PALEOCEANOGRAPHY, VOL. 19, PA4014, doi:10.1029/2004PA001072, 2004

Biases in the estimation of transfer function prediction errors

R. J. Telford,[1] C. Andersson,[1] H. J. B. Birks,[1,2,3] and S. Juggins[4]

Received 8 July 2004; revised 27 August 2004; accepted 1 September 2004; published 10 November 2004.

[1] In the quest for more precise sea-surface temperature reconstructions from microfossil assemblages, large modern training sets and new transfer function methods have been developed. Realistic estimates of the predictive power of a transfer function can only be calculated from an independent test set. If the test set is not fully independent, the error estimate will be artificially low. We show that the modern analogue technique using a similarity index (SIMMAX) and the revised analogue method (RAM), both derived from the modern analogue technique, achieve apparently lower root mean square error of prediction (RMSEP) by failing to ensure statistical independence of samples during cross validation. We also show that when cross validation is used to select the best artificial neural network or modern analogue model, the RMSEP based on cross validation is lower than that for a fully independent test set. *INDEX TERMS:* 3030 Marine Geology and Geophysics: Micropaleontology; 4267 Oceanography: General: Paleoceanography; 4294 Oceanography: General: Instruments and techniques; *KEYWORDS:* transfer functions, quantitative paleoenvironmental reconstructions

Citation: Telford, R. J., C. Andersson, H. J. B. Birks, and S. Juggins (2004), Biases in the estimation of transfer function prediction errors, *Paleoceanography*, *19*, PA4014, doi:10.1029/2004PA001072.

1. Introduction

[2] Quantitative sea-surface temperature (SST) estimates calculated from foraminiferal, and other microfossil, assemblages are widely used to provide insights into Quaternary environmental changes [e.g., *CLIMAP Project Members*, 1984; *Pflaumann et al.*, 2003]. Understandably, palaeoecologists want the most precise SST reconstructions possible, and since *Imbrie and Kipp* [1971] first introduced quantitative transfer functions, new methods and larger modern training sets have apparently substantially increased precision (Table 1).

[3] The Imbrie and Kipp approach of factor analysis and associated multiple regression is based on an underlying linear species-environment model, although the initial data normalization used in the method seems able to cope with nonlinear relationships [*ter Braak*, 1995]. Since many foraminifera species show a nonlinear unimodal response to temperature, it is not surprising that transfer function methods based on an explicit unimodal species-environment response (such as correspondence analysis regression [*Roux*, 1979] and weighted-averaging partial least squares [*ter Braak and Juggins*, 1993]) outperform Imbrie and Kipp factor analysis [*Birks*, 1995]. Still greater reductions in RMSEP have resulted from the use of transfer function techniques that lack any implicit species-environment response model such as the modern analogue technique (MAT) [*Prell*, 1985] and artificial neural networks (ANN) [*Malmgren et al.*, 2001]. The conceptual simplicity of MAT has encouraged the development of derivatives, including SIMMAX [*Pflaumann et al.*, 1996], which geographically weights analogues, and the revised analogue method (RAM) [*Waelbroeck et al.*, 1998], which allows the selection of analogues from a response surface, and has special analogue selection rules: both SIMMAX and RAM report a substantially lower prediction error than simple MAT.

[4] Transfer function model choice is most commonly guided by training set performance statistics, especially the root mean square error (RMSE: the square root of the mean of the squared differences between observed and predicted SST). As the true RMSE is invariably under-estimated when based solely on the training set [*Birks*, 1995], some form of cross validation with an independent test set is required to derive a more reliable and realistic estimate of prediction error (RMSEP) and hence to evaluate the predictive abilities of the transfer function model [*Birks et al.*, 1990]. If many transfer function models are produced, with different settings, and RMSEP is used to select the best model, then this RMSEP is not independent of model choice [*ter Braak*, 1995], and is a biased estimate. In this paper we assess the magnitude of this bias by calculating the RMSEP for both the data set used to select the best model (the optimization set) and a fully independent test set. We also test the SIMMAX and RAM methods to discover if the reported reductions in RMSEP are real.

2. Methods and Data Sets

[5] All analyses were performed on the 947 sample Atlantic foraminiferal database (ATL947; *Pflaumann et al.* [2003]) and we use winter SSTs. The data set was split, at random, into three parts: a modeling (or training) set of 747 samples; a 100 sample optimization (or selection set), used to select the model with the lowest RMSEP for methods

[1] Bjerknes Centre for Climate Research, Bergen, Norway.
[2] Also at Department of Biology, University of Bergen, Bergen, Norway.
[3] Also at Environmental Change Research Centre, University College London, London, UK.
[4] School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne, UK.

0883-8305/04/2004PA001072

Table 1. Number of Samples in Selected Atlantic Foraminiferal Training Sets, Transfer Function Method Used, and RMSEP of Winter Season SST

Source	Method	Number of Samples	RMSEP, °C
Imbrie and Kipp [1971]	I and K	61	2.26[a]
Prell [1985]	MAT	356	1.02
Pflaumann et al. [1996]	SIMMAX	738	0.9
Waelbroeck et al. [1998]	RAM	615	0.7
Pflaumann et al. [2003]	SIMMAX	947	0.75

[a]*Imbrie and Kipp* [1971] factor analysis cross-validated using C2 [*Juggins*, 2003].

generating multiple models; and a 100 sample-independent test set, for which the RMSEP was calculated using the model selected by the optimization set. This procedure was repeated 100 times to generate 100 estimates of the RMSEP. RMSEPs were also calculated for the modeling set with leave-one-out cross validation with RAM and MAT with the predetermined options listed below.

[6] MAT was calculated, using functions written in the R statistical package [*R Development Core Team*, 2004], with between 1 and 20 analogues, or a jump threshold of between 0.02 and 1 (see below): these options cover the range of values likely to be chosen. RAM [*Waelbroeck et al.*, 1998] was calculated using winRAM, with a spacing and selection radius of 0.2°C, a maximum of ten analogues, and a jump threshold of 0.1: settings similar to those used by *Waelbroeck et al.* [1998] and *Malmgren et al.* [2001]. RAM was calculated in both 2 dimensions (summer and winter SST) and 1 dimension (by setting both axes of the grid to winter temperature). SIMMAX [*Pflaumann et al.*, 1996] was re-implemented in R, and calculated with 10 analogues with and without geographic distance weighting. ANN was calculated using the nnet library for R [*Venables and Ripley*, 2002], with a single hidden layer of between 5 and 25 neurons (in increments of 5). The significance of differences in RMSEP between methods were tested with Wilcoxon tests.

3. Results and Discussion

3.1. Model Choice and Independence

3.1.1. Modern Analogue Technique

[7] The modern analogue technique (MAT), perhaps the most widely used transfer function for reconstructing SSTs, is based on the premise that faunal assemblages that resemble one another are derived from similar environments [*Prell*, 1985]. This is quantified by selecting the k-nearest neighbors in the modeling set, using an appropriate distance or dissimilarity metric, and calculating the mean (or a dissimilarity-weighted mean) of the environmental parameter of interest. There are different criteria for choosing k. We consider two techniques here. The first is to use the same value of k for each sample, with k chosen to minimize the RMSEP in the optimization set. The second, referred to here as threshold MAT (T.MAT), is to sort the ten best analogues according to their dissimilarity distance from the test sample and compare each value with the previous one [*Waelbroeck et al.*, 1998; *Sawada et al.*, 2004]. If the proportional increase in distance is larger than a threshold, the preceding analogues are retained, and the remainder discarded. The threshold can be chosen to minimize the RMSEP of the optimization set.

[8] The median number of analogues selected by the optimization set for MAT was five. In only 16% of trials were ten or more analogues selected. The median threshold chosen by T.MAT was 0.17. Figure 1 shows that optimization set RMSEPs are significantly lower than the independent test set RMSEPs for both methods of choosing k (median RMSEP 0.94 versus 1.02 for MAT ($p < 0.0001$); 0.94 versus 1.00 for T.MAT ($p < 0.001$)). There is no significant difference between the optimization set RMSEP for the different methods; the threshold method has a slightly, but not significantly, lower independent test set RMSEP.

3.1.2. Artificial Neural Networks

[9] Artificial neural networks are algorithms that, by mimicking biological neural networks, have the ability to learn by example. They learn by iteratively adjusting a large set of parameters, which are initially set at random values, to minimize the error between the predicted and actual output. They can approximate any continuous function [*Hornik et al.*, 1989] and provide a flexible way to generalize a linear regression function [*Venables and Ripley*, 2002]. If trained for too long, ANNs can over-fit the data,

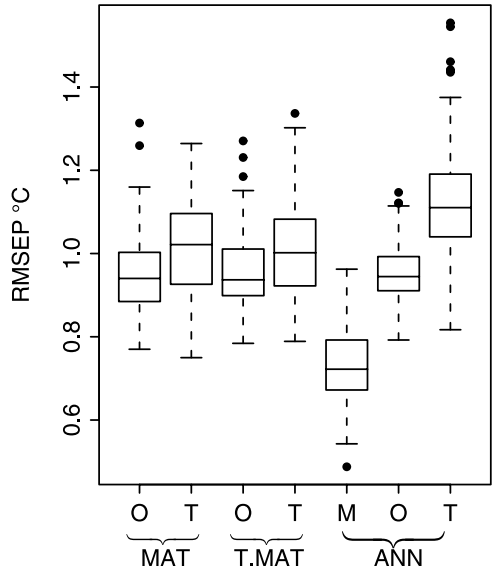

Figure 1. Box plots of RMSEP (100 trials) of modeling (M), optimization (O) and test (T) set calculated using MAT, MAT with a jump threshold (T.MAT), and artificial neural networks (ANN) from the ATL947 data set.

learning particular features of the modeling set rather than the general rules. This is normally controlled by using a second data set and stopping the training when the model stops reducing the RMSEP of this data set. Typically many ANN models are generated from different random initial conditions and configurations and the best model used. ANNs have proved useful for problems where the underlying structure is not well understood. *Malmgren et al.* [2001] apply neural networks to reconstruct SSTs and report that the ANN-model RMSEPs are not significantly lower than MAT.

[10] Each network configuration (number of hidden layer neurons) was selected by approximately the same number of trials, with a slight tendency toward selecting 15 hidden layer neurones. As expected, the RMSE of the modeling set (Figure 1) is much lower than the RMSEP of the optimization set (median RMSEP 0.72 versus 0.94°C). The RMSEP of the independent test set is larger again by almost the same amount (median RMSEP 1.11°C). While the optimization set RMSEPs are similar for MAT and ANN (median RMSEP is 0.94°C for both), the independent test set RMSEPs are significantly worse for ANN (median RMSEP 1.02 versus 1.11°C; $p < 0.0001$). This presumably reflects the greater potential for ANN to be tuned to the optimization set.

[11] The modeling set used here is two hundred samples smaller than the original data set. This will cause a reduction in the efficiency of the transfer function and larger expected RMSEP. This effect could be minimized by using smaller optimization and test sets, but the variability of the RMSEP of these sets will increase as they become smaller, so more trials would be needed for a reliable estimate of the median RMSEP. In a large data set, like ATL947, the bias caused by the reduced size of the modeling set is small, but it would be more serious in small data sets. If the MAT analogue selection criteria are set a priori, there is no need for an optimization set, so more samples are available for the modeling set. The ANN parameters cannot be set a priori, so three data partitions should always be used.

3.2. Performance of MAT Variants

3.2.1. SIMMAX

[12] SIMMAX (modern analogue technique using a similarity index; *Pflaumann et al.* [1996]) differs from MAT in two important ways. First, the best analogues are selected using the scalar product of the standardized faunal percentages as the distance metric, rather than the squared-chord distance recommended by *Overpeck et al.* [1985] and *Gavin et al.* [2003]. Second, the modern analogues are weighted according to the reciprocal of the geographical distance to the unknown sample. The RMSEP of winter SST, based on 10 analogues, using the ATL947 foram data set [*Pflaumann et al.*, 2003], is 0.75°C with geographical distance weighting, and 1.29°C without distance weighting [*Pflaumann et al.*, 2003].

[13] The poor performance of geographically unweighted-SIMMAX relative to MAT (Figure 2) suggests that the scalar product distance metric is less well suited to foraminiferal assemblages than the squared-chord distance metric. Geographically weighted-SIMMAX has a much lower prediction error than either the unweighted variant or MAT (Figure 2). However, there are concerns over the use of the geographical distance weighting [e.g., *Trend-Staid and Prell*, 2002], given its potential to bias results by weighting most heavily the modern analogue closest to the core site. This will bias the reconstruction toward the modern value of the site. This problem is most extreme when there is an analogue very close to the core site, the high weighting given to this sample can mean that the reconstructed temperature does not deviate at all down-core [*Malmgren et al.*, 2001].

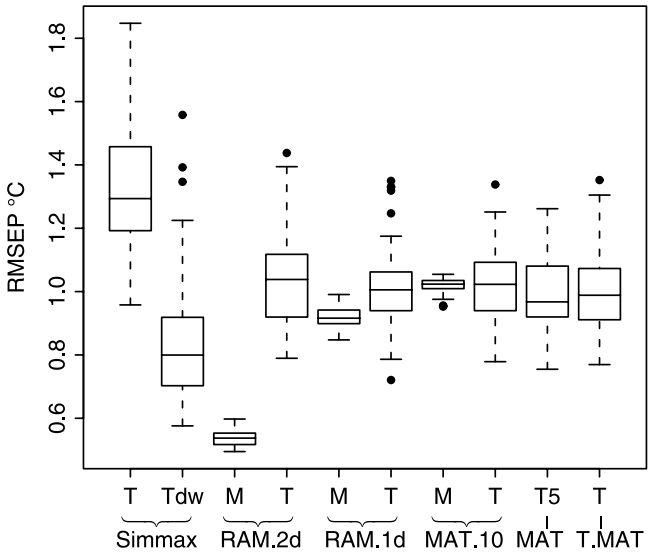

Figure 2. Box plots of RMSEP (100 trials) for SIMMAX without (T) and with geographic distance weighting (Tdw) for the independent test set; modeling (M) and test (T) set RMSEPs for RAM based on two and one environmental dimensions and MAT with 10 analogues; and test set RMSEPs for MAT with five analogues (T5) and MAT with the same analogue selection criteria as RAM (T.MAT).

[14] Ocean surface temperatures are highly autocorrelated: SIMMAX uses this structure to reduce prediction error. This removes the statistical independence [*Legendre*, 1993] between the sample being tested and the remainder of the data set during cross validation. Because of this the RMSEP, even under our more rigorous data splitting, will be artificially low and will not be a reliable indication of the prediction error. Even if foraminiferal assemblages were identical over the whole ocean, a geographical distance weighted approach would still yield a low RMSEP: this affect can be seen by the improvements in estimation at the cold end of the gradient [*Pflaumann et al.*, 2003], where assemblages are all dominated by *Neogloboquadrina pachyderma* (sin.).

[15] The problem of MAT choosing geographically inappropriate analogues, which SIMMAX attempts to fix [*Pflaumann et al.*, 2003], has been addressed by *Trend-Staid and Prell* [2002] by limiting the geographical range from which analogues can be chosen. This approach, although ad hoc, is probably to be preferred. The problem of multiple analogues, where similar assemblages can occur in different temperature waters in different geographical regions, can

also be addressed within a Bayesian framework [e.g., *ter Braak et al.*, 1996], or by accepting increased uncertainty in reconstructions.

3.2.2. RAM

[16] RAM (revised analogue method; *Waelbroeck et al.* [1998]) has two important differences from MAT. First, the data set of possible analogues is expanded by mapping the original data onto a grid of summer and winter SSTs. At each grid node an artificial assemblage is constructed from the samples within a specified radius, weighted by the reciprocal of their Euclidean distance from the grid point. The rationale behind this is that more appropriate analogues can be chosen from the larger, more homogeneous data set [*Waelbroeck et al.*, 1998]. Second, RAM attempts to find the optimum number of analogues for each site by searching for jumps in the distance to the next analogue, with a minimum of two and a maximum of ten analogues. *Waelbroeck et al.* [1998] report that RAM gives a substantially lower RMSEP than MAT (0.7°C versus 1.13°C), but *Malmgren et al.* [2001] find only a slight improvement.

[17] RMSEPs calculated for the modeling set (Figure 2) with RAM, using two environmental dimensions, are much lower than those for MAT with 10 analogues. The reduction in RMSEP is less marked with one environmental dimension, but still highly significant ($p < 0.0001$). However, the independent test set RMSEPs calculated with two-dimensional RAM, are no smaller than for MAT. The one-dimensional RAM test set RMSEPs are lower than those for MAT, but higher, though not significantly, than those for T.MAT, the variant of MAT calculated with the same analogue selection criteria as RAM. The difference between the performance of RAM under leave-one-out cross validation (the method used by winRAM) and with an independent test set indicates that the low RMSEP reported by *Waelbroeck et al.* [1998] is erroneous, and that the grid of artificial samples generates no improvement in one dimension, and is detrimental in two. RAM's analogue selection rules reduce the average number of analogues used. MAT with five analogues rather than the usual default of ten slightly outperforms RAM and T.MAT.

