–1462.

Davis, R. B. & Anderson, D. S. 1985 Methods of pH calibration of sedimentary diatom remains for reconstructing history of pH in lakes. *Hydrobiologia* **120**, 69–87.

Davis, R. B. & Smol, J. P. 1986 The use of sedimentary remains of siliceous algae for inferring past chemistry of lake water-problems, potential and research needs. In *Diatoms and lake acidity* (ed. J. P. Smol, R. W. Battarbee, R. B. Davis & J. Meriläinen), pp. 291–300. Dordrecht: Dr W. Junk.

Diaconis, P. & Efron, B. 1983 Computer-intensive methods in statistics. *Scient. Am.* **248**(5), 96–109.

Efron, B. 1982 The jackknife, the bootstrap, and other resampling plans. *SIAM NSF-CBMS Monograph* **38**, 1–92.

Efron, B. 1983 Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. statist. Ass.* **78**, 316–331.

Efron, B. & Gong, G. 1983 A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Statist.* **37**, 36–48.

Flower, R. J. 1986 The relationship between surface sediment diatom assemblages and pH in 33 Galloway lakes: some regression models for reconstructing pH and their application to sediment cores. *Hydrobiologia* **143**, 93–103.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Gallant, A. R. 1975 Nonlinear regression. *Am. Statist.* **29**, 73–81.
Hill, M. O. 1979 *DECORANA – a FORTRAN program for detrended correspondence analysis and reciprocal averaging.* Cornell University, Ithaca, New York: Section of ecology and systematics.
Holmes, R. W., Whiting, M. C. & Stoddard, J. L. 1989 Changes in diatom-inferred pH and acid neutralizing capacity in a dilute, high elevation, Sierra Nevada lake since A.D. 1825. *Freshwater Biol.* **21**, 295–310.
Imbrie, J. & Webb, T. III 1981 Transfer functions: calibrating micropaleontological data in climatic terms. In *Climatic variations and variability: facts and theories* (ed. A. Berger), pp. 125–134. Dordrecht: D. Reidel.
Jones, V. J., Stevenson, A. C. & Battarbee, R. W. 1989 Acidification of lakes in Galloway, South West Scotland: a diatom and pollen study of the post-glacial history of the Round Loch of Glenhead. *J. Ecol.* **77**, 1–23.
Juggins, S. 1988 A diatom/salinity transfer function for the Thames Estuary and its application to waterfront archaeology. Ph.D. thesis, University of London.
Kitchell, J. A., Estabrook, G. & MacLeod, N. 1987 Testing for equality of rates of evolution. *Paleobiology* **13**, 272–285.
Lwin, T. & Maritz, J. S. 1982 An analysis of the linear-calibration controversy from the perspective of compound estimation. *Technometrics* **24**, 235–242. (Minor correction in *Technometrics* **27**, 445.)
Martinelle, S. 1970 On the choice of regression in linear calibration. Comments on a paper by R. G. Krutchkoff. *Technometrics* **12**, 157–161.
Middleton, A. C. & Rovers, F. A. 1976 Average pH. *J. Wat. Pollut. Control Fed.* **48**, 395–396.
Oksanen, J., Läära, E., Huttunen, P. & Meriläinen, J. 1988 Estimation of pH optima and tolerances of diatoms in lake sediments by the methods of weighted averaging, least squares and maximum likelihood, and their use for the prediction of lake acidity. *J. Paleolimnol.* **1**, 39–49.
Payne, C. D. (ed.) 1986 *The GLIM System Release 3.77.* Oxford: Numerical Algorithms Group.
Picard, R. R. & Cook, R. D. 1984 Cross-validation of regression models. *J. Am. Statist. Ass.* **79**, 575–583.
Prentice, I. C. 1980 Multidimensional scaling as a research tool in Quaternary palynology: a review of theory and methods. *Rev. Palaeobot. Palynol.* **31**, 71–104.
Renberg, I. & Hellberg, T. 1982 The pH history of lakes in southwestern Sweden, as calculated from the subfossil diatom flora of the sediments. *Ambio* **11**, 30–33.
Snee, R. D. 1977 Validation of regression models: methods and examples. *Technometrics* **19**, 415–428.
Stone, M. 1974 Cross-validatory choice and assessment of statistical predictions (with discussion). *Jl R. statist. Soc.* B **36**, 111–147.
ter Braak, C. J. F. 1985 Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics* **41**, 859–873.
ter Braak, C. J. F. 1987*a* Calibration. In *Data analysis in community and landscape ecology* (ed. R. H. G. Jongman, C. J. F. ter Braak & O. F. R. van Tongeren), pp. 78–90. Wageningen: Pudoc.
ter Braak, C. J. F. 1987*b* *Unimodal models to relate species to environment.* Doctoral thesis, University of Wageningen.
ter Braak, C. J. F. 1988 CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). Technical Report LWA-88-02, GLW, Wageningen, 95 pp.
ter Braak, C. J. F. & Barendregt, L. G. 1986 Weighted averaging of species indicator values: its efficiency in environmental calibration. *Math. Biosci.* **78**, 57–72.
ter Braak, C. J. F. & Looman, C. W. N. 1986 Weighed averaging logistic regression and the Gaussian response model. *Vegetatio* **65**, 3–11.
ter Braak, C. J. F. & Looman, C. W. N. 1987 Regression. In *Data analysis in community and landscape ecology* (ed. R. H. G. Jongman, C. J. F. ter Braak & O. F. R. van Tongeren), pp. 29–77. Wageningen: Pudoc.
ter Braak, C. J. F. & Prentice, I. C. 1988 A theory of gradient analysis. *Adv. ecol. Res.* **18**, 271–317.
ter Braak, C. J. F. & van Dam, H. 1989 Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* **178**, 209–223.
Wallach, D. & Goffinet, B. 1989 Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecol. Modelling* **44**, 299–306.

