Quaternary Science Reviews 30 (2011) 1272–1278

Contents lists available at ScienceDirect

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev

A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages

R.J. Telford [a,b,*], H.J.B. Birks [a,b,c,d]

[a] Department of Biology, University of Bergen, Thormøhlensgate 53 A, N-5006 Bergen, Norway
[b] Bjerknes Centre for Climate Research, Allégaten 55, N-5007 Bergen, Norway
[c] School of Geography and the Environment, University of Oxford, UK
[d] Environmental Change Research Centre, University College London, UK

ARTICLE INFO

Article history:
Received 16 December 2010
Received in revised form
28 February 2011
Accepted 1 March 2011
Available online 23 March 2011

Keywords:
Palaeoecology
Palaeoclimate
Reconstruction
Significance test
Transfer functions

ABSTRACT

We present a method to test the statistical significance of a quantitative palaeoenvironmental reconstruction inferred from biotic assemblages with a transfer function. A reconstruction is considered statistically significant if it explains more of the variance in the fossil data than most reconstructions derived from transfer functions trained on random environmental data. Given reconstructions of several environmental variables from the same fossil proxy, the method can determine which is the best reconstruction, and if there is sufficient information in the proxy data to support multiple independent reconstructions. Reconstructions that fail this test have limited credibility and should be interpreted with considerable caution.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Since Imbrie and Kipp (1971) introduced the methodology, quantitative reconstructions have been made for scores of environmental variables, using dozens of proxies, at thousands of sites. Many of these reconstructions are valid and useful. Some are not. The difficulty is to determine which. This is a two-part problem. The first part is to decide if a variable could *potentially* be reconstructed. The second is to determine if a variable can be reconstructed at a *specific* site.

The first problem is normally addressed by using a constrained ordination to test if a variable explains a significant proportion of the variance in the biotic data in the modern training-set, and then choosing an appropriate transfer function method and estimating its performance with cross-validation (Birks, 1995). Recent research has highlighted the potential pitfalls caused by spatial autocorrelation (Telford and Birks, 2005, 2009), model selection bias (Telford et al., 2004), and unevenly sampled gradients (Telford and Birks, submitted for publication), that can potentially make ecologically unimportant variables appear possible to reconstruct, and make model performance statistics over-optimistic.

Passing the first test is necessary but not sufficient. With a judicious choice of enough training-set sites, maximising variation in the environmental variable of interest, and minimising nuisance gradients (Birks, 1995), any ecologically relevant variable will be a significant predictor of the biotic data. Statistical significance of an environmental variable as a predictor of the training-set biota is a property of the modern training-set. It reveals which variables can potentially be reconstructed, but it does not provide a guarantee that a variable can be meaningfully reconstructed at a specific site. Any fossil data can be run through the transfer function and, provided there are species in common, numeric results will be obtained. But these results are not necessarily useful. On rare occasions, it is possible to validate the reconstruction against instrumental records (e.g., Lotter, 1998). While reassuring when successful, validation is valid only for the site, and time-window, for which it was performed. For most sites, no validation is possible, and instead a range of metrics have been used to assess reconstructions. These include the presence of good analogues for the fossil samples in the training-set, and the goodness-of-fit of the fossil observations passively placed in an ordination of the modern observations and the environmental variable being reconstructed (Birks, 1995). While these can highlight reconstructions that are

[*] Corresponding author. Department of Biology, University of Bergen, Thormøhlensgate 53 A, N-5006 Bergen, Norway. Tel.: +47 55583422.
E-mail address: richard.telford@bio.uib.no (R.J. Telford).

0277-3791/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.quascirev.2011.03.002

unlikely to be reliable because the fossil assemblages have aberrant compositions, they cannot identify sites where changes in the environmental variables being reconstructed were not important drivers of assemblage composition.

Many authors reconstruct multiple variables from the same fossil data. There are two ways to interpret these multiple reconstructions. The first is when a group of variables is reconstructed that are expected to be highly correlated. For example, Telford et al. (1999) reconstructed conductivity, pH, anion ratio and cation ratio from diatoms in a soda lake. These should be viewed as multiple alternative reconstructions, little more than a rescaling and change of units. The second interpretation is that the each reconstruction contains unique information about the palaeoenvironment. Unfortunately, it is seldom clear which of these two interpretations the author believes, and the assumption of the second interpretation, that each reconstruction contains unique information, is rarely, if ever, tested. Indeed, except where validation against instrumental data is possible, the utility of a single reconstruction is not assessed either.

Since it is always possible to obtain numeric results for a reconstruction of any variable, regardless of its ecological relevance, some test of utility is required. We propose the following test, that a single reconstruction should explain more of the variation in the fossil data in a constrained ordination than a transfer function trained on random environmental data applied to the same fossil data. Multiple independent reconstructions should each explain more of the fossil data than a random variable after the other reconstructions have been partialled out. The remainder of this paper explores the properties of these tests using transfer functions from several training-sets on a variety of fossil data. In particular, we wish to test: are reconstructions that most palaeoecologists would think robust statistically significant; are reconstructions that we think questionable not significant; and how robust is the method to various possible causes of Type I and Type II errors?

2. Data and methods

Published reconstructions from a range of organisms and transfer function methods, and using training-sets with different properties were analysed (Table 1).

The proportion of variance in the fossil data explained by a single reconstruction is estimated using a constrained ordination. We use redundancy analysis (RDA) as the species turnover is relatively low in most of our examples (Table 1). Then, using the biotic data from the same training-set, reconstructions are inferred from transfer functions trained on random environmental variables drawn from a uniform distribution. The proportion of the variance explained by these random reconstructions is estimated. If the observed environmental variable explains more of the variance than 95% of the random reconstructions, the reconstruction is deemed statistically significant. The proportion of the variance explained by the first axis of a principal components analysis (PCA) is also recorded, as this represents the maximum proportion of the variance in the fossil data that the reconstruction could possibly explain.

If there are multiple reconstructions, a forward selection procedure is adopted. First, the reconstruction that explains the most variance is accepted. Then other reconstructions are tested to determine if they explain significantly more variance than random reconstructions when the first reconstruction is partialled out. This procedure can be repeated until there are no more significant reconstructions.

In all cases, 999 random environmental variables were generated to produce the null distribution. Significance values are given as the fraction of random variables that explain as much as or more of the fossil data than the observed variable.

In this paper, we chose to develop the null distribution by using uniformly distributed random variables rather than by permuting the environmental variable, the procedure used by CANOCO *(*ter Braak and Šmilauer, 2002*)*. This choice allows us to compare multiple reconstructions against the same null distribution, and, more importantly, to use spatially structured random variables when the training-set is autocorrelated. Quantile–quantile plots (not shown) of the proportion of variance explained by reconstructions derived from permuted or uniformly distributed variables do not deviate from the 1:1 line. This result is not method dependent.

We use weighted averaging (WA) with inverse deshrinking on square-root transformed data and the modern analogue technique (MAT) with squared chord distance and five analogues. The

Table 1
Reconstructions analysed. Effective number of species is estimated using N2 (Hill, 1973). Turnover is estimated as the length of the first detrended correspondence analysis axis (standard deviation units). Significance is based on 999 trials. WA – weighted averaging, MAT – modern analogue technique.

Site Name	Proxy	Effective No. Species	No. observations	Turnover SD	Time span	Variable (s)	Transfer Function	Significance	Training-set	Reconstruction Citation
Round Loch of Glenhead (RLGH)	Diatoms	15.6	20	0.7	Industrial	pH	WA	0.006	SWAP (Birks et al., 1990)	Allott et al. (1992)
Kråkenes	Chironomids	9.9	44	3.2	Lateglacial	Temperature	WA	0.004	Brooks and Birks (2001)	Brooks and Birks (2000b)
Bjørnfjelltjørn	Chironomids	12.8	61	1.4	Holocene	Temperature	WA	0.183	Brooks and Birks (2001)	Brooks (2006)
Whitrig Bog	Chironomids	14.8	76	2.4	Lateglacial	Temperature	WA	0.001	Brooks and Birks (2001)	Brooks and Birks (2000a)
Holebudalen	Chironomids	12.5	53	1.8	Holocene	Temperature	WA	0.009	Brooks and Birks (2001)	Velle et al. (2005)
Vøring Plateau MD95-2011	Planktonic foraminifera	3.3	380	1.4	Lateglacial – modern	Annual SST	MAT	0.001	Pflaumann et al. (2003)	Risebrobakken et al. (2003)
Chukchi Sea	Dinoflagellates	4.0	86	1.2	Holocene	Summer SST, sea-ice duration, summer salinity	MAT	0.003, 0.007, 0.146	Radi and de Vernal (2008)	McKay et al. (2008)
Bjørnfjelltjørn	Pollen	4.8	63	1.2	Holocene	July temperature, January temperature, precipitation	MAT	0.029, 0.160, 0.743	(Bjune et al., 2010)	(Birks and Peglar, unpublished).
Holebudalen	Pollen	8.0	94	1.2	Holocene	July temperature, January temperature, precipitation	MAT	0.001, 0.173, 0.153	(Bjune et al., 2010)	Eide et al. (2006)

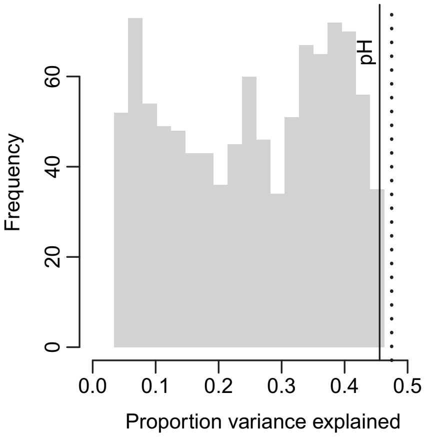

Fig. 1. Histogram of the proportion of variance in the RLGH diatom record explained by 999 transfer functions trained with random data. Solid black line marks the proportion of variance explained by the SWAP pH transfer function. The dotted line marks the proportion of variance explained by the first axis of a PCA of the fossil data.

dinoflagellate cyst data were log-transformed following Radi and de Vernal (2008). All calculations were performed using the statistical language R version 2.11.1 (R Development Core Team, 2010). Transfer functions were fitted with the rioja library version 0.5–6 (Juggins, 2009). RDA and PCA were run with the vegan library version 1.17–2 (Oksanen et al., 2010).

3. Results

3.1. Are reconstructions we trust statistically significant?

Many reconstructions have a good ecological basis, are applied to sensitive sites where change is expected, and the range of the reconstructed variable is at least as large as the root mean square error of prediction (RMSEP). It seems reasonable to anticipate that these will be statistically significant.

Diatom-inferred pH changes in the Round Loch of Glenhead (RLGH) since the start of the industrial period are statistically significant (Fig. 1, Table 1), as are planktonic foraminifera-inferred sea-surface temperature (SST) changes since the Younger Dryas at the Vøring Plateau and Lateglacial chironomid-inferred July temperature changes from two Norwegian lakes.

3.2. Are reconstructions we do not trust statistically significant?

Some reconstructions are more problematic, either because the reconstructed changes are small relative to the RMSEP, the site is of questionable sensitivity, or the environmental variable being reconstructed is of uncertain ecological significance. We expect that at least some of these reconstructions will not be statistically significant.

Holocene chironomid-inferred July air temperature reconstructions from Norwegian lakes are somewhat idiosyncratic (Velle et al., 2005, 2010), suggesting that factors other than temperature may be important in at least some lakes. Here we test two Holocene reconstructions (Table 1). The reconstruction at Holebudalen, which shows a decline in temperature throughout the Holocene consistent with changes in orbital forcing, is significant. Bjørnfjelltjørn, which has a more erratic reconstruction, is not statistically significant.

3.3. Multiple reconstructions

When attempting to generate multiple reconstructions, the first test is to determine which, if any, of the reconstructions are significant on their own. McKay et al. (2008) reconstruct salinity, summer SST and duration of sea-ice cover from dinoflagellate cysts in a core in the Chukchi Sea. We find that the salinity reconstruction is not significant (Fig. 2a). This is not unexpected as the robustness of dinoflagellate-salinity transfer functions was questioned by Telford (2006). Both summer SST and sea-ice duration are significant as single reconstructions; of these, summer SST explains marginally more of the variance, but both explain much less than the first unconstrained axis of the RDA suggesting that they do not capture the major trends in the fossil data. The robustness of the order of importance of the variables could be established by bootstrapping the observations in the training-set; this possibility will be explored elsewhere. Once summer SST is partialled out, sea-ice duration is not significant (Fig. 2b).

Pollen-July temperature reconstructions are significant from both Holebudalen and Bjørnfjelltjørn if MAT is used. Neither precipitation nor January temperature is significant at either site.

4. Discussion

The significance tests we develop here fill a major gap in the range of numeric procedures available to palaeoecologists. Our

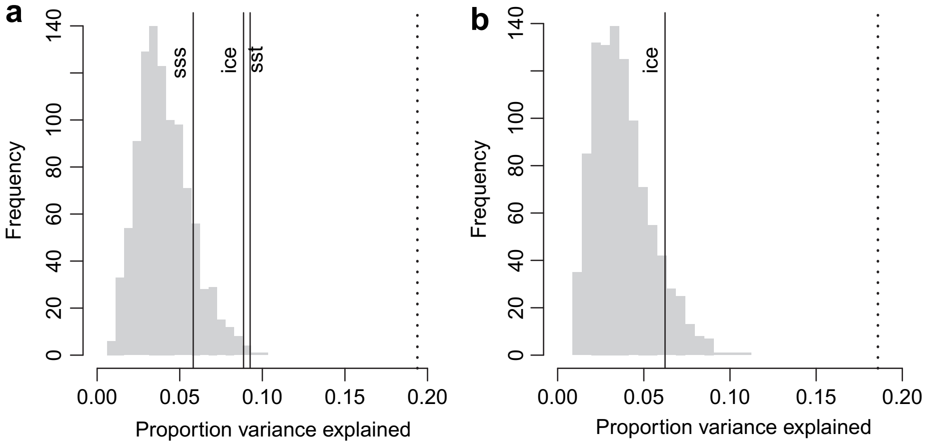

Fig. 2. a) Proportion of variance in the Chukchi Sea fossil dinocyst data explained by salinity (sss), sea-ice duration (ice) and summer SST (sst). Dashed line shows the proportion explained by the first axis of a PCA. b) Proportion of variance explained by sea-ice duration after summer SST is partialled out. In both graphs, the histograms show the amount of variance explained by 999 reconstructions derived from transfer functions trained on random data.

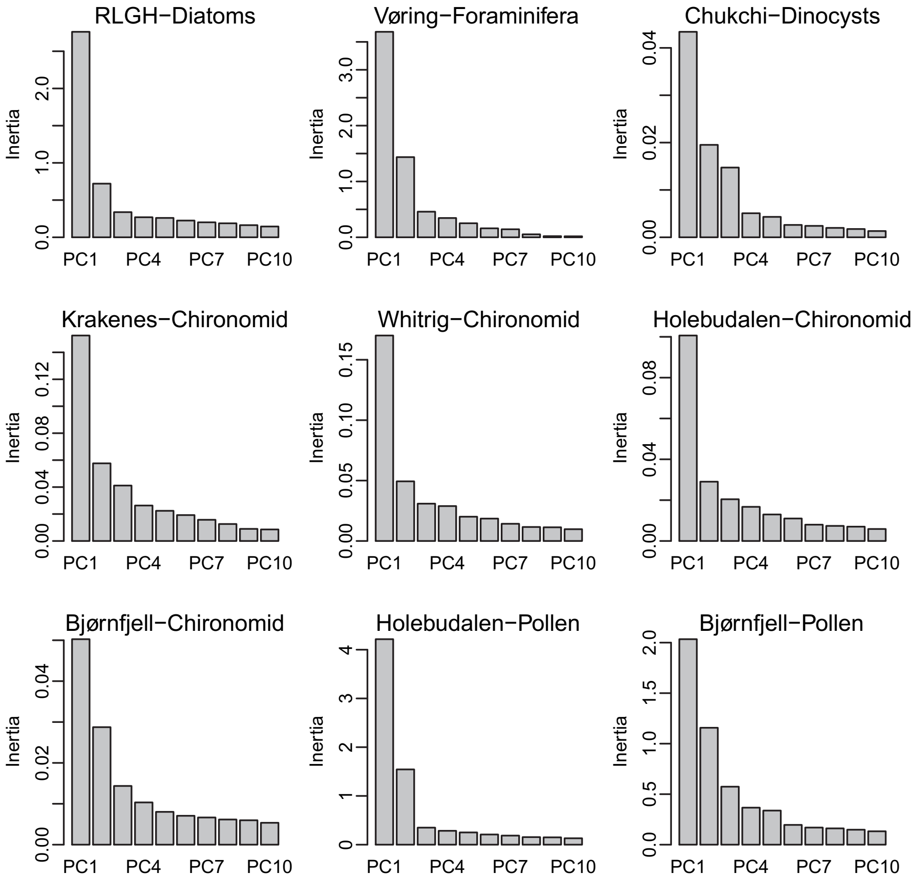

Fig. 3. Screeplots of PCA results of the different fossil data used. Broken stick model results (not shown) tend to indicate several axes are marginally significant, perhaps because of temporal autocorrelation.

inability to reject reconstructions that are not statistically significant has made it difficult to interpret divergent reconstructions either from the same proxy at different sites or different proxies at the same site. Uncritical acceptance of reconstructions can result in physically unlikely scenarios (Telford, 2006) that potentially divert the research community away from more fruitful lines of enquiry.

Reconstructing multiple variables from the same proxy is common practice. Although our survey is far from exhaustive, it would appear that it is seldom possible to reconstruct more than one, or perhaps two, independent variables from a proxy. This is perhaps not surprising given that screeplots of PCA of fossil data tend to reveal only one or two interpretable axes (Fig. 3). It may, however, be possible to reconstruct synthetic variables composed of two or more variables such as warm-wet vs. cool-dry. We will explore this possibility elsewhere.

If multiple independent reconstructions are unlikely to be statistically significant, care need to be taken to select sites for reconstruction. Sites should ideally be chosen where *a priori* we expect the variable of interest to be the dominant driver of biotic change (Velle et al., 2005), as it may be difficult to successfully reconstruct any statistically significant variables from sites where biotic change is driven by several factors.

Rejecting the null hypothesis does not prove that the alternative hypothesis is correct: that a reconstruction performs significantly better than the null distribution does not guarantee that the reconstruction is valid. For example, if several environmental variables have similar effects on the biotic assemblages, changes in one of these variables may result in significant, but wrong, reconstructions for any of the others. For instance, Korhola et al. (2000) argued that the chironomid response to lake depth could be confounded by changes in temperature or hypolimnetic oxygen concentration.

As with all statistical methods, there is a risk that the null hypothesis will be incorrectly rejected, or that the null hypothesis will be accepted when the alternative hypothesis is correct. Below we explore the behaviour of the methods under various possible causes of such errors.

4.1. Type I errors

Type I errors occur when the null hypothesis is incorrectly rejected and represent a “false positive”. Transfer function performance statistics have been shown to be over-optimistic if the training-set is spatially autocorrelated (Telford and Birks, 2005, 2009), with an inflated risk of a Type I error. Here we test if spatial autocorrelation also biases the reconstruction’s significance tests. We generate and test reconstructions for the Vøring foraminifera record using transfer functions trained on the simulated environmental fields developed in Telford and Birks (2005) for the North Atlantic planktonic foraminifera training-set. Of the one hundred simulated environmental fields with an effective range of 10,000 km, 16% of the MAT reconstructions and 23% of the WA reconstructions, are apparently significant at $p = 0.05$.

The finding that WA reconstructions have a higher risk of Type I errors than MAT when the training-set is spatially autocorrelated is the opposite of that found for training-sets, where MAT is more susceptible to autocorrelation inflating performance statistics (Telford and Birks, 2005). This apparent contradiction requires explanation.

Fig. 4 shows reconstructions of three random variables for the Vøring Plateau foraminiferal record using either MAT or WA.

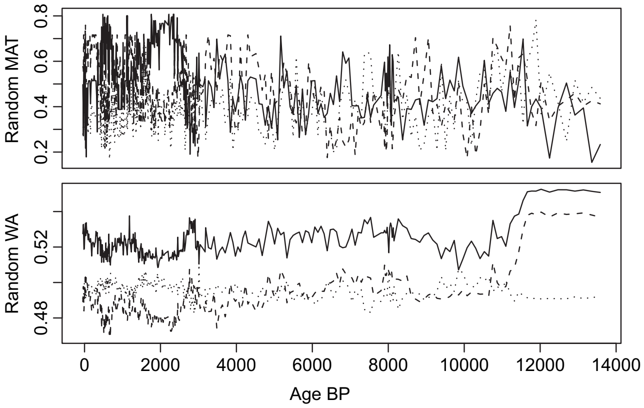

Fig. 4. Reconstructions of three random variables for the Vøring Plateau foraminifera record using MAT (upper) and WA (lower).

Although none of the random variables were spatially autocorrelated, the reconstructions are temporally autocorrelated. The MAT reconstruction is autocorrelated because some of the same training-set observations are selected as analogues for adjacent fossil observations in the core. If the analogues that are swapped between adjacent fossil observations are spatially close, then increasing spatial autocorrelation in the environmental variable will increase the similarity of these analogues, regardless of the ecological importance of the variable. This results in an increased Type I error with the autocorrelated random variables.

The random WA reconstruction is temporally autocorrelated because it is the weighted average of the fossil abundances down core and these are temporally autocorrelated. The axes of the RDA of the fossil data are a weighted sum of the species abundances, so the reconstruction (a weighted average) will be a good predictor of the fossil assemblages if the WA optima correlate well with the first axis species scores of the RDA. It might be anticipated that, for the taxa found in the fossil data, the correlation between the WA optima of the environmental variable of interest and the WA optima of a random variable would be near zero. In fact, the correlation is rather strong, because taxa with similar optima are typically found at the same sites (Fig. 5). The strength of this correlation increases if the environmental variable is spatially autocorrelated. This results in an elevated Type I error rate.

The elevated risk of a Type I error when a variable is reconstructed from a transfer function trained on an autocorrelated training-set implies that the null distribution should be derived from spatially structured random variables rather than uncorrelated random variables. Methods for generating suitable spatially structured random variables are discussed in Telford and Birks (2009). The Type I error rate will be reconstruction specific.

Nothing in this analysis affects the conclusions of Telford and Birks (2005) regarding the elevated risk of Type I errors for spatially autocorrelated training-sets.

4.2. Type II errors

Type II errors occur when the null hypothesis is incorrectly accepted and represent a "false negative". There are several factors that could elevate the risk of a Type II error. We explore these by manipulating some of the data used above, especially the diatom record from the RLGH and the chironomid record from Holebudalen. These were chosen as the reconstructed environmental change is only somewhat higher than the RMSEP of the respective transfer functions, unlike the reconstructions covering the Lateglacial. This should make them more fragile. We define the Type II error rate as high when it is above 0.2.

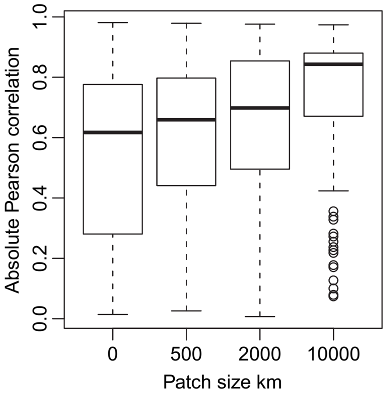

Fig. 5. Boxplot of absolute Pearson correlations between the SST WA optima and random WA optima with different amounts of spatial structure for foraminifera taxa found in the Vøring Plateau core. Results are from 100 trials.

There is the possibility that species-poor assemblages have a higher risk of a Type II error. The effective number of species, estimated with Hill's (1973) N2, is probably a better guide than species richness. We explore this risk by randomly removing increasing numbers of species, stratified by abundance class, from a fossil assemblage, to determine how low the effective number of species can be before the Type II error rate becomes high.

The RLGH record has 15.6 effective taxa. These can be reduced to seven before the risk of a Type II error becomes high (Fig. 6). The Holebudalen record has 12.5 effective species. These can be reduced to nine before the risk of a Type II error becomes high. The high risk of a high Type II error for assemblages with low diversity may explain why the WA reconstructions are not significant for the pollen, Vøring, and Chukchi records which all have eight or fewer effective species.

It is possible that sites with few fossil observations will have a greater risk of a Type II error. We explore this risk by taking a statistically significant reconstruction and progressively discarding observations, whilst maintaining the range of the reconstructed values, until the reconstruction is not significant. The more fossil observations that can be dropped, the more robust the method is to this risk.

The 20-observation RLGH diatom record can be reduced to seven observations before the risk of a Type II error for the pH reconstruction is high. The Holebudalen chironomid record can be reduced from 53 to 16 observations before the risk of a Type II error is high. Most high-resolution reconstructions contain sufficient observations that they should be relatively robust to this risk.

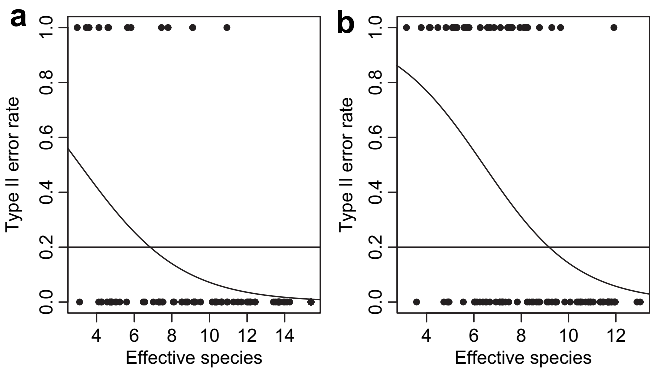

Fig. 6. Type II error rate with reduced number of species for a) RLGH diatom-pH reconstruction and b) Holebudalen chironomid-temperature reconstruction. Black dots represent the success or failure of individual trials.

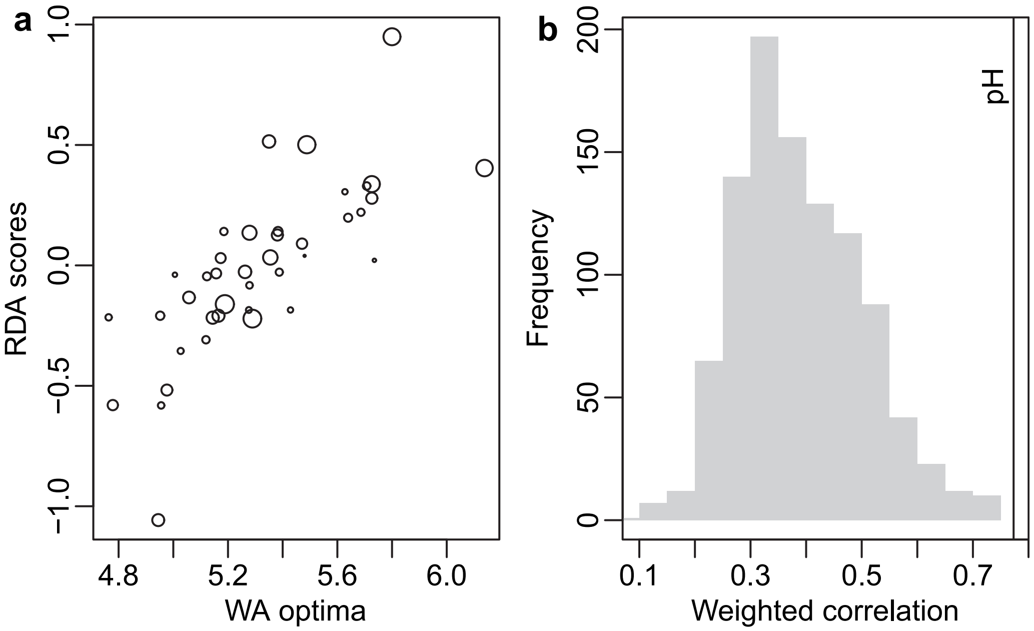

Fig. 7. a) RLGH RDA axis 1 species scores against SWAP pH WA optima. b) Histogram of absolute weighted correlation of RDA species scores against WA optima for 999 random variables. The observed correlation, shown by the vertical line, exceeds all these simulations.

Reconstructions from periods of relatively small changes in the environmental variable of interest, at least relative to other changes in other variables, may have a high risk of a Type II error. This risk is of particular importance given the societal relevance of reconstructing late-Holocene climate changes (Mann et al., 2008) to increase our understanding of natural climatic variability as we enter a period with increasing anthropogenic forcings. We explore this risk by progressively omitting fossil observations furthest from the median reconstruction value, thus shrinking the range of reconstructions.

The WA reconstructed pH range at the RLGH is 0.37, a little larger than the leave-one-out RMSEP of 0.32. Dropping samples, the reconstruction remains significant at $p = 0.05$ until the range falls below 0.27 pH units. The Holebudalen reconstruction ceases to be significant once the range is truncated to below 1.15 °C, well below the RMSEP of 1.54 °C.

A final risk of an enhanced Type II arises from an inadequate or inappropriate training-set. If the training-set is small, optima will be poorly estimated and there will be few available analogues. We explore this risk by reducing the size of the training-set by selecting observations at random, stratified by the environmental gradient. We expect MAT reconstructions to be more susceptible to Type II errors with a small training-set due to the paucity of suitable analogues.

The SWAP training-set has 167 observations. Provided at least 20 are retained, the Type II error rate for the RLGH WA reconstruction is low. In contrast, the RLGH MAT reconstruction is only marginally significant even when all training-set observations are retained. The Atlantic foraminifera training-set contains 973 observations. The Type II error rate for the Vøring reconstruction is low provided at least 25 of these are retained. That such small training-sets can generate significant reconstructions suggests that inappropriate training-sets, i.e. ones that are taxonomically or environmentally remote from the site being reconstructed, are likely to be a more serious problem than training-set size.

4.3. Transfer function method independent?

The tests developed here can be used with any transfer function method, but the test can be interpreted in different ways for different methods.

The axes of the RDA of the fossil data are a weighted sum of the species abundances, so a WA reconstruction should explain a statistically significant part of the variance in the fossil data if the WA optima correlate well with the first axis species scores of the RDA. This test can be applied directly (Fig. 7), and might have some attractive properties, especially when there are very few fossil observations, but is limited to WA. MAT reconstructions are significant if the environmental variable for the succession of analogues selected is coherent with changes in the biota. These different underlying modes of operation affect the susceptibility to Type II errors. WA reconstructions have many degrees-of-freedom if they have many effective species. MAT reconstructions have many degrees-of-freedom if they have a large turnover of analogues.

We therefore recommend that a MAT reconstruction be tested if there are few species and many potential analogues, and that a WA reconstruction should be tested if the assemblages are diverse and the training-set relatively small. These choices should minimise the risk of a Type II error. It is not, however, appropriate to test both methods and select the one with the greatest significance unless a correction for multiple testing is made. If there are few species, so a Type II error is expected for WA, it is perhaps reasonable to test the significance of the MAT reconstruction, but still present the WA reconstruction, provided the two reconstructions are comparable.

The tests may not be computationally feasible with all transfer function methods. For example, generating many random reconstructions will be impossibly slow with neural networks. This is not sufficient reason to avoid using these significance tests, rather it would seem to be a further reason to avoid neural networks, given that their performance is not outstanding, and they are not very robust against spatial autocorrelation (Telford and Birks, 2005).

5. Conclusions

If reconstructions inferred from transfer functions trained on random environmental variables explain as much of the variance of the fossil data as reconstructions from transfer functions trained on the observed environmental data, we cannot be certain that our reconstruction is better than a random one. Reconstructions that fail this test should be interpreted with considerable caution.

It is inevitable that some reconstructions, perhaps the result of many months of effort, will be found not to be statistically significant. This situation, while new to palaeoecological reconstructions, is far from unique to palaeoecology, as negative results are common in many disciplines. Palaeoecology and palaeoclimatology cannot claim exemption from the usual strictures to interpret significant results only, especially if we wish to claim that our results are policy relevant.

We strongly recommend that the significance tests presented here are used whenever a reconstruction is published.

Acknowledgements

This work was supported by the Norwegian Research Council projects ARCTREC and PES. This is publication no. A318 from the Bjerknes Centre for Climate Research. R code is available from http://cran.r-project.org/web/packages/palaeoSig for the functions developed in this paper. We are grateful to Steve Brooks for providing data, to Gaute Velle for discussions, and to Cajo ter Braak for comments.

References

Allott, T.E.H., Harriman, R., Battarbee, R.W., 1992. Reversibility of lake acidification at the Round Loch of Glenhead, Galloway, Scotland. Environmental Pollution 77, 219–225.

Birks, H.J.B., 1995. Quantitative palaeoenvironmental reconstructions. In: Maddy, D., Brew, J.S. (Eds.), Statistical Modelling of Quaternary Science Data. Quaternary Research Association, Cambridge, pp. 116–254.

Birks, H.J.B., Line, J.M., Juggins, S., Stevenson, A.C., ter Braak, C.J.F., 1990. Diatoms and pH reconstruction. Philosophical Transactions of the Royal Society B-Biological Sciences 327, 263–278.

Bjune, A.E., Birks, H.J.B., Peglar, S.M., Odland, A., 2010. Developing a modern pollen–climate calibration data set for Norway. Boreas 39, 674–688.

Brooks, S.J., 2006. Fossil midges (Diptera: chironomidae) as palaeoclimatic indicators for the Eurasian region. Quaternary Science Reviews 25, 1894–1910.

Brooks, S.J., Birks, H.J.B., 2000a. Chironomid-inferred Lateglacial air temperatures at Whitrig Bog, southeast Scotland. Journal of Quaternary Science 15, 759–764.

Brooks, S.J., Birks, H.J.B., 2000b. Chironomid-inferred lateglacial and early-Holocene mean July air temperatures for Kråkenes Lake, western Norway. Journal of Paleolimnology 23, 77–89.

Brooks, S.J., Birks, H.J.B., 2001. Chironomid-inferred air temperatures from Late-glacial and Holocene sites in north-west Europe: progress and problems. Quaternary Science Reviews 20, 1723–1741.

Eide, W., Birks, H.H., Bigelow, N., Peglar, S., Birks, H.J.B., 2006. Holocene forest development along the Setesdal valley, southern Norway, reconstructed from macrofossil and pollen evidence. Vegetation History and Archaeobotany 15, 65–85.

Hill, M.O., 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54, 427–432.

Imbrie, J., Kipp, N.G., 1971. A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core. In: Turekian, K.K. (Ed.), The Late Cenozoic Glacial Ages. Yale University Press, New Haven, pp. 71–181.

Juggins, S., 2009. rioja: an R Package for the Analysis of Quaternary Science Data, Version 0.5-6.

Korhola, A., Olander, H., Blom, T., 2000. Cladoceran and chironomid assemblages as qualitative indicators of water depth in subarctic Fennoscandian lakes. Journal of Paleolimnology 24, 43–54.

Lotter, A.F., 1998. The recent eutrophication of Baldeggersee (Switzerland) as assessed by fossil diatom assemblages. The Holocene 8, 395–405.

Mann, M.E., Zhang, Z., Hughes, M.K., Bradley, R.S., Miller, S.K., Rutherford, S., Ni, F., 2008. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. Proceedings of the National Academy of Sciences 105, 13252–13257.

McKay, J.L., de Vernal, A., Hillaire-Marcel, C., Not, C., Polyak, L., Darby, D., 2008. Holocene fluctuations in Arctic sea-ice cover: dinocyst-based reconstructions for the eastern Chukchi Sea. Canadian Journal of Earth Sciences 45, 1377–1397.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2010. vegan: Community Ecology Package. R Package Version 1.17-2.

Pflaumann, U., Sarnthein, M., Chapman, M., d'Abreu, L., Funnell, B., Huels, M., Kiefer, T., Maslin, M., Schulz, H., Swallow, J., van Kreveld, S., Vautravers, M., Vogelsang, E., Weinelt, M., 2003. Glacial North Atlantic: sea-surface conditions reconstructed by GLAMAP 2000. Paleoceanography 18, 1065.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Radi, T., de Vernal, A., 2008. Dinocysts as proxy of primary productivity in mid-high latitudes of the Northern Hemisphere. Marine Micropaleontology 68, 84–114.

Risebrobakken, B., Jansen, E., Andersson, C., Mjelde, E., Hevrøy, K., 2003. A high-resolution study of Holocene paleoclimatic and paleoceanographic changes in the Nordic Seas. Paleoceanography 18, 1017.

Telford, R.J., 2006. Limitations of dinoflagellate cyst transfer functions. Quaternary Science Reviews 25, 1375–1382.

Telford, R.J., Andersson, C., Birks, H.J.B., Juggins, S., 2004. Biases in the estimation of transfer function prediction errors. Paleoceanography 19 (Artn Pa4014).

Telford, R.J., Birks, H.J.B., 2005. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. Quaternary Science Reviews 24, 2173–2179.

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

Telford, R.J., Birks, H.J.B. Effect of uneven sampling along an environmental gradient on transfer-function performance. Journal of Paleolimnology, submitted for publication.

Telford, R.J., Lamb, H.F., Umer Mohammed, M., 1999. Diatom-derived palaeoconductivity estimates for Lake Awassa, Ethiopia: evidence for pulsed inflows of saline groundwater. Journal of Paleolimnology 21, 409–422.

ter Braak, C.J.F., Šmilauer, P., 2002. CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5). Microcomputer Power, Ithaca, NY, USA.

Velle, G., Brodersen, K.P., Birks, H.J.B., Willassen, E., 2010. Midges as quantitative temperature indicator species: essons for palaeoecology. The Holocene 20, 989–1002.

Velle, G., Brooks, S.J., Birks, H.J.B., Willassen, E., 2005. Chironomids as a tool for inferring Holocene climate: an assessment based on six sites in southern Scandinavia. Quaternary Science Reviews 24, 1429–1462.