[18] Remapping the data onto a grid is a concept similar to the smooth response surfaces used to investigate pollen-climate relationships [*Bartlein et al.*, 1986; *ter Braak*, 1995]. However, RAM does not produce smooth grids. There are two reasons for this. First, by mapping the data over two environmental variables, RAM suffers from the "curse of dimensionality," and a large number of grid points are derived from few samples. Of 1144 grid points generated using the entire ATL947 data set, only 4% are derived from more than five samples, and over 50% consist of single samples. This problem is less significant if only one environmental dimension is used. Second, the reciprocal Euclidean distance weighting does not yield a smooth curve. If a grid point coincides with a sample, that sample has an arbitrarily high weighting; and a sample 0.01°C away has twice the weighting of a sample 0.02°C even though the inherent uncertainty in the temperature estimates is probably higher than this. Together, these problems mean that rather than producing a smooth grid, the grid is very noisy, and many points actually, or effectively, derive from single samples. Generating the smooth grid is computer intensive, and RAM calculates it just once. The leave-one-out cross validation in winRAM is not therefore an independent test as the sample being tested can find grid samples that it has contributed to. The resulting RMSEPs are artificially low: this effect disappears when RAM is tested using a completely independent test set. Under true cross validation, the affect of the grid is to give some modeling set samples much more weight than others.

[19] RAM is a novel hybrid between classical and inverse approaches to calibration [see *ter Braak*, 1995], and the implementation problems above could be addressed. If it had the advantages of both approaches, it could offer improved reconstructions. If a finely spaced grid is used, the large number of interpolated smooth grid points will dominate the reconstruction, at least for some samples, and reconstructions will tend to resemble the response surface; conversely a coarse grid will have relatively little influence and the reconstruction will resemble a MAT reconstruction. The contributions can only be balanced if two reconstructions are made, one MAT and one with response surfaces, and a consensus reconstruction calculated [*Bartlein and Whitlock*, 1993].

4. Conclusions

[20] Realistic and robust estimates of the predictive power of a transfer function are only possible when evaluated with an independent test set. Independence can be compromised in many ways to give artificially low prediction errors. When a transfer function method generates multiple models, the RMSEP of the data set used to select the best model is a biased estimate. An independent test set should be used to get a more robust estimate of the true uncertainty.

[21] ANN does not out perform MAT when independent test sets are used. Given the computational demands of ANN, the difficulty of training, the susceptibility to overfitting, and the lack of interpretability of the resulting network, it cannot be recommended for routine use.

[22] SIMMAX uses a poorly performing distance metric, and uses geographic information to achieve low RMSEPs. This reduces the independence of the test set as SSTs are autocorrelated, so SIMMAX will give artificially low RMSEPs and reconstructions are biased toward the modern temperature.

[23] RAM achieves apparently low RMSEPs as a result of its flawed cross validation procedure. It fails to outperform MAT with an independent test set. The low RMSEP reported by both RAM and SIMMAX should therefore be treated with caution.

[24] **Acknowledgments.** This research was supported by NORPAST-2, funded by the Norwegian Research Council, and PACLIVA, an EU Framework 5 Programme project (EVK2-CT2002-000143). This is publication nr. A64 from the Bjerknes Centre for Climate Research. We thank two anonymous reviewers for their comments on an earlier version of this manuscript.

References

Bartlein, P. J., and C. Whitlock (1993), Paleoclimatic interpretation of the Elk Lake pollen record, *Geol. Soc. Am. Spec. Pap.*, *276*, 275–293.

Bartlein, P. J., I. C. Prentice, and T. Webb (1986), Climatic response surfaces from pollen data for some eastern North American taxa, *J. Biogeogr.*, *13*, 35–57.

Birks, H. J. B. (1995), Quantitative palaeoenvironmental reconstructions, in *Statistical Modelling of Quaternary Science Data*, *Tech. Guide 5*, edited by D. Maddy and J. S. Brew, pp. 116–254, Quat. Res. Assoc., Cambridge, UK.

Birks, H. J. B., J. M. Line, S. Juggins, A. C. Stevenson, and C. J. F. ter Braak (1990), Diatoms and pH reconstructions, *Philos. Trans. R. Soc. London B*, *327*, 263–278.