Appendix 1

Bootstrap estimation of sample-specific mean-squared error for pH *reconstructions by weighted averaging*

The notation and bootstrap procedure are described in the main text. In addition, $\mathrm{AVE}(\hat{x}_{i,\mathrm{boot}})$ and $\mathrm{MS}(x_i - \hat{x}_{i,\mathrm{boot}})$ denote the mean and mean square, respectively, of the argument across all the bootstrap cycles where sample i does not belong to the bootstrap training set.

The mean-squared error of the inferred pH of training sample i is estimated by

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

$\mathrm{MS}(x_i - \hat{x}_{i,\mathrm{boot}})$. This estimator has some importance; (i) it does not suffer from resubstitution bias and (ii) its mean across the training samples yields the bootstrap estimate of overall mean-squared error of prediction. But the corresponding formula cannot be calculated for individual fossil samples, simply because the observed value (x_i) is not available. To obtain a sample-specific error estimator we use the following decomposition:

$$\mathrm{MS}(x_i - \hat{x}_{i,\mathrm{boot}}) = \mathrm{MS}(\hat{x}_{i,\mathrm{boot}} - \mathrm{AVE}(\hat{x}_{i,\mathrm{boot}})) + (x_i - \mathrm{AVE}(\hat{x}_{i,\mathrm{boot}}))^2,$$

which we write in shorthand as:

$$v_i = v_{i1} + v_{i2}.$$

The first part, v_{i1}, can be calculated from the bootstrap cycles for each sample, both fossil and training samples. It represents the effects that the variability of the taxon parameters in the calibration function have on the inferred pH for sample i. It reduces in magnitude as the size of the training set increases. But it does so in a sample-specific way. This error component is likely to be relatively small for fossil assemblages consisting of taxa that are frequent and abundant in the training set and to be relatively large for assemblages consisting of taxa that are infrequent and rare in the training set.

The second part, v_{i2}, can be calculated for the training samples only. It includes the error caused by imperfections in the calibration function, even if the parameters are known without error. Diatom assemblages vary even among lakes with the same pH or, conversely, because lakes with the same diatom assemblage may differ in pH. Model specification error also enters v_{i2}. By using multiple regression we investigated whether v_{i2} depends, in a systematic way, on pH, the number of taxa, and the inhomogeneity of an assemblage. For pH we used a second-order polynomial in $\mathrm{AVE}(\hat{x}_{i,\mathrm{boot}})$ and for inhomogeneity the variance ('tolerance') of the optima of the taxa present in the assemblage (Hill 1979). These predictors are suggested from the theory of linear (Martinelle 1970) and WA (ter Braak & Barendregt 1986) calibration. The predictors, however, explained less than 10% of the variance of v_{i2} in the training set. Transformation of the variables (except pH) to logarithms did not improve the fit. Apparently, the second error component is mainly due to other factors. For fossil samples it was therefore taken as a constant, namely the mean v_{i2} across the training set.

The above derivation ignores terms of order $1/n_{\mathrm{boot}}$ with n_{boot} being the number of bootstrap cycles. These terms are negligible with our choice of $n_{\mathrm{boot}} = 1000$.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES