Understanding past ocean circulations: a nonparametric regression case study

Richard Samworth* and Heather Poore†

August 16, 2005

Abstract

Oceanographers study past ocean circulations and their effect on global climate through carbon isotope records obtained from microfossils deposited on the ocean floor. An initial goal is to estimate the carbon isotope levels for the Pacific, Southern and North Atlantic Oceans over the last 23 million years, and to provide confidence bands. We consider a nonparametric regression model, and demonstrate how several recent developments in methodology make local linear kernel regression an attractive approach for tackling the problem. The results are used to estimate a quantity called the proportion of Northern Component Water and its effect on global climate. Several interesting and important geophysical and oceanographic conclusions are suggested by the study.

Keywords: Confidence bands, errors-in-variables, local linear kernel estimator, model checking, nonparametric regression, Northern Component Water, SIMEX algorithm

*Address for correspondence: Richard Samworth, Statslab, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB. Email: rjs57@cam.ac.uk. Tel: +44 1223 337950. Fax: +44 1223 337956.

†Heather Poore, Department of Earth Sciences, Bullard Laboratries, Madingley Road, Cambridge, CB3 0EZ

1 Introduction

Understanding the way that global climate has evolved over time is important to scientists trying to predict future climates. Changes in climatic conditions are strongly linked to the way in which oceans distribute heat and moisture to the atmosphere, and this depends upon the prevailing ocean currents.

Oceanographers study past changes in the ocean circulation system by examining microfossils that record the isotopic composition of water at the time at which they live. In particular, they measure the ratio of ^{13}C to ^{12}C, compared to a standard, in the shell of the fossil, and refer to this as the δ^{13}C value. This can be used to determine the direction in which ocean currents flowed during the lifetime of the organism. Our initial goal is to study the evolution of these δ^{13}C levels over the last 23 million years (Ma) in the Pacific, Southern and North Atlantic oceans, and to deduce information about ocean current flow and global climate during this interval, which geologists refer to as the Neogene period.

Of course, errors are inherent in the measurement process, and our approach to the regression problem is to use a local linear kernel estimator (Wand and Jones, 1995; Bowman and Azzalini, 1997). Recent advances in methodology have made such an approach very attractive, with an increased level of sophistication in analyses now possible. One aim of this article is to highlight and analyse some of the most recent developments in the context of a practical problem, and show how confidence bands, model checking and errors-in-variables problems can all now be handled.

The data collection is a time-consuming and costly process which began in the 1970s and is still ongoing; see Section 2 for further details. The data presented in Figure 1 are the most recent data available, and are reported on the latest Geological time scale (Lourens et al., 2004), which allows considerably higher precision in dating samples.

Many authors have studied older data sets, often covering shorter age ranges and a subset of the oceans considered here, including Oppo and Fairbanks (1987), Mix et al. (1995), Wright and Miller (1996), Billups, Ravelo and Zachos (1997), Zachos et al. (2001) and Billups, Channell and Zachos (2002). In some cases, the data are smoothed using the locally constant Nadaraya–Watson estimator with a Gaussian kernel and constant bandwidth. This has the disadvantage that the bandwidth must be chosen sufficiently large to cover gaps in the data, and tends to result in oversmoothing elsewhere. Other smoothing techniques have also been employed: Billups, Ravelo

and Zachos (1997) and Zachos et al. (2001) use three- and five-point running means respectively, while Oppo and Fairbanks (1987) smooth their data with a five-point weighted least-squares estimator.

Unfortunately, none of the papers mentioned above provides any estimate of the uncertainty in the resulting regression curves. There is now a large literature on the important problem of constructing confidence bands in nonparametric regression, and in Section 3 we consider in detail those proposed by Xia (1998) and Claeskens and Van Keilegom (2003) for our nonparametric model. In Section 4, we examine the validity of the assumptions of our model using the hypothesis test proposed by Einmahl and Van Keilegom (2004). Another feature of this section is an attempt to tackle the issue of incorporating the age uncertainty into our regression estimates using the SIMEX algorithm, introduced for parametric settings by Cook and Stefanski (1994) and discussed in nonparametric contexts by Carroll, Ruppert and Stefanski (1995).

One of the main reasons for studying δ^{13}C records is to compute estimates of the proportion of the volume of water in the Southern Ocean originating from the North Atlantic (the rest is assumed to come from the Pacific). This proportion is referred to as the proportion of Northern Component Water, and can be thought of as a measure of the efficiency of oceanic heat transfer from the equator to the poles. A conservation of mass argument (cf. Section 5) allows us to express this proportion as a simple function of the δ^{13}C values in the three oceans. While Oppo and Fairbanks (1987), Venz and Hodell (2002) and Wright and Miller (1996) have all estimated this proportion before, we are able to assess the uncertainty in the estimate using the delta method. This provides additional insight and suggests several geophysical and oceanographic conclusions, as well as implying that the estimate is meaningless in certain age ranges. Some concluding remarks are presented in Section 6.

2 Data Collection

The data are obtained by extracting large cores from the ocean floor and picking particular species of microfossils called foraminifera from small slices of sediment at different depths within the core. The foraminifera are then cleaned before analysing the δ^{13}C values in their shells, which are made of calcium carbonate.

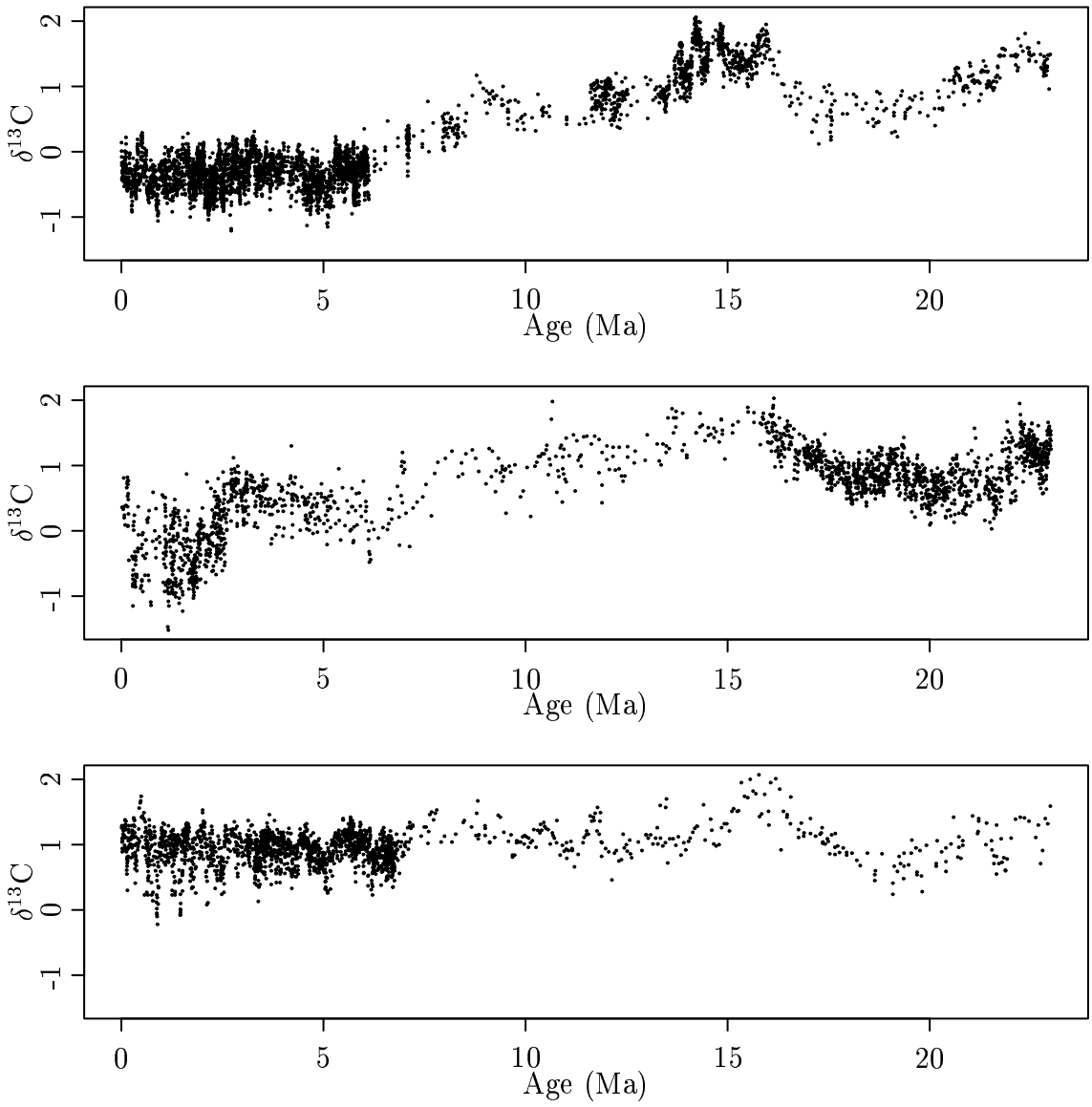

Figure 1: The data show δ^{13}C measurements over time, measured in millions of years, for the Pacific Ocean (top), Southern Ocean (middle) and North Atlantic Ocean (bottom). A colour version of this figure, showing the extent of the overlap of the different age ranges covered by the different cores, appears on the journal website.

For each ocean, we have used several cores taken at different times and in different locations in order to cover the age range of interest. The colour version of Figure 1, which appears on the journal website, shows that there is not a great deal of overlap between the different cores. The Pacific, Southern and North Atlantic data were made up of six, five and five cores, yielding totals of 4020, 2102 and 2079 data points respectively. It is standard practice in marine geology to simply pool the data from different cores together, under the assumption that the systematic variations in samples at carefully chosen locations will be small by comparison with the measurement error and differences between oceans. This assumption will be examined in Section 4.

It is a complex task to convert the depth of a sample, which can be measured accurately, to age. The procedure is to find depths at which there is some kind of age marker, between which marine geologists assume a constant sedimentation rate, and so are able to linearly interpolate to obtain age estimates for the depths at which δ^{13}C values have been measured. The markers which are used to generate the age models in this study come from three different geological disciplines:

1. *Biostratigraphy*: Some types of fossil organisms preserved within the sediment core have known ages at which they evolved or became extinct. By finding the level at which they appear, disappear, or are especially common within the core, biostratigraphers assign an age to a particular depth. In practice, samples are taken perhaps every 1.5m, and the depth–age marker is constrained at these points.

2. *Magnetostratigraphy*: When sediment is deposited, magnetic components of the sediment are in alignment with the Earth's magnetic field. Occasionally, the Earth's magnetic field reverses its direction, and so at certain levels within the core, the sediment changes its polarity alignment. These magnetic reversals have been dated. Some sediment cores do not retain their magnetic signature during drilling, and so this method is not always applicable.

3. *Isotope Stratigraphy*: It is possible to use features of the isotope records themselves (in particular globally-recognised spikes in the ratio of ^{18}O to ^{16}O) to assign ages to depths in the core of interest. This requires well-constrained isotope records, and assumes that the isotope event recorded is synchronous around the globe.

Further details on the techniques used to date samples can be found in the books by

Gradstein, Ogg and Smith (2004) and Shackleton, McCave and Weedon (1999).

3 Main Analysis

For each of the three oceans, the data consist of pairs $\{(x_i, y_i) : i = 1, \ldots, n\}$, with x_i giving the age inferred for the ith sample, and y_i giving the corresponding δ^{13}C measurement. We assume that the pairs are realisations of independent and identically distributed random vectors (X, Y), and for the moment consider the model given by

$$Y = m(X) + \epsilon \tag{3.1}$$

where $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma^2$ and ϵ is independent of X. We write $f(x)$ for the density of X. The initial goal is to estimate the regression function $m(x) = \mathbb{E}(Y|X = x)$.

There are several popular techniques for the estimation of the regression function, including splines (Ruppert, Wand and Carroll, 2003), Fourier series smoothers, Wavelets (Ogden, 1996) and other orthogonal series methods. We choose to apply a local linear kernel estimator defined in (3.2) below, which is discussed in the monographs by Fan and Gijbels (1996) and Chapter 5 of Wand and Jones (1995). This has several attractive features: it is easy to implement, flexible and conceptually simple. Moreover, its mathematical properties are well understood, so some of the more complicated problems mentioned in Section 1 can be dealt with in a way which would not be possible with other techniques. The local linear estimator is given by

$$\hat{m}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{s_2 - s_1(x_i - x)\} K_h(x_i - x)}{s_2 s_0 - s_1^2} y_i, \tag{3.2}$$

where

$$s_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)^r K_h(x_i - x), \quad r = 0, 1, 2,$$

and the scaled kernel $K_h(\cdot) = h^{-1} K(\cdot/h)$ satisfies $\int K_h = 1$. Of course, other local polynomial estimators are also possible, although several authors, e.g. Fan and Gijbels (1996), pp.76–83, have recognised that fitting linear polynomials locally suffices for many applications where primary interest is in recovery of the regression function itself, rather than its derivatives.

Except when estimating the bias of $\hat{m}_h(x)$ in the implementation of the Xia (1998) confidence bands and in the SIMEX algorithm, we use the Epanechnikov kernel

$$K(x) = \frac{3}{4\sqrt{5}}\Big(1 - \frac{x^2}{5}\Big)\mathbb{1}_{\{x^2 \leq 5\}}. \tag{3.3}$$

In all smoothing problems the choice of smoothing parameter, in this case the bandwidth h, is critical. Many papers have been written on automatic rules for bandwidth selection, from which we briefly mention Fan and Gijbels (1995), Härdle and Kelly (1987), Hurvich, Simonoff and Tsai (1998), Ruppert, Sheather and Wand (1995) and Hengartner, Wegkamp and Matzner-Løber (2002). However, most of these papers discuss only constant bandwidths, and we are not aware of any proposals for confidence bands in a variable bandwidth setting. It is clear from Figure 1 that we would like a variable bandwidth due to the marked differences in the density of observations at different ages. We sidestep this problem by dividing the x-axis into blocks of approximately equal point density, and choose a bandwidth for each block separately. The Pacific, Southern and North Atlantic oceans were divided into 21, 16 and 12 regions respectively, and a bandwidth chosen for each. The random-division cross-validation bandwidth selector of Hengartner, Wegkamp and Matzner-Løber (2002) was used as a starting point as it performed well in their simulation comparison. Their simulations were, however, based on a uniform distribution for the design points, and it should be noted that the algorithm tended to pick particularly large bandwidths when the point density was not close to uniform. Although this problem was not too serious after taking care to divide the x-axis into blocks of approximately uniform designs, we found it convenient to make small adjustments to the suggested bandwidth 'by eye'.

An alternative to the procedure described above would be to make a transformation of age based on the (smoothed) empirical distribution function. A single bandwidth could then be used to smooth the data before making the inverse transformation. This amounts to using a variable bandwidth to smooth the original data. While this would avoid the need to divide the x-axis into blocks, we decided against making the transformation because theoretical justification for some of the algorithms (e.g. SIMEX) is lacking in such situations, and because the computational burden is significantly increased when dealing with all of the data simultaneously.

Hereafter, we consider the model (3.1) for each block, with pairs in different blocks assumed independent. For simplicity of notation, we suppress mention of the index of a block, so the model (3.1) applies to a generic block as well as a generic ocean.

Moreover, we assume that the data $\{(x_i, y_i) : i = 1, \ldots, n\}$ in each block are ordered so that $x_1 < \ldots < x_n$. Note that we now allow a different residual variance in each block, which can be estimated using the first-order differences method of Rice (1984) by

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n} (y_i - y_{i-1})^2.$$

We now turn attention to the problem of finding confidence bands. To our knowledge, the first confidence bands applicable to local linear kernel regression were developed by Knafl, Sacks and Ylvisaker (1985), followed by Hall and Titterington (1988) and Sun and Loader (1994). We looked in detail at the bands proposed more recently by Xia (1998) and the two proposed by Claeskens and Van Keilegom (2003).

Xia notes that the leading term in the bias of $\hat{m}_h(x)$ is $m''(x)h^2/2$, and subtracts an estimate $\hat{m}_b''(x)h^2$ of this quantity, where $\hat{m}_b''(x)$ is an estimate of $m''(x)/2$ obtained by local cubic polynomial fitting with a Gaussian kernel $K_0(x) = (2\pi)^{-1/2} e^{-x^2/2}$ and a second bandwidth b, chosen by cross-validation. The confidence limits are found by approximating the bias-corrected and standardised process

$$\{nhf(x)\}^{1/2} \big\{ \hat{m}_h(x) - \hat{m}_b''(x)h^2 - m(x) \big\}$$

by a Gaussian process $Y_n(x)$, and applying the famous result of Bickel and Rosenblatt (1973) which gives the asymptotic Gumbel distribution of $\sup_x |Y_n(x)|$. Before we can compute the limits, however, we require an estimate of the variance of $\hat{m}_h(x)$. We refer to Xia's paper for the precise definition of this estimator, and here denote it by $\hat{V}\big(\hat{m}_h(x)\big)$. The resulting confidence band is

$$\hat{m}_h(x) - \hat{m}_b''(x)\hat{h}^2 \pm L_\alpha(x),$$

where

$$L_\alpha(x) = \hat{V}\big(\hat{m}_h(x)\big) \big\{ (-2\log h)^{1/2} + (-2\log h)^{-1/2} (A - \chi_\alpha) \big\},$$

$A = -\log(2^{3/2}\pi)$ for the kernel (3.3) and $\chi_\alpha = \log\{-\log(1-\alpha)/2\}$.

Claeskens and Van Keilegom (2003) propose two confidence bands, the first of which has two differences from that of Xia. Firstly they omit the explicit bias correction, preferring a slight undersmoothing of the data. This is manifested in the technical assumption that $nh^5 \log n \to 0$ as $n \to \infty$, whereas the optimal choice of h for curve estimation has h of precise order $n^{-1/5}$. They also require specification of the

distribution of ϵ in order to define their three possible estimators of the variance of $\hat{m}_h(x)$. It is most natural for us to assume that $\epsilon \sim N(0, \sigma^2)$, though since there appears to be the possibility of heavy tails in the error distribution (cf. Figure 4), we used the most robust of the three estimators.

The main concern over both confidence bands is possible undercoverage in small and moderate samples. This is based on results in Hall (1979, 1991) demonstrating the slow rate of convergence to normal extremes. Claeskens and Van Keilegom demonstrate the undercoverage numerically, and argue in favour of a smoothed bootstrap procedure. An unusual feature of their proposal is that they bootstrap the leading term in an asymptotic expansion of the difference $\hat{m}_h(x) - m(x)$, rather than the difference itself, in order to reduce the computational burden.

Despite the theoretical and numerical justification for the bootstrap bands in the Claeskens and Van Keilegom paper, we found their width to be very sensitive to small changes in bandwidth. The numerical simulations in their paper used an optimally chosen bandwidth. The performance of the Xia bands and the other Claeskens and Van Keilegom bands was largely similar, although the Xia bands tended to be slightly wider, and the Claeskens and Van Keilegom bands occasionally became very narrow indeed. The bias correction was only significant at clear peaks and troughs of the regression function. Since the Xia bands appear to suffer less from undercoverage and do not require specification of the error distribution, we prefer the Xia bands for this problem.

Figure 2 shows the regression estimate together with 95% Xia confidence bands for all three oceans. The different blocks are indicated by different shades of grey for the confidence bands, and have been slightly separated so as to distinguish the different bands more easily. Although there is no guarantee that the estimates of the regression function will line up at the block boundaries, we found that in practice they were very close indeed. To appreciate better the local variation, we present the band for the Pacific ocean in the 0–0.68 Ma block on a finer scale in Figure 3. The block in Figure 3 has 277 data points, so undercoverage should not be much of a problem, but the bands in blocks with few data points should be treated with caution.

The results indicate higher levels of δ^{13}C in the North Atlantic than the other two oceans in the 0–10 Ma interval, with similar levels in all three oceans previously. There is a noticeable trough in the records for all three oceans between 15 Ma and 23 Ma. Oceanographers expect deep water to flow from regions of high δ^{13}C to regions

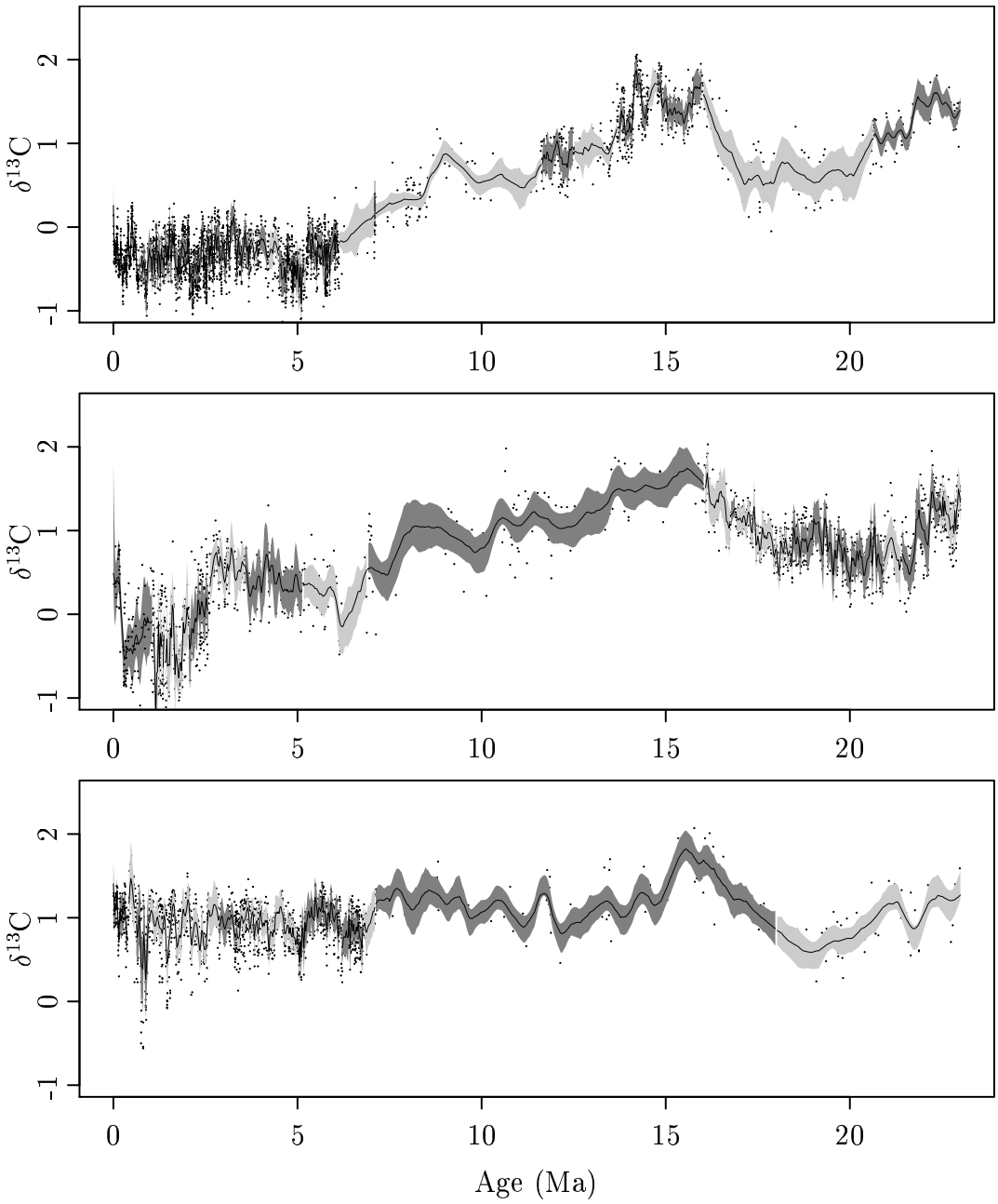

Figure 2: Regression estimate and 95% Xia confidence band for the Pacific (top), Southern (middle) and North Atlantic (bottom) oceans.

of low δ^{13}C. This is because deep waters collect decayed organic matter over time, thereby becoming enriched in ^{12}C and lowering the δ^{13}C value. Thus, a low δ^{13}C value is an indication that water is far from its surface source, and conversely a high δ^{13}C level indicates proximity to a source of deep ocean water. The relative levels of δ^{13}C in the three oceans are studied in greater detail in Section 5.

One of the advantages of using a local linear regression estimator rather than the local constant version is the fact that the bias is of the same order of magnitude in the interior region as in the boundary regions. We are able almost to eliminate boundary effects entirely, except at the 0 Ma boundary, by using some extra data from the neighbouring block in our calculations of the regression estimate and the confidence bands.

Figure 3 suggests the possibility of a periodic component in the data. In fact, geologists expect to see periodicities at the periods of perturbations of the Earth's orbit, namely $T_1 = 1.25$ Ma, $T_2 = 0.4$ Ma, $T_3 = 0.1$ Ma, $T_4 = 0.041$ Ma, $T_5 = 0.023$ Ma and $T_6 = 0.019$ Ma. We investigated various simple linear models for each block with some or all of the above periodicities, and with parameters fitted by least squares. In a few short blocks of high resolution data, such as the 0–0.68 Ma block for the Pacific data shown in Figure 3, the fit was reasonably good. However, in general the sparsity of data necessitated omitting some of the higher frequency terms in the model, and the fit was very poor. We do not pursue parametric models further here.

4 Checking the Model

Recall that our nonparametric model $Y = m(X) + \epsilon$ for the data $\{(x_i, y_i) : i = 1, \ldots, n\}$ in each block assumed that ϵ and X were independent. A recent discussion paper by Einmahl and Van Keilegom (2004) addresses precisely this assumption. They construct a hypothesis test based on a bivariate Kolmogorov–Smirnov statistic, whose asymptotic distribution under the null hypothesis of independence is that of the supremum of a particular Gaussian process.

We carried out the hypothesis test on each block (49 in total), and found the test statistic to be significant at the 5% level in three blocks, with one such block in each ocean. The statistic for the Pacific Ocean block was just significant at the 1% level, while the other two were not significant at this level. Since this significance rate

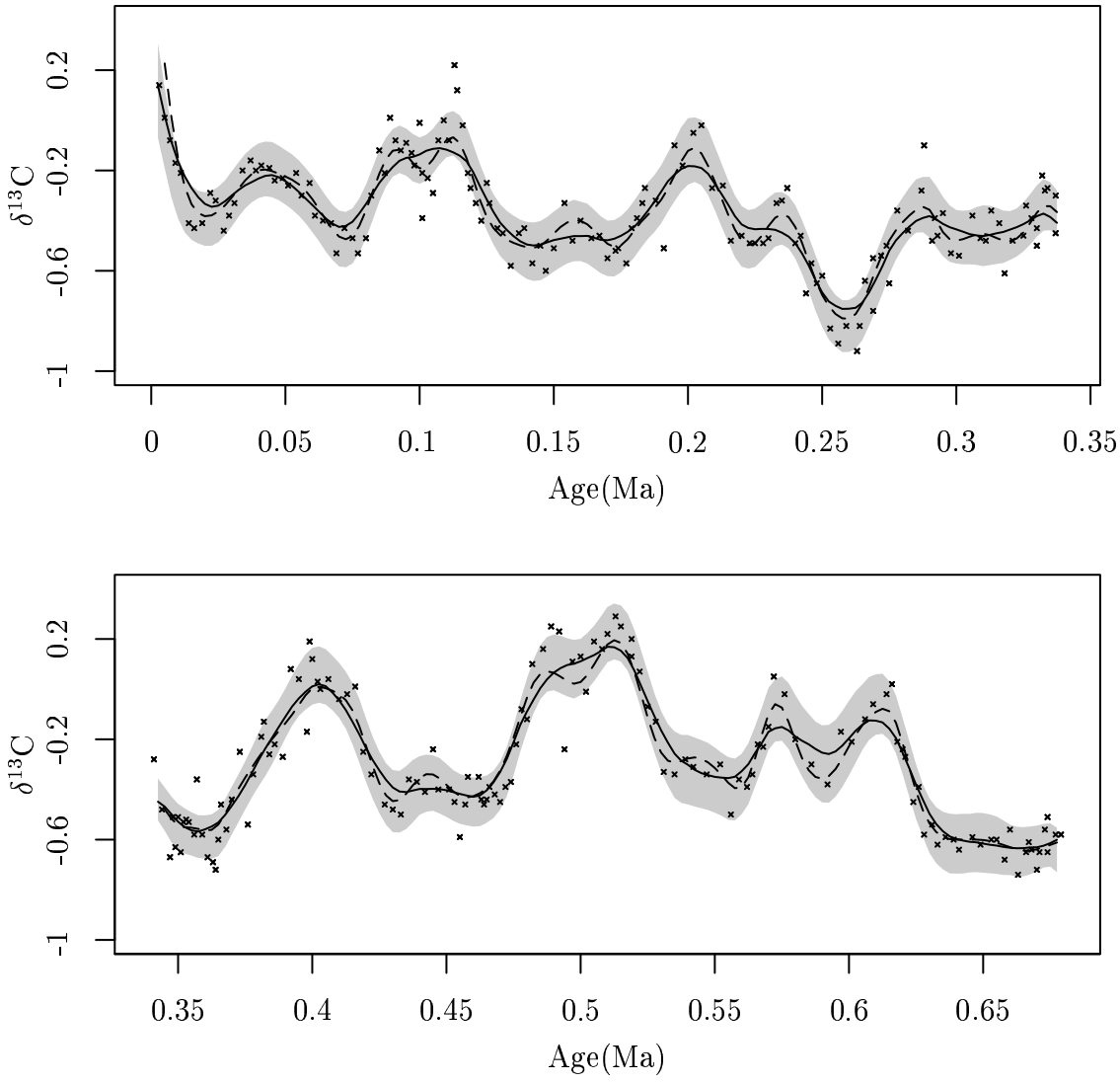

Figure 3: Regression estimate (solid) and Xia confidence band for the Pacific ocean data from 0-0.68 Ma. The dashed line is the output from the SIMEX algorithm discussed in Section 4.

is almost exactly what we would expect under the null hypothesis for each block, and since a Bonferroni adjustment for multiple testing would conclude that none of the test statistics was significant, we find no reason to doubt this assumption of our model. Several tests of the weaker null hypothesis of homoscedasticity, which focuses only on the conditional variance rather than the full conditional distribution of ϵ given X, are also available (Liero, 2003; Dette and Munk, 1998; Dette, 2002); see also Stute (1997).

To study the error distribution, we plotted the quantiles of the studentised residuals against the quantiles of the standard normal distribution. The quantile-quantile (qq) plots of original data set suggested heavy tails at both ends, but particularly at the negative end (cf. Figure 4 for the Pacific Ocean data). While such extreme measurements could be due to random variation, it is also possible that certain samples had become corrupted by a process called diagenesis. This occurs when water whose isotopic composition has been altered by bacteria percolates and then partially recrystallises the microfossil shell, and will tend to reduce the δ^{13}C value. Some such samples could easily be identified and removed by inspection, for instance when other measurements were available at almost identical ages. Another aid was the fact that the δ^{18}O level (the ratio of ^{18}O to ^{16}O compared to a standard) in each sample was also measured. If diagenesis has occurred, both isotope measurements will be affected. By fitting a regression estimate to the Oxygen isotope data in the same way, we removed those points for which both studentised residuals were less than -1.645, or both were greater than 1.645. The qq plot of the revised data set is also shown in Figure 4.

The qq plot of the revised data indicates that a normal distribution is quite a good fit for the error distribution, but suggests that the possibility of heavy tails remains. A bootstrap hypothesis test, developed by Neumeyer, Dette and Nagel (2004), may now be used to examine this possibility more formally.

It was mentioned in Section 2 that the data from different cores was pooled together. To investigate the possibility of a 'core effect', we compared two models for the fitted residuals. In the first, the fitted residuals were approximated by independent and identically distributed normal random variables, while in the second we allowed a random effect for core, fitted using `lme(Resid~1,random=~1|Core)` in R. Anova tests revealed non-significant test statistics at the 5% level for the Southern and North Atlantic Oceans, while the statistic for the Pacific Ocean was not significant after a correction for multiple testing. We therefore retain our model (3.1), though

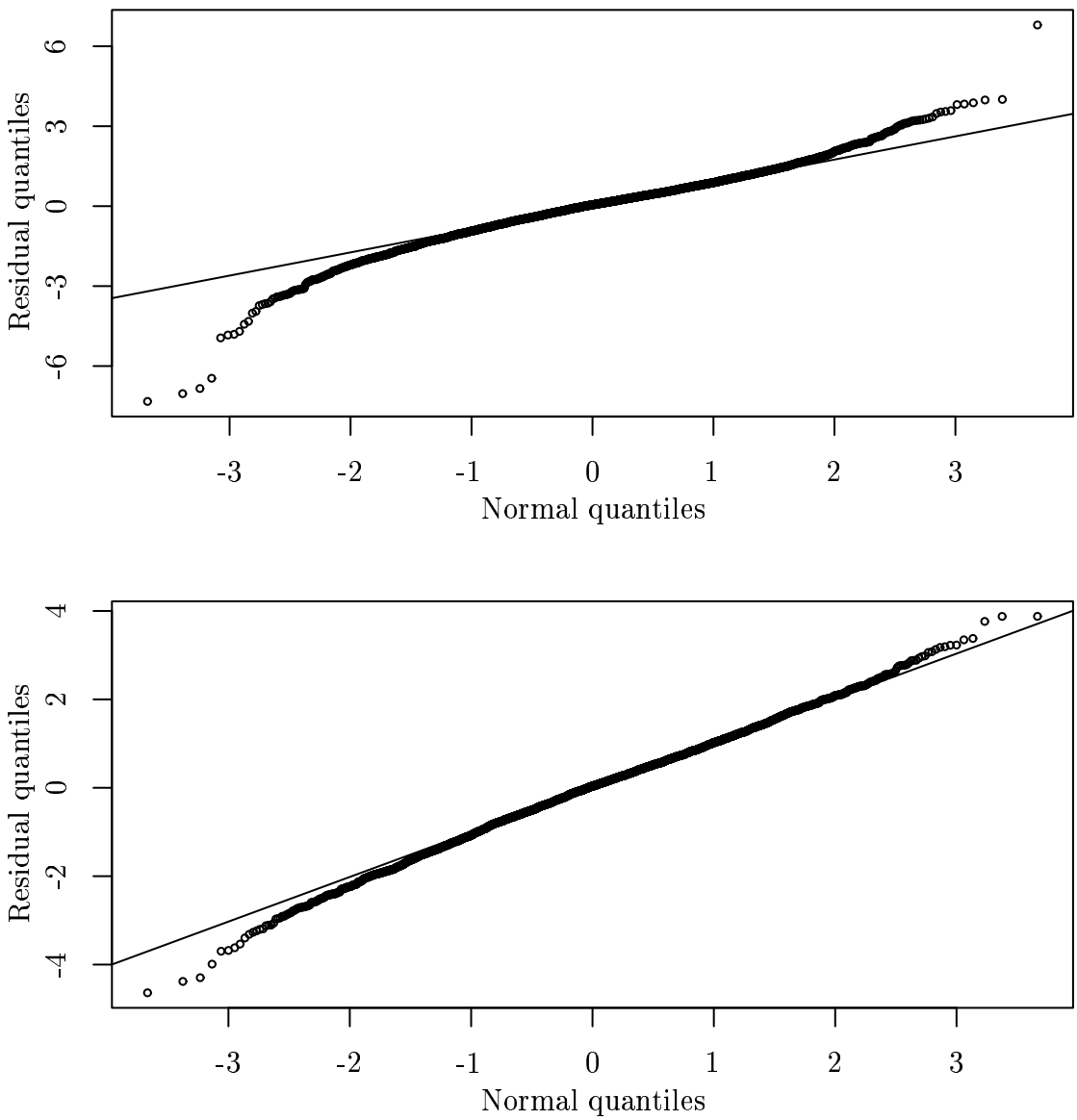

Figure 4: Plots comparing quantiles of the studentised residuals with quantiles of the standard normal distribution for the Pacific Ocean data both before removing samples suspected of contamination (top) and after (bottom). The straight lines go through the first and third quartiles.

since there is a relatively small overlap for the different age ranges covered by the cores, the possibility of core effects in other similar data sets should not be ruled out and would be an interesting topic for further study.

As well as the error in the response, there are also errors associated with the ages assigned to samples. These arise from several sources, in particular from the facts that the marker levels are not dated precisely, and that the assumption of constant sedimentation rate, which underlies the use of linear interpolation between marker events, is merely a good approximation rather than being exact. The dating methods employed result in dependent errors which increase with distance from the marker levels. Such difficulties mean that it is not realistic to be able to model the age errors accurately, and we are forced to make some simplifications before studying their effect.

We consider the model

$$X_i = W_i + U_i, \quad i = 1, \ldots, n, \tag{4.1}$$

where X_i is the age measured for the ith sample, W_i is the unobserved true age and $U_1, \ldots, U_n$ are independent and identically distributed $N(0, \tau^2)$ random variables. Expert opinion based on knowledge of the origin of each core, the sample resolution and expected periodicities suggested that taking $\tau = h$ was reasonable, and in fact probably conservative, i.e. too large, in most blocks. Since ageing errors are closely related to the sampling resolution of the core, and since the sampling resolution directly affects the bandwidth, this choice goes some way to incorporating a core-specific ageing error into the model. This is particularly the case in view of the relatively small overlap of the age ranges of different cores. The SIMEX algorithm attempts to adjust the regression curve estimate for the age-measurement error by first examining the effect of additional age-error noise and then extrapolating backwards to estimate the effect of no noise. The algorithm is as follows:

1. For $\lambda \in \Lambda = \{0, 1/4, 1/2, \ldots, 2\}$ and $b \in 1, \ldots, B = 50$, define modified ages $X_{ib}(\lambda) = x_i + \lambda^{1/2} U_{ib}$, where the U_{ib} are independent $N(0, \tau^2)$ random variables.

2. For $\lambda \in \Lambda$ and $b \in 1, \ldots, B$, fit the local linear kernel regression estimate $\hat{m}_h(x; \lambda, b)$ to the data $\big\{\big(X_{ib}(\lambda), y_i\big) : i = 1, \ldots, n\big\}$.

3. Let $\hat{m}_h(x; \lambda)$ be the sample mean of $\hat{m}_h(x; \lambda, b)$.

4. For each x, fit a quadratic curve by least squares through the points $\big\{\hat{m}_h(x, \lambda) : \lambda \in \Lambda\big\}$ considered as a function of λ. Extrapolate the curve back to $\lambda = -1$ to give the SIMEX estimator of $m(x)$.

In computing the regression estimates in step (b), we used a Gaussian kernel, so as to avoid problems where the largest gap in the modified ages happened to be larger than twice the kernel-window size. To illustrate the effect of the SIMEX algorithm, we performed a simulation in which a sample of $n = 100$ pairs (W_i, Y_i) were generated from a model with $W_i \sim \text{Beta}(1.1, 1.1)$ and $Y_i = m(W_i) + \epsilon_i$, where $m(x) = e^{-x}\sin(4\pi x)$ and $\epsilon_i \sim N(0, \sigma^2)$, with $\sigma = 0.05$. Next we set $X_i = W_i + U_i$, with U_i independent of W_i and having a $N(0, \tau^2)$ distribution with $\tau = 0.05$. The automatic bandwidth selection algorithm of Hengartner, Wegkamp and Matzner-Løber (2002) was applied to the observed data (X_i, Y_i), yielding the choice $h = 0.046$ (replication of the experiment verified that this value was fairly typical). Both the naive regression estimate and the SIMEX modification are shown in Figure 5, along with the true regression function $m(x)$.

From Figure 5, we see that the SIMEX estimate improves on the naive regression estimate at interior points (the boundary effects are not that critical for our purposes, for the reason given in the penultimate paragraph of Section 3). The effect of the SIMEX algorithm is much like bias correction: peaks in the regression estimate are made higher, while troughs are made lower. Intuitively, adding extra noise to the x-variable makes peaks and troughs less well-defined, so extrapolating the effect back to $\lambda = -1$ should clarify these features.

We applied the SIMEX estimator to each block of the δ^{13}C data separately, and the results for the 0–0.68 Ma block of the Pacific Ocean are shown in Figure 3. We found the SIMEX estimate to be very satisfactory in all blocks, suggesting that the model (4.1), though crude, is nevertheless effective.

5 Northern Component Water estimation

As was mentioned in the introduction, a quantity of particular interest to oceanographers is the history of the proportion of Northern Component Water (NCW). At a particular time x, this is defined as the proportion of the volume of water in the Southern Ocean originating from the North Atlantic, and is denoted by NCW(x). Assuming that all Southern Ocean water originates from either the North Atlantic or the Pacific, an estimate of this quantity may be derived as follows. At times x for which (in obvious notation), $m^{SO}(x)$ lies between $m^{PO}(x)$ and $m^{NA}(x)$, conservation

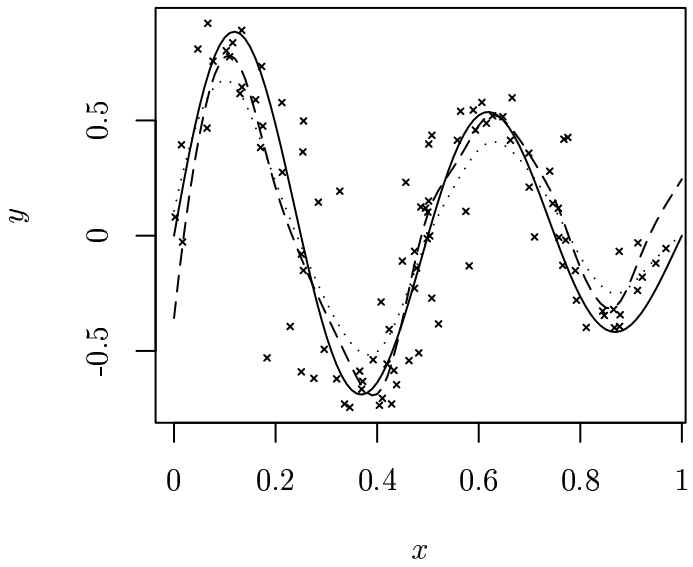

Figure 5: Simulated data showing the effect of the SIMEX algorithm. The lines show the true regression function (solid), the naive regression estimate (dotted) and the SIMEX estimate (dashed).

of mass gives that

$$\mathrm{NCW}(x)\, m^{NA}(x) + \{1 - \mathrm{NCW}(x)\} m^{PO}(x) = m^{SO}(x),$$

whence an estimate of $\mathrm{NCW}(x)$ is given by

$$\widehat{\mathrm{NCW}}(x) = \frac{\hat{m}_h^{SO}(x) - \hat{m}_h^{PO}(x)}{\hat{m}_h^{NA}(x) - \hat{m}_h^{PO}(x)}. \tag{5.1}$$

It is clear that this equation is meaningless unless $\mathrm{NCW}(x)$ lies between 0 and 1. It could therefore be argued that it would be preferable to modify the estimator to avoid this problem, though we did not do this as the quantity defined in (5.1) has been considered by several different authors, and is one with which oceanographers are familiar. It is also clear from (5.1) that $\widehat{\mathrm{NCW}}(x)$ is particularly sensitive to uncertainties in the regression estimates for each separate ocean. This is one reason why appropriate bandwidth choice in Section 3 was so important. We now apply the delta method to compute the asymptotic variance of $\widehat{\mathrm{NCW}}(x)$.

For the purposes of this calculation, we suppress the argument x, and denote the asymptotic variances of the three oceans by $V(\hat{m}_h^{SO})$, $V(\hat{m}_h^{PO})$ and $V(\hat{m}_h^{NA})$. The regression estimates are independent for different oceans and, subject to minor regularity conditions, are asymptotically normally distributed. It follows that the vector consisting of the numerator and denominator of (5.1) has an asymptotic bivariate normal distribution, with mean vector μ and covariance matrix Σ given by

$$\mu = \begin{pmatrix} m^{SO} - m^{PO} \\ m^{NA} - m^{PO} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} V(\hat{m}_h^{SO}) + V(\hat{m}_h^{PO}) & V(\hat{m}_h^{PO}) \\ V(\hat{m}_h^{PO}) & V(\hat{m}_h^{NA}) + V(\hat{m}_h^{PO}) \end{pmatrix}.$$

Applying Theorem A on p.122 of Serfling (1980), the delta method, to the function $g(x, y) = x/y$ then gives that $\widehat{\mathrm{NCW}}$ is asymptotically normally distributed with mean NCW and variance

$$\frac{A}{\{m^{NA} - m^{PO}\}^2} + \frac{B}{\{m^{NA} - m^{PO}\}^4} - \frac{2C}{\{m^{NA} - m^{PO}\}^3},$$

where $A = V(\hat{m}_h^{SO}) + V(\hat{m}_h^{PO})$, $B = \{m^{SO} - m^{PO}\}^2 \{V(\hat{m}_h^{NA}) + V(\hat{m}_h^{PO})\}$ and $C = \{m^{SO} - m^{PO}\} V(\hat{m}_h^{PO})$. We can therefore estimate the variance of $\widehat{\mathrm{NCW}}$ by replacing $V(\hat{m}_h^{PO})$ by $\hat{V}(\hat{m}_h^{PO})$, and replacing m^{PO} by $\hat{m}_h^{PO}$ etc..

Panel (a) of Figure 6 shows an estimate of the NCW proportion in the 0–8 Ma age range, with variability bands of width two estimated standard deviations (computed

using the above formula). The estimate is not meaningful in the 8-23 Ma age range, and even in certain parts of the range presented, the width of the bands indicates a large uncertainty in the estimate. Of course, our calculation of $\widehat{\mathrm{NCW}}(x)$ in (5.1) relies on the assumptions that the cores chosen accurately reflect the relevant deep water in each ocean and that there is no core effect, as discussed in Section 4.

There are several interesting features in the NCW estimate, which of course has time evolving from right to left. A major event affecting global climate in the last 8 Ma was the closure of the Panama Isthmus, which was a gateway for surface current flow between the equatorial Atlantic and Pacific Oceans. Prior to this closure, conditions in the North Atlantic were less suitable for generation of southbound deep ocean currents (Murdock, Weaver and Fanning, 1997), so one expects a significant long-term increase in the proportion of NCW after closure. The increase in heat transfer efficiency from the equator would result in warmer poles. Previous dates assigned to this event include 4.36 Ma (Billups, Ravelo and Zachos, 1998), 4.6 Ma (Haug and Tiedemann, 1998), 5–6 Ma (Lear, Rosenthal and Wright, 2003) and finally 6–6.6 Ma (Billups , 2002), whose approach based on ^{13}C gradients is most similar to ours. There remains uncertainty in the precise timing of closure of the Isthmus, but surface currents would appear to be restricted from about 4.5 Ma onwards (Haug and Tiedemann, 1998). This does not appear to correlate with a major change in the proportion of NCW. Instead, there is a large increase in the proportion of NCW from 6.2 Ma. We therefore examine a possible control of deep southbound water by the gateway around Iceland, the Greenland Scotland Ridge. This ridge forms a barrier to the passage of southbound water from the seas north of Iceland into the Southern Ocean.

From their study of older δ^{13}C data over 0–25 Ma, Wright and Miller (1996) show a strong negative correlation between the proportion of NCW and a quantity called residual depth, which is plotted in panel (b) of Figure 6. This is a measure of the effect of uplift from the Icelandic plume, a rising convection current formed within the Earth. If the plume is hot, it will cause uplift of the ocean floor, thereby restricting southward deep water flow. Wright and Miller conclude that the Icelandic plume has a strong controlling effect on the proportion of NCW. The residual depth record here is adjusted for the time it takes for uplift to reach the Denmark Straits from the centre of the plume. The Denmark Straits is the closest deep channel to the plume centre. The increase in the proportion of NCW at about 6 Ma correlates to a decrease in residual depth because of plume activity and suggests the Denmark Straits is the most important gateway for deep water in the North Atlantic.

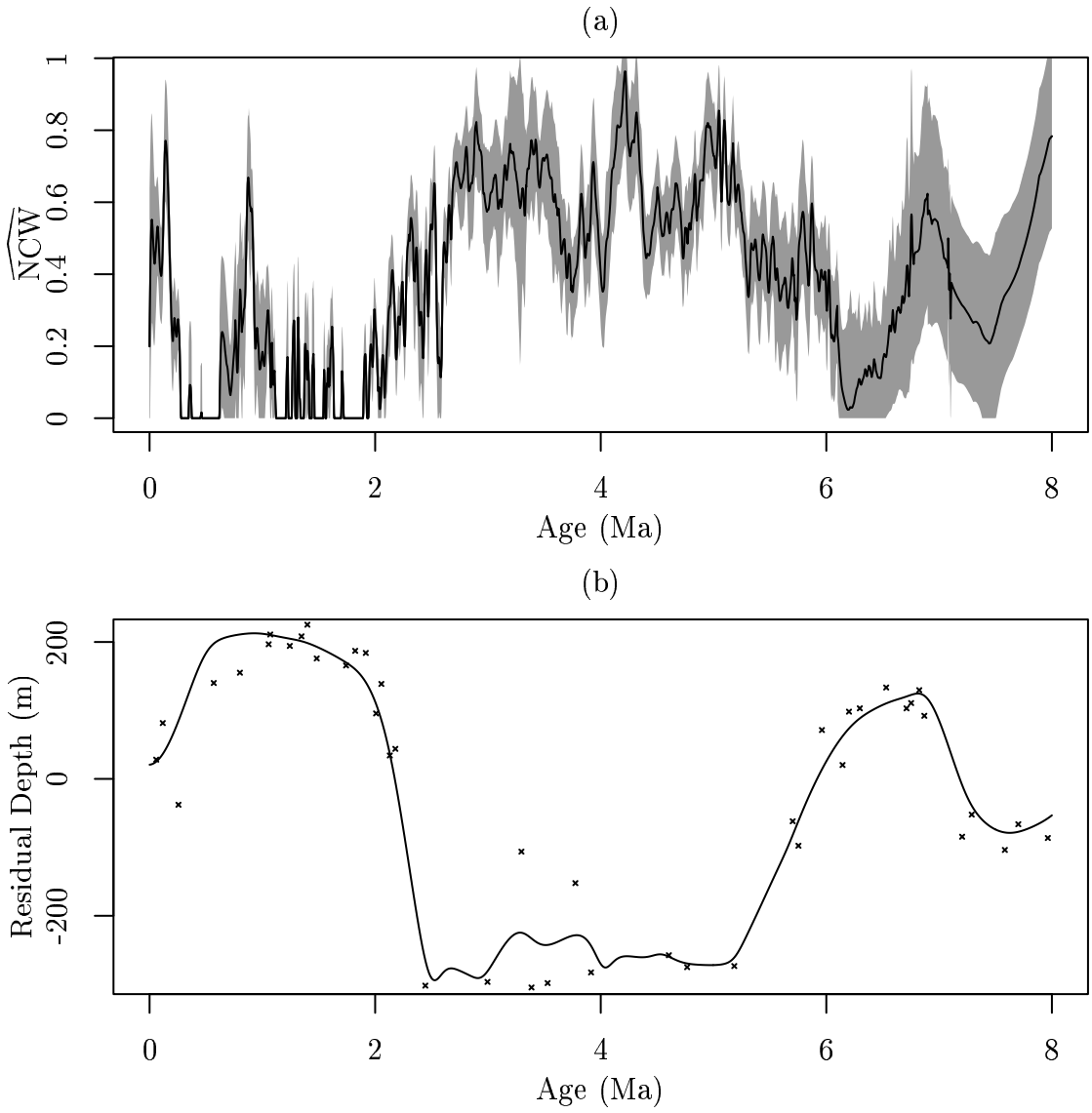

Figure 6: Panel (a) shows the Northern Component Water estimate for 0-8 Ma with variability bands of width two estimated standard deviations. Panel (b) gives the residual depth data together with a smoothed estimate.

A second broad-scale feature of the NCW record is the long-term decrease in the 2–2.8 Ma interval. This is in close agreement with other indicators of deep water production, such as poor carbonate preservation beginning at 2.8 Ma (Henrich et al., 2002). It also coincides with an increase in the areal extent of Northern Hemisphere Glaciation at about 2.7 Ma (Maslin et al., 1998). The decrease in the proportion of NCW correlates with an increase in residual depth and suggests plume uplift is responsible for the reduction in the proportion of NCW and perhaps for the increased glaciation in the Northern Hemisphere.

One final point to observe concerns the periodicities present in the NCW record. The 2.1–6.1 Ma interval in particular shows a strong 0.041 Ma periodicity, as well as periodicities of moderate amplitude at 1.25 Ma and 0.4 Ma. These are all known orbital periodicities of the Earth. In the 0-2.1 Ma interval, the NCW estimate is meaningless in places, which hinders spectral analysis, while in the 6.1-8 Ma interval, the resolution of the data is too low to detect these periodicities. Corresponding Oxygen isotope records appear to show only very weak 1.25 Ma and 0.4 Ma periodicities (Zachos et al., 2001), and this unexpected difference between the isotope records for the different elements is an interesting avenue for further research.

6 Concluding remarks

In this article, we have shown the effectiveness of modern nonparametric regression techniques, in particular local linear kernel estimation, for handling a variety of problems arising in practice. An appreciation of the uncertainty of the estimates was especially important so as not to overstate the conclusions of the study. Nevertheless, the uncertainty in the 0-8 Ma interval is sufficiently small to suggest a number of significant conclusions.

In regions of high sampling resolution, the response variations present at the Earth's orbital periods suggest the basis of a parametric model. Though there were too few such regions in our data to warrant a more detailed analysis, this should be borne in mind in the light of any data collected in the future. Undoubtedly, the coefficients controlling each periodic component would not be constant over time.

Finally, we mention that the methods employed in this paper could be applied to a variety of other related problems. For instance, the inclusion of data from the Indian

Ocean may help to date the closure of the Tethys Ocean in the Mediterranean region, which has been suggested as a source of deep ocean currents from 5–23 Ma (Savin et al., 1981).

Acknowledgements: We are grateful to Katharina Billups and Helen Pfuhl for providing us with recent, unpublished data, and to Simon Crowhurst, Nick McCave, Nick Shackleton and Nicky White for helpful conversations. Finally, we are grateful for the thorough reviews and constructive comments of the anonymous referees.

References

Bickel, P. J. and Rosenblatt, M. (1973), *On some global measures of the deviations of density function estimates*, Ann. Statist., **1**, 1071–1095.

Billups, K., Ravelo, A. C. and Zachos, J. C. (1997), *Early Pliocene deep-water circulation: stable isotope evidence for enhanced northern component deep water*, Proc. Ocean Drilling Program, Scientific Results, **154**, 319–330.

Billups, K., Ravelo, A. C. and Zachos, J. C. (1998), *Early Pliocene climate: A perspective from the western equatorial Atlantic warm pool*, Paleoceanography, **13**, 459–470.

Billups, K., Channell, J. E. T. and Zachos, J. C. (2002), *Late Oligocene to early Miocene geochronology and paleoceanography from the subantarctic South Atlantic*, Paleoceanography, **17**, 4-1–4-11.

Billups, K. (2002), *Late Miocene through early Pliocene deep water circulation and climate change viewed from the sub-Antarctic South Atlantic*, Palaeogeography, Palaeoclimatology, Palaeoecology, **185**, 287–307.

Bowman, A. W. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis*, Oxford University Press.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, Chapman and Hall, London.

Claeskens, G. and Van Keilegom, I. (2003), *Bootstrap confidence bands for regression curves and their derivatives*, Ann. Statist., **31**, 1852–1884.

Cook, J. R. and Stefanski, L. A. (1994), *Simulation-extrapolation estimation in parametric measurement error models*, J. Amer. Statist. Assoc., **89**, 1314–1328.

Dette, H. (2002), *A consistent test for heteroscedasticity in nonparametric regression based on the kernel method*, J. Statist. Plann. Infer., **103**, 311–329.

Dette, H. and Munk, A. (1998), *Testing heteroscedasticity in nonparametric regression*, J. Roy. Statist. Soc., Ser. B, **60**, 693–708.

Einmahl, J. H. and Van Keilegom, I. (2004), *Goodness-of-fit tests in nonparametric regression*, CentER Discussion Paper No. 2004–12, Institut de Statistique, Université Catholique de Louvain. http://ssrn.com/abstract=556975

Fan, J. and Gijbels, I. (1995), *Data-driven selection in local polynomial fitting: variable bandwidth and spatial adaptation*, J. Roy. Statist. Soc., Ser. B, **57**, 371–394.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, Chapman and Hall, London.

Gradstein, F., Ogg, J. and Smith, A. G. (2004), *A geologic time scale 2004*, Cambridge University Press.

Hall, P. (1979), *On the rate of convergence of normal extremes*, J. Appl. Probab., **16**, 433–439.

Hall, P. (1991), *On convergence rates of suprema*, Probab. Theor. Rel. Fields., **89**, 447–455.

Hall, P. and Titterington, D. M. (1988), *On confidence bands in nonparametric density estimation and regression*, J. Mult. Anal., **29**, 163–179.

Härdle, W. and Kelly, G. (1987), *Nonparametric kernel regression estimation – optimal choice of bandwidth*, Statistics, **18**, 21–35.

Hengartner, N. W., Wegkamp, M. H. and Matzer-Løber, E. (2002), *Bandwidth selection for local linear regression smoothers*, J. Roy. Statist. Soc., Ser. B., 791–804.

Henrich, R., Baumann, K.-H., Huber, R. and Meggers, H. (2002), *Carbonate preservation records of the past 3 Myr in the Norwegian–Greenland Sea and the northern North Atlantic: implications for the history of NADW production*, Marine Geology, **184**, 17–39.

Haug, G. H. and Tiedemann, R. (1998), *Effect of the formation of the Isthmus of Panama on Atlantic Ocean thermohaline circulation*, Nature, **393**, 673–676.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998), *Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion*, J. Roy. Statist. Soc., Ser. B, **60**, 271–293.

Knafl, G., Sacks, J. and Ylvisaker, D. (1985), *Confidence bands for regression functions*, J. Am. Statist. Assoc., **80**, 683–691.

Lear, C. H., Rosenthal, Y. and Wright, J. D. (2003), *The closing of a seaway: ocean water masses and global climate change*, Earth and Planetary Science Lett., **210**, 425–436.

Liero, H. (2003), *Testing heteroscedasticity in nonparametric regression*, J. Nonparam. Statist., **15**, 31–51.

Lourens, L. J., Hilgen, F. J., Laskar, J., Shackleton, N. J. and Wilson, D. (2004), *The Neogene Period*, in *A Geologic time scale 2004* (Gradstein, Ogg and Smith Eds), Cambridge Univ. Press.

Maslin, M. A., Li, X. S., Loutre, M.-F. and Berger, A. (1998), *The contribution of orbital forcing to the progressive intensification of Northern Hemisphere Glaciation*, Quaternary Science Reviews, **17**, 411–426.

Mix, A. C., Pisias, N. G., Rugh W., Wilson J., Morey A. and Hagelberg T. K. (1995), *Benthic foraminifer stable isotope record from Site 849 (0-5 Ma); local and global climate changes*, Proc. Ocean Drilling Program; Scientific Results, **138**, 371–412.

Murdock, T. Q., Weaver, A. J. and Fanning, A. F. (1997), *Paleoclimatic response of the closing of the Isthmus of Panama in a coupled ocean-atmosphere model*, Geophysical Research Lett., **24**, 253–256.

Neumeyer, N., Dette, H. and Nagel, E.-R. (2004), *Bootstrap tests for the error distribution in linear and nonparametric regression models*, Austral. & New Zealand J. Statist., to appear.

Ogden, R. T. (1996), *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhäuser, Boston.

Oppo, D. W. and Fairbanks, R. G. (1987), *Variability in the deep and intermediate water circulation of the Atlantic Ocean during the past 25,000 years: Northern Hemisphere modulation of the Southern Ocean*, Earth Planet. Sci. Lett. **86**, 1–15.

Rice, J. (1984), *Bandwidth choice for nonparametric regression*, Ann. Statist., **12**, 1215–1230.

Ruppert, D., Shether, S. J. and Wand, M. P. (1995), *An effective bandwidth selector for local least squares regression*, J. Am. Statist. Assoc., **90**, 1257–1270.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

Savin, S. M., Douglas, R. G., Keller, G., Killingley, J. S., Shaughnessy, L., Sommer, M. A., Vincent, E. and Woodruff, F. (1981), *Miocene benthic foraminiferal isotope records: a synthesis*, Marine Micropaleontology, **6**, 423–450.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Shackleton, N. J., McCave, I. N. and Weedon, G. P. (Eds) (1999), *Astronomical (Milankovitch) Calibration of the geological time scale*, Phil. Trans. Roy. Soc. London, Ser. A, **357**, 1733–2007.

Stute, W. (1997), *Nonparametric model checks for regression*, Ann. Statist., **25**, 613–641.

Sun, J. and Loader, C. R. (1994), *Simultaneous confidence bands for linear regression and smoothing*, Ann. Statist., **22**, 1328–1345.

Venz, K. A. and Hodell, D. A. (2002), *New evidence for changes in Plio-Pleistocene deep water circulation from Southern Ocean ODP Leg 177 Site 1090*, Palaeogeography, Palaeoclimatology, Palaeoecology, **182**, 197–220.

Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.

Wright, J. D. and Miller, K. G. (1996), *Control of North Atlantic deep water circulation by the Greenland-Scotland ridge*, Paleoceanography, **11**, 157–170.

Xia, Y. (1998), *Bias-corrected confidence bands in nonparametric regression*, J. Roy. Statist. Soc., Ser. B, **60**, 797–811.

Zachos, J., Pagani, M., Sloan, L., Thomas, E. and Billups, K. (2001), *Trends, rhythms, and aberrations in global climate 65 Ma to present*, Science, **292**, 686–693.