CLIMAP Project Members (1984), The last interglacial ocean, *Quat. Res.*, *21*, 123–244.

Gavin, D. G., W. W. Oswald, E. R. Wahl, and J. W. Williams (2003), A statistical approach to evaluating distance metrics and analog assignments for pollen records, *Quat. Res.*, *60*, 356–367.

Hornik, K., M. Stinchcombe, and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, *2*, 359–366.

Imbrie, J., and N. G. Kipp (1971), A new micropaleontological method for quantitative paleoclimatology: Application to a late Pleistocene Caribbean core, in *The Late Cenozoic Glacial Ages*, edited by K. K. Turekian, pp. 71–181, Yale Univ. Press, New Haven, Conn.

Juggins, S. (2003), C2 user guide: Software for ecological and palaeoecological data analysis and visualisation, 69 pp., Univ. of Newcastle, Newcastle upon Tyne, UK.

Legendre, P. (1993), Spatial autocorrelation: Trouble or new paradigm?, *Ecology*, *74*, 1659–1673.

Malmgren, B. A., M. Kucera, J. Nyberg, and C. Waelbroeck (2001), Comparison of statistical and artificial neural network techniques for estimating past sea surface temperatures from planktonic foraminifer census data, *Paleoceanography*, *16*, 520–530.

Overpeck, J. T., T. Webb III, and I. C. Prentice (1985), Quantitative interpretation of fossil pollen spectra: Dissimilarity coefficients and the method of modern analogs, *Quat. Res.*, *23*, 87–108.

Pflaumann, U., J. Duprat, C. Pujol, and L. D. Labeyrie (1996), SIMMAX: A modern analog technique to deduce Atlantic sea surface temperatures from planktonic foraminifera in deep-sea sediments, *Paleoceanography*, *11*, 15–35.

Pflaumann, U., et al. (2003), Glacial North Atlantic: Sea-surface conditions reconstructed by GLAMAP 2000, *Paleoceanography*, *18*(3), 1065, doi:10.1029/2002PA000774.

Prell, W. L. (1985), The stability of low-latitude sea-surface temperatures: An evaluation of the CLIMAP reconstruction with emphasis on the positive SST anomalies, 60 pp., Dept. of Energy, Washington, D. C.

R Development Core Team (2004), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna.

Roux, M. (1979), Estimation des paléoclimats d'apres l'ecologie des foraminifers, *Cah. Anal. Données*, *4*, 61–79.

Sawada, M., A. E. Viau, G. Vettoretti, W. R. Peltier, and K. Gajewski (2004), Comparison of North-American pollen-based temperature and global lake-status with CCCma AGCM2 output at 6 ka, *Quat. Sci. Rev.*, *23*, 225–244.

ter Braak, C. J. F. (1995), Non-linear methods for multivariate statistical calibration and their use in paleoecology—A comparison of inverse (*k*-nearest neighbours, partial least-squares and weighted averaging partial least-squares) and classical approaches, *Chem. Intell. Lab. Syst.*, *28*, 165–180.

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

ter Braak, C. J. F., H. van Dobben, and G. di Bella (1996), On inferring past environmental change from species composition data by non-linear reduced-rank models, paper presented at XVIIIth International Biometric Conference, Int. Biometric Soc., Amsterdam, 1 –5 July.

Trend-Staid, M., and W. L. Prell (2002), Sea surface temperature at the Last Glacial Maximum: A reconstruction using the modern analog technique, *Paleoceanography*, *17*(4), 1065, doi:10.1029/2000PA000506.

Venables, W. N., and B. D. Ripley (2002), *Modern Applied Statistics with S*, 4th ed., 495 pp., Springer-Verlag, New York.

Waelbroeck, C., L. Labeyrie, J. C. Duplessy, J. Guiot, M. Labracherie, H. Leclaire, and J. Duprat (1998), Improving past sea surface temperature estimates based on planktonic fossil faunas, *Paleoceanography*, *13*, 272–283.

C. Andersson and R. J. Telford, Bjerknes Centre for Climate Research, Allégaten 55, N-5007 Bergen, Norway. (richard.telford@bjerknes.uib.no)

H. J. B. Birks, Department of Biology, University of Bergen, Allégaten 41, N-5007 Bergen, Norway.

S. Juggins, School of Geography, Politics and Sociology, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK.