


All age–depth models are wrong, but are getting better

The Holocene
1–10
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0959683616675939
hol.sagepub.com


Mathias Trachsel^{1,2} and Richard J Telford^{1,3}

Abstract

The construction of accurate age–depth relationships and a realistic assessment of their uncertainties is one of the fundamental prerequisites for comparing and correlating late Quaternary stratigraphical proxy records. Four widely used age–depth modelling routines – CLAM, OxCal, Bacon and Bchron – were tested using radiocarbon dates simulated from varved sediment stratigraphies. All methods produce mean age–depth models that are close to the true varve age, but the uncertainty estimation differs considerably among models. Age uncertainties are usually underestimated by CLAM, whereas age uncertainties produced by Bchron are often too large. With OxCal and Bacon, the setting of model-specific parameters influences the estimated uncertainties, which vary from too large to too small. The variability of sediment accumulation rates is underestimated by CLAM but overestimated by Bacon and Bchron. Bayesian age–depth models mainly improve the assessment of uncertainties of age–depth models.

Keywords

age–depth modelling, Bayesian inference, sediments, uncertainty

Received 23 April 2016; revised manuscript accepted 30 August 2016

Introduction

Accurate age–depth models are essential for comparing proxy records. With the exception of incrementally dated archives, the age–depth relationship is typically estimated from a relatively small number of dated levels using an age–depth model. Telford et al. (2004a) compared different age–depth modelling techniques for radiocarbon-dated lake sediments and found that uncertainty is often underestimated and the errors surprisingly large, especially when there are few dated levels. Over the last decade, considerable advances have been made in age–depth modelling: the once common intercept method for calibrating radiocarbon dates (Telford et al., 2004b) is now scarcely used; Monte Carlo–based uncertainty estimates for age–depth models (Bennett, 1994) are more widely used due to the availability of convenient code (Blaauw, 2010); and Bayesian age–depth modelling techniques (Blaauw and Christen, 2011; Bronk Ramsey, 2008; Haslett and Parnell, 2008) have been developed.

The three Bayesian age–depth modelling procedures are based on different depositional models that seek to mimic sediment deposition processes, have different parameters that need to be set and the resulting age–depth models and uncertainties are different. There is currently little guidance as to which of the three Bayesian age–depth modelling procedures performs best or how the parameters should be set. This study aims to fill this gap and help users choose a method and parameterise their age–depth modelling routines. We summarise the advantages and disadvantages of the different age–depth modelling routines.

This study follows the methods developed by Telford et al. (2004a): age–depth models are fitted using the different procedures to radiocarbon dates simulated from varved sediment stratigraphies and then compared with the varve age–depth relationship to gauge their performance. We tested three Bayesian age–depth modelling routines – OxCal (Bronk Ramsey, 2008),

Bchron (Haslett and Parnell, 2008) and Bacon (Blaauw and Christen, 2011) – and for comparison with classical age–depth modelling methods, we also tested CLAM (Blaauw, 2010). We tested the methods on three varved stratigraphies with different properties using different numbers of simulated radiocarbon dates and different settings for parameters in the Bayesian age–depth modelling routines. All age–depth modelling routines return best estimate age–depth models along with their uncertainties. We tested the accuracy of the best estimate age–depth models and also assessed how realistic the uncertainty estimates are. In a final step, we compared the estimated accumulation rates, which are necessary to calculate sediment fluxes or pollen accumulation rates (e.g. Jeffers et al., 2015), with the varve-based accumulation rates.

Methods

Age–depth modelling

Varved lake sediments are especially suited for testing age–depth modelling routines since the varve stratigraphy provides realistic sedimentation rates. Although varve counting has uncertainties, we treated varve ages as perfect. We used three

¹Department of Biology, University of Bergen, Norway

²Currently at Department of Geology, University of Maryland, College Park, USA

³Bjerknes Centre for Climate Research, University of Bergen, Norway

Corresponding author:

Mathias Trachsel, Department of Geology, University of Maryland, College Park, College Park, MD 20742-4454, USA.
Email: mtrachs@umd.edu

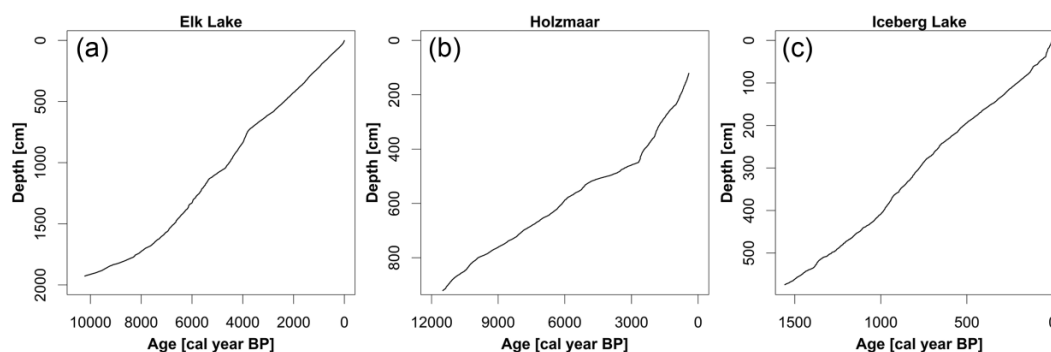


Figure 1. Varve age–depth relationships for (a) Elk Lake, (b) Holzmaar and (c) Iceberg Lake.

varved sediment sequences: Holzmaar (Zolitschka et al., 1998) and Elk Lake (Bradbury et al., 1993) as in Telford et al. (2004a) and Iceberg Lake (Loso, 2009). Holzmaar and Elk Lake span the entire Holocene, whereas the sequence of Iceberg Lake only covers the last 1500 years. The three sediment sequences are shown in Figure 1. Between 6 and 41 dates were equally spaced along the cores. These dates with true varve ages were then decalibrated using the IntCal13 calibration curve (Reimer et al., 2013) provided with the CLAM script for age–depth modelling (Blaauw, 2010). To simulate the uncertainty of these dates, a number drawn from a normal distribution with mean 0 and standard deviation of 30 years was added to the real radiocarbon age, as 30 years is the typical uncertainty of accelerator mass spectrometry (AMS) radiocarbon dates. These radiocarbon ages and the associated depths were passed to four age–depth modelling routines: (1) the classical age–depth modelling routine (CLAM; Blaauw, 2010), (2) OxCal (Bronk Ramsey, 2008), (3) Bchron (Haslett and Parnell, 2008) and (4) Bacon (Blaauw and Christen, 2011). All these age–depth modelling routines have internal procedures for calibrating radiocarbon dates.

The first step in constructing age–depth models with CLAM (Blaauw, 2010) is to calibrate all the radiocarbon dates in the sequence. Then, a single year in the probability density function (PDF) of each calibrated radiocarbon date is sampled according to the probability density. Interpolation between these years is then possible with a variety of regression-based techniques. The latter two steps are repeated many times using a Monte Carlo procedure. In this study, we chose a smoothing spline model with smoothing parameter set to 0.4. This procedure was repeated 1000 times to get an ensemble of possible age–depth models (i.e. a multitude of internally consistent age–depth models). The smoothing parameter was chosen arbitrarily and is a trade-off between models that do not fit the data well and ones that are too flexible. Since PDFs of calibrated radiocarbon dates at different depths overlapped at Iceberg Lake and CLAM does not have a monotonicity constraint (models with inversions are removed after their construction by CLAM), CLAM was only applied to sequences from Holzmaar and Elk Lake.

In contrast to classical age–depth modelling techniques, the Bayesian routines model the depositional process explicitly, and this process is monotonic.

OxCal has four options for modelling depositional processes: the D- and V-sequence models are designed for situations where the age-gap between two points is exactly (D, tree rings) or approximately (V, varves) known; the U-sequence model assumes a uniform depositional process; and the P-sequence model allows for variable sediment accumulation. We chose the P-sequence model which models sediment deposition as a Poisson process. This depositional model is analogous to a rain gauge where Poisson-distributed raindrops fall into the gauge leading to discrete increases in the water level and not to a continuous increase in

water level (Bronk Ramsey, 2008). The number of depositional events, raindrops in the analogy, per unit length is controlled by a parameter k . The higher k is, the smaller the depositional events are and the more uniform the modelled sediment deposition is. With a small k , more variable sediment deposition can be modelled (Bronk Ramsey, 2008).

Choosing a reasonable value for k is not a trivial task; Bronk Ramsey (2008) describes how k can, in principle, be determined from the variability in thickness of horizons in a profile. In this study, k was set to six different values: 0.1, 0.25, 0.5, 1, 2 and 3. It is also possible to set a prior distribution for k and let the data determine the most probable values of k (the posterior distribution of k ; Bronk Ramsey et al., 2010). For this purpose, a nominal k_0 is set and then a prior for $\log_{10}(k/k_0)$ is defined (Bronk Ramsey and Lee, 2013). We used a prior uniformly distributed between -2 and 1 , that is, $\log_{10}(k/k_0) \sim U(-2, 1)$, with $k_0 = 1$, so k varies between 0.01 and 10.

The quality of OxCal age–depth models is expressed by calculating an agreement index (A). The agreement index for individual radiocarbon ages is calculated using the PDF of the calibrated radiocarbon age and the PDF of the ages at the same depth when the depositional process has been modelled. The agreement index is defined as the ratio between the integral of the product of the PDF of the radiocarbon age and the distribution of modelled ages and the integral of the square of the PDF of the radiocarbon date (Bronk Ramsey, 1995). If the probability density of both distributions is high simultaneously, then usually $A > 100\%$, whereas if the probability density of both distributions is high at different ages, then usually $A < 100\%$. For the entire age–depth model, an overall agreement index is calculated as a modified geometric mean of all the individual agreement indices (using the reciprocal of the square root of the number of individual agreement indices, instead of the reciprocal of the number of individual agreement indices). Overall, agreement indices higher than 60% are considered as indicative of acceptable models.

Bchron and Bacon are based on modelling piecewise linear accumulations. In Bchron, ‘the process is piecewise linear, based on additive independent gamma increments arriving in a Poisson fashion’ (Haslett and Parnell, 2008). The total number of increments between dates is drawn from a Poisson distribution, with time intervals and height increments between steps drawn from two gamma distributions. Consecutive intervals between steps and increments are independent, allowing abrupt changes in accumulation rates.

In Bacon (Blaauw and Christen, 2011), the accumulation rates of the linear segments are controlled by a gamma autoregressive semi-parametric model (Blaauw and Christen, 2011): the accumulation rate in the current segment depends on the accumulation rate in the previous segment. For this model, two prior distributions have to be defined: the accumulation rate (expressed in yr cm^{-1}) and the autocorrelation of the accumulation rates between

piecewise linear segments (memory). Both of these distributions (gamma distribution for the accumulation rate and beta distribution for the accumulation memory) are defined by a parameter for the mean and a parameter for the shape. Additionally, the length of the linear segments must be defined. We set the mean distribution of the accumulation rate prior to the total number of varve years divided by the core length (cm) and the shape parameter to 1.4, the default value suggested by Goring et al. (2012), for Elk Lake and Iceberg Lake, and to 3 for Holzmaar as the Markov Chain Monte Carlo (MCMC) algorithm converged slowly in preliminary tests using the default of 1.4. For Holzmaar, we subsequently had to change the distribution of the radiocarbon dates from a t -distribution to a normal distribution because some of the correct dates were considered outliers by the algorithm. The two parameters of the prior distribution of the accumulation autocorrelation distribution were left at default (0.7 and 4). We ran Bacon models with four different segment lengths: 5, 10, 15 and 20 cm.

We tested the effect of changing the prior for the accumulation rate (changing the shape of the accumulation rate prior) and changing segment length in Bacon. We ran this experiment on the sediments of Holzmaar using 21 radiocarbon dates. The two parameters were set to two different levels resulting in four possible combinations of parameters. The levels were 1.4 (default) and 3.5 for the accumulation shape and 5 and 15 cm for the segment length. This experiment was run nine times, with nine different sets of radiocarbon dates. The influence of the parameters on the proportion of varves between the 16.6% and 83.3% quantiles was assessed by two-way analysis of variance (ANOVA), with accumulation shape and segment length being treated as factors. To remove the effects of the different sets of radiocarbon ages, we compared the values obtained for the four combinations of factor levels to their mean per set of radiocarbon dates. We did not test the effect of the distribution for accumulation autocorrelation, as preliminary tests showed that the accumulation persistence prior had little influence on the age–depth models provided all prior probabilities were greater than zero.

Assessment of age–depth models

To assess the quality of the age–depth models, we used four approaches:

1. As scientists are usually interested in having one mean age–depth model, we compared the mean age–depth model to the true varve-based age and calculated the root mean square error of prediction (RMSEP). The four age–depth modelling routines employed produced ensembles or PDFs of age–depth models. Therefore, we tested how well these ensembles represent the true age known from varve counting.

CLAM, Bacon and Bchron produce a multitude of internally consistent age–depth models, whereas OxCal only explicitly returns ages at dated and other selected depths. However, OxCal returns a PDF of probable ages at each depth. This PDF is sufficient to implement the second and third methods we used to assess the quality of age–depth modelling routines.

2. As a measure of spread, we assessed the proportion of radiocarbon dates that have their varve age within the uncertainty of the modelled date, given by the 16.6% and 83.3% quantiles of the ensembles of age–depth models. Hence, 66.7% of the true varve ages should lie within these bounds. If age–depth modelling routines are too pessimistic, ensembles will be too wide and more than 66.7% of the real ages will be found between the 16.6% and 83.3% quantiles, whereas over-optimistic models will generate narrow ensembles with too few real ages in this interval.

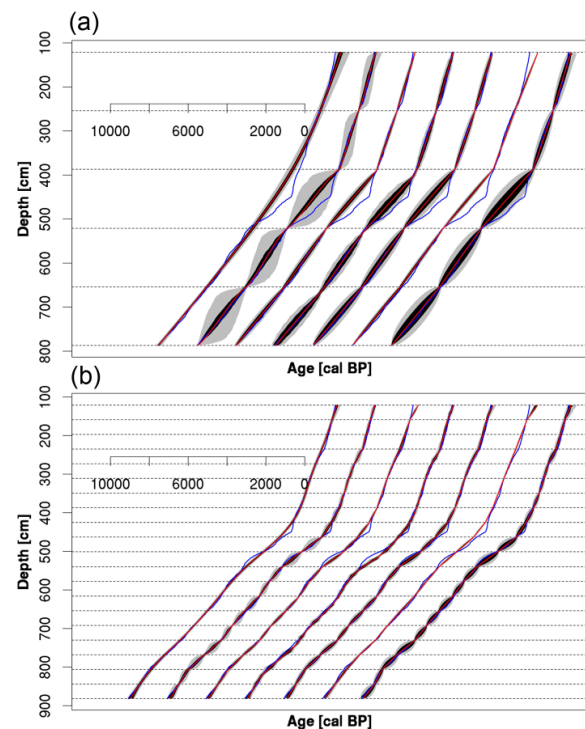


Figure 2. Age–depth models of Holzmaar with (a) 6 and (b) 21 radiocarbon dates. From left to right: CLAM, Bchron and Bacon, segment length 10 cm; Bacon, segment length 20 cm; OxCal $k = 0.25$, OxCal $k = 2$ and OxCal with prior for k . Red line: mean of age–depth model ensembles. Blue line: varve ages. Grey shading represents the 2.5% and 97.5% quantiles of the modelled ages, and black shading represents the 16.6% and 83.3% quantiles of the modelled ages. Dashed horizontal lines indicate dated depths.

3. We also used the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007; Hersbach, 2000) to assess the quality of age–depth models. The CRPS is a verification tool for probabilistic predictions. It is a proper scoring rule and is, therefore, expected to be optimal (lowest) for the best model. The CRPS considers both the uncertainty and the coverage probability of a probabilistic forecast.
4. In addition to the mean age–depth model and uncertainties of the age–depth models, scientists are often interested in sediment accumulation rates to calculate pollen accumulation rates or rates of change in sedimentological variables. We compared the number of varves contained in 4 cm sections with accumulation rates obtained using age–depth modelling algorithms (accumulation rates sensu Blaauw and Christen, 2011). Since OxCal does not explicitly model the age–depth relationship at each depth but only at dated and some selected depths, this test was only possible for CLAM, Bacon and Bchron (OxCal with all depths set to produce ensembles failed to converge).

For the four quality measures, we only used those parts of the cores that were covered by simulated radiocarbon ages; hence, we did not test the ability of the age–depth modelling routines to extrapolate.

Results

Age–depth models for Holzmaar are shown in Figure 2, illustrating the performance of the different age–depth modelling routines, the effect the number of radiocarbon dates has and the effect of the different settings for segment length in Bacon and k in

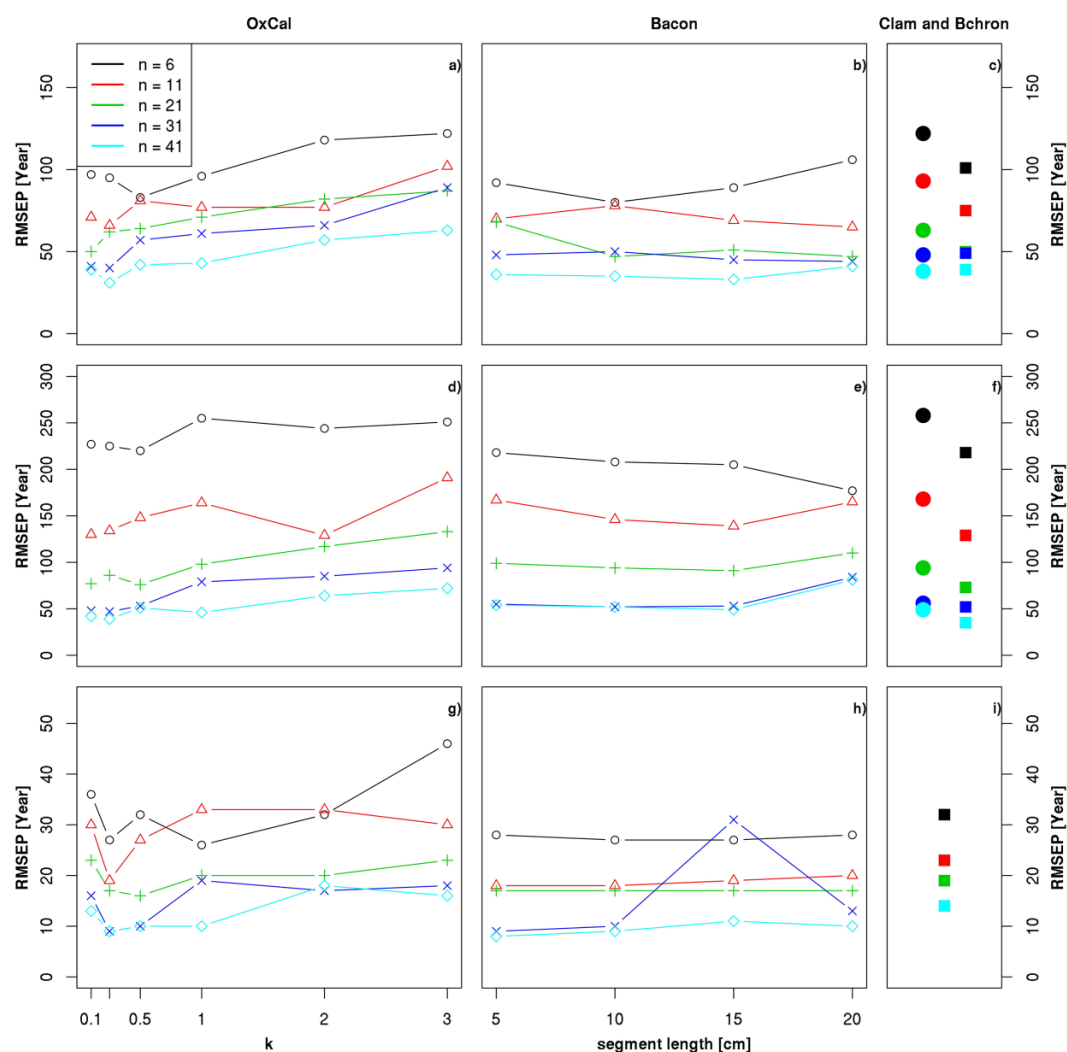


Figure 3. Root mean square error of prediction (RMSEP) of age–depth modelling routines. Left panels: RMSEP of OxCal models as a function of k for (a) Elk Lake, (d) Holzmaar and (g) Iceberg Lake. Middle panels: RMSEP of Bacon models as a function of segment length for (b) Elk Lake, (e) Holzmaar and (h) Iceberg Lake. Right panels: RMSEP of CLAM (circles) and Bchron (squares) for (c) Elk Lake, (f) Holzmaar and (i) Iceberg Lake (only Bchron). All colours as in the legend for (a).

OxCal. For all modelling routines, the mean age model gets closer to the true varve age when more radiocarbon dates are available. Uncertainties increase between dated levels for Bayesian age–depth models but remain constant when using CLAM. Uncertainties are largest for Bchron and OxCal with priors set for k . For Bacon, a longer segment length produces wider uncertainty bounds that can take clearly unrealistic shapes with many radiocarbon dates (Figure 2b). For OxCal, the error bands depend mainly on k , with larger k producing narrower error bands.

The RMSEPs for the four age–depth modelling routines are given in Figure 3. RMSEPs are lowest for age–depth models based on 31 or 41 radiocarbon dates. RMSEP of Bacon and OxCal are shown as a function of segment length and k , respectively: for OxCal, RMSEP increases with higher k . RMSEP of CLAM is generally higher than RMSEP of Bchron, OxCal with optimal k and Bacon with optimal segment length. We did not test different settings of CLAM since exhaustive tests of CLAM were beyond the scope of this study.

Figure 4 shows the proportion of varves with their true age within the bounds of the 16.6% and 83.3% quantiles of age–depth model ensembles. For CLAM, the ensembles of age–depth models are too narrow; fewer than the expected 66.7% of the true varve ages are between the quantile bounds of the ensemble of age–depth models. Bchron models have more than the expected 66.7% of the true varve ages between the quantile bounds. The parameter k has a systematic influence on the age–depth models

produced by OxCal: high k ensembles are too narrow, whereas very low values of k are too wide. Models from OxCal with a prior set for k are also pessimistic for all lakes and number of radiocarbon ages (Table 1), which results in a visual resemblance between Bchron models and OxCal models with a prior set for k . Changing the segment length in Bacon age–depth models influences the proportion of varves with their true age between the 16.6% and 83.3% quantiles, but not systematically. For Holzmaar, some of the segment lengths prescribed when running Bacon were longer than the distance between the radiocarbon dates when using 31 or 41 radiocarbon samples, which causes increased RMSEP, overly narrow age–depth model ensembles and increased RMSEP for these age–depth models.

For Iceberg Lake, where PDFs of calibrated radiocarbon ages at different depths overlap, age–depth model ensembles produced by Bchron are no longer too wide, in contrast to most models produced by Bacon which are too wide.

CRPS values (Figure 5) decrease with increasing number of ^{14}C dates, which is indicative of better models. For OxCal, CRPS changes systematically as a function of k : increasing with increasing k . CRPS is largely unaffected by the segment length using Bacon. CRPS of CLAM is generally higher than CRPS of the Bayesian age–depth modelling routines. The difference in CRPS between the methods decreases with increasing number of radiocarbon dates. The lowest CRPS for each number of radiocarbon

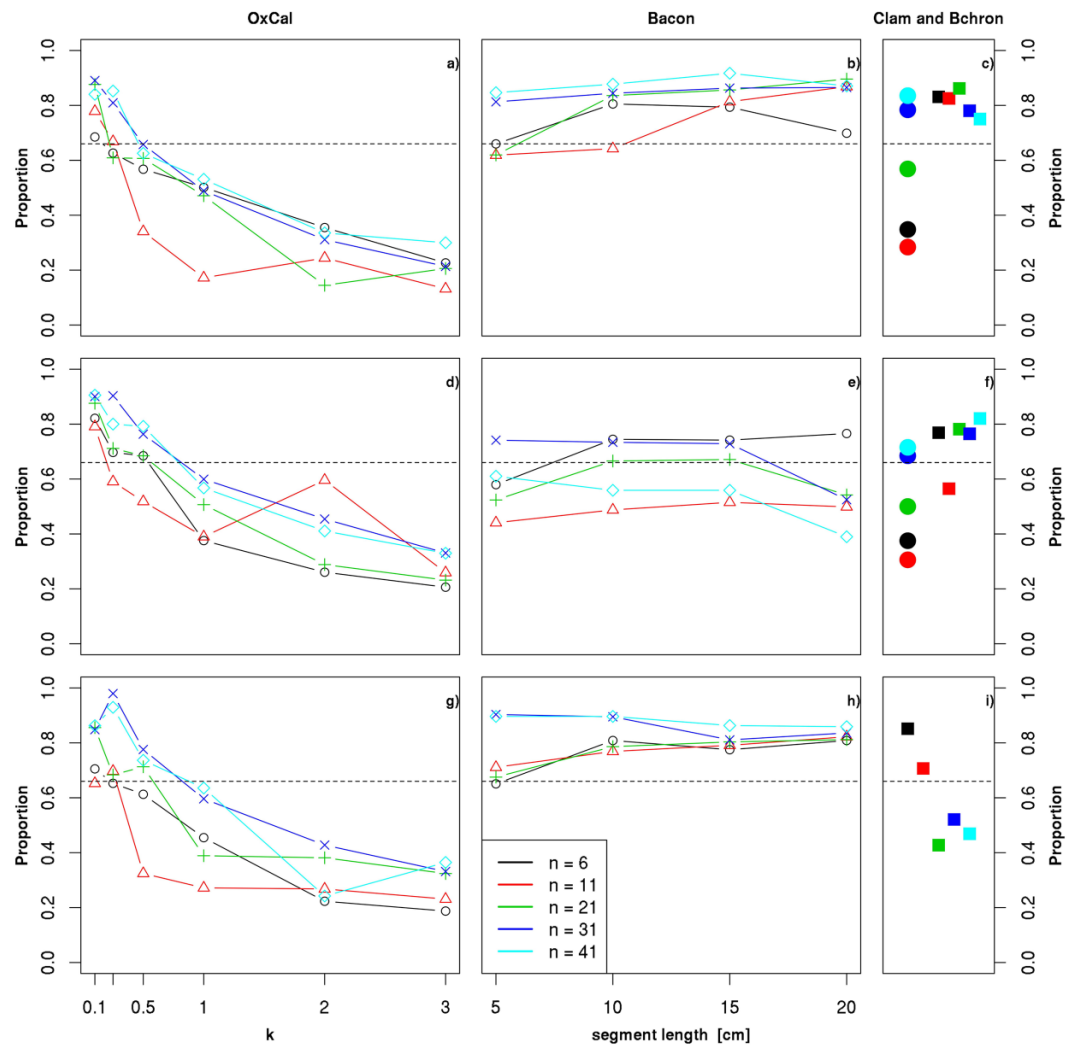


Figure 4. Probability coverage of age–depth modelling routines. Left panels: proportion of true varve ages between 16.6% and 83.3% quantiles of OxCal age–depth model ensembles as a function of k for (a) Elk Lake, (d) Holzmaar and (g) Iceberg Lake. Middle panels: proportion of true varve ages between 16.6% and 83.3% quantiles of Bacon age–depth model ensembles as a function of segment length for (b) Elk Lake, (e) Holzmaar and (h) Iceberg Lake. Proportion of true varve ages between 16.6% and 83.3% quantiles of CLAM (circles) and Bchron (squares) age–depth models for (c) Elk Lake, (f) Holzmaar and (i) Iceberg Lake (only Bchron). All colours as in the legend for (h). Dashed horizontal lines indicate the expected proportion (66.7%) of true varve ages between the 16.6% and 83.3% quantiles.

Table 1. Results of OxCal setting priors for k (means of three model runs using the same radiocarbon dates, differences among runs are minor, either indicating convergence or similar failure to converge).

No. of ^{14}C dates	Elk			Holzmaar			Iceberg		
	LB	UB	Proportion	LB	UB	Proportion	LB	UB	Proportion
6	0.03	2.88	0.62	0.01	0.08	0.97	0.02	10	0.64
11	0.02	0.21	0.85	0.02	0.13	0.92	0.05	6.83	0.63
21	0.04	0.17	0.81	0.04	0.16	0.84	0.3	4.2	0.55
31	0.05	0.19	0.93	0.05	0.18	0.94	0.08	0.83	0.97
41	0.05	0.16	0.9	0.07	0.2	0.86	0.12	0.82	0.92

LB: lower bound of k ; UB: upper bound of k ; proportion: proportion of varves between the 16.6% and 83.3% quantiles.

dates and lake is always found for one of the Bayesian age–depth modelling routines, but minima are found for all three methods. Within OxCal and Bacon, the lowest CRPS is found at different values of k and segment length, respectively, for different numbers of radiocarbon dates.

Figure 6 shows the PDFs of different age–depth models with 5 or 21 radiocarbon dates for Elk Lake at a depth of 500 cm, between two radiocarbon dates. The PDFs are wider when 5 radiocarbon dates are available than with 21 radiocarbon dates.

CLAM and OxCal with $k = 2$ always produce the narrowest age–depth models, while Bchron models are widest. OxCal models with $k = 0.25$ and Bacon models with segment length = 5 cm are indistinguishable. Bacon models with segment length = 15 cm are wider than models with segment length = 5 cm. The mode of the PDF of the OxCal age–depth model with $k = 2$ is distinctly different from the modes of PDFs produced by the other age–depth modelling routines, a tendency generally visible in Figure 2 (see also supplementary material, available online).

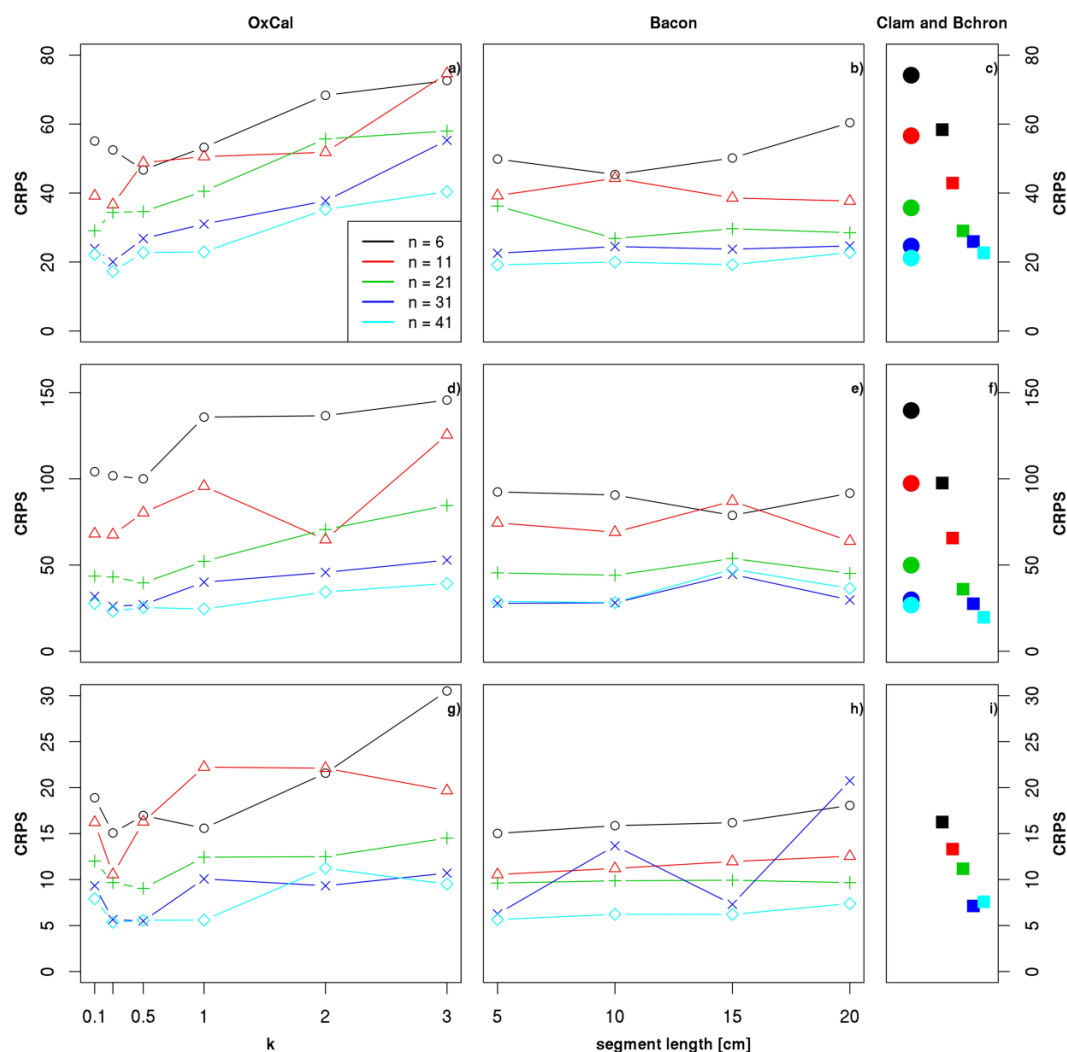


Figure 5. Continuous ranked probability score (CRPS). Left panels: CRPS of OxCal age–depth models as a function of k for (a) Elk Lake, (d) Holzmaar and (g) Iceberg Lake. Middle panels: CRPS of Bacon age–depth models as a function of segment length for (b) Elk Lake, (e) Holzmaar and (h) Iceberg Lake. Right panels: CRPS of CLAM (circles) and Bchron (squares) age–depth models for (c) Elk Lake, (f) Holzmaar and (i) Iceberg Lake (only Bchron). All colours as in the legend for (a).

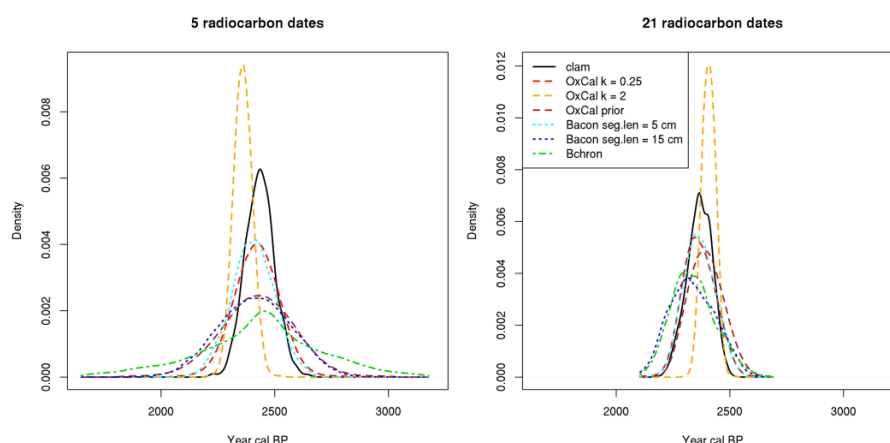


Figure 6. Probability density function (PDF) of age–depth models at depth 500 cm for Elk Lake using 5 (left) and 21 (right) radiocarbon dates.

The accumulation rates produced by a smoothing spline model in CLAM are very smooth (Figure 7) and look as if a smoothing spline has been fitted to the real accumulation rates. The true accumulation rates are often outside the 2.5% and 97.5% quantiles of accumulation rates produced by CLAM. Accumulation rates produced by Bacon are more variable than accumulation rates modelled by CLAM, and the real

accumulation rates are always inside the 2.5% and 97.5% quantiles of the modelled accumulation rates (Figure 7). The accumulation rates calculated by Bchron show conspicuous peaks at the 97.5% quantile of the accumulation rates for each dated level. For Holzmaar, Bacon and Bchron are able to find realistic median accumulation rates when 41 radiocarbon ages are available in age–depth modelling (Figure 7).

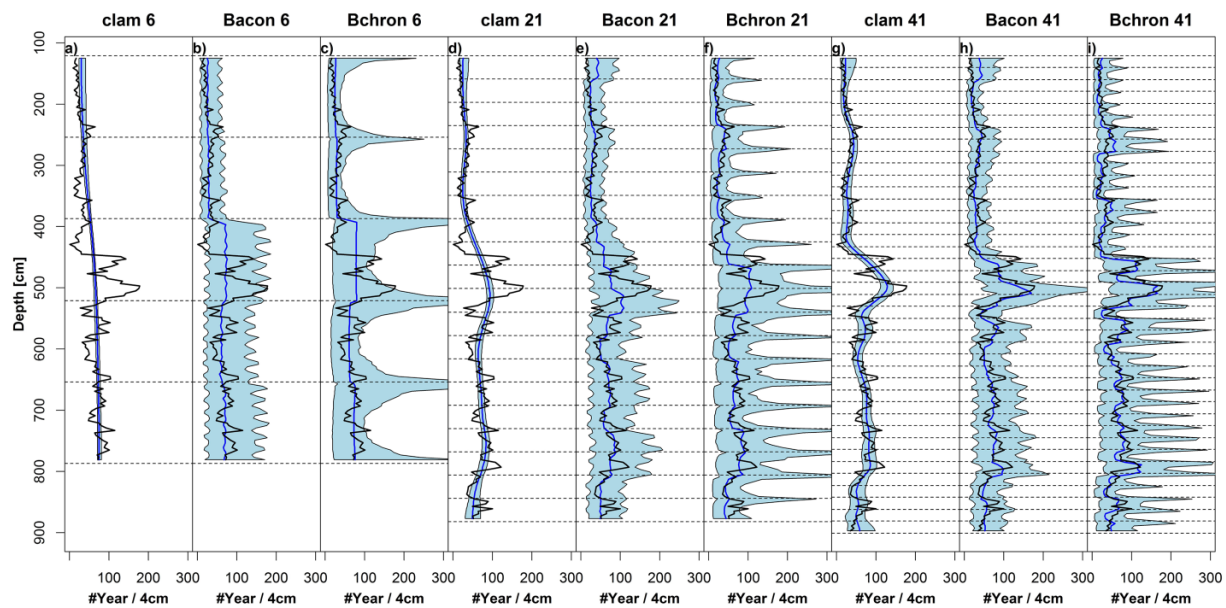


Figure 7. Accumulation rates for Holzmaar modelled by CLAM (a, d and g), Bacon (b, e and g) and Bchron (c, f and i) using 6 (a–c), 21 (d–f) and 41 (g–i) radiocarbon dates. Black lines are real accumulation rates, blue lines are median accumulation rates and lower and upper bounds are the 2.5% and 97.5% quantiles of the accumulation rates at one depth. Dashed lines indicate dated levels

Age–depth models obtained using OxCal with a prior set for k are generally too wide (Table 1). Six radiocarbon dates are not enough to constrain k . For instance, with Iceberg Lake, setting a uniform prior for k between 0.01 and 10 and using 6 radiocarbon dates result in lower and upper bounds of the posterior for k of 0.02 and 10.

The overall agreement index returned by OxCal favours models with $k = 0.1$, the lowest k tested in this study, and the value of $k = 0.1$ is always within the 95% Bayesian credible interval of the posterior distribution of k when setting a prior for k . But the measures of skill used in this study do not favour $k = 0.1$. Especially, the proportion of true varve ages found between the 16.6% and the 83.3% quantiles of the model ensemble is too high (Figure 4); hence, the uncertainty estimates are too pessimistic. The proportion of varves between the 16.6% and 83.3% quantiles is in favour of $k = 0.25$ or even $k = 0.5$ in some cases, while the RMSEP is largely indifferent to changes of k between 0.1 and 0.5 (Figures 3 and 4).

To test the effects of changing accumulation shapes and segment length, we used treatment contrasts in ANOVA with accumulation shape = 1.4 and segment length = 5 cm as the reference level. The effects of changing accumulation shape and segment length as well as their interaction are all significant ($p < 0.01$; all variables treated as factors) for the proportion of true ages between the 16.6% and 83.3% quantiles. The effect sizes are also considerable: changing the accumulation shape results in an average decrease in the proportion of varves between the 16.6% and 83.3% quantiles of 0.17, while changing the segment length results in an increase of 0.05. Notably, the effect of changing both is not additive, resulting in an interaction of 0.1. These changes are large given that the proportion of true varve ages between the 16.6% and 83.3% quantiles can vary between 0 and 1.

Choosing default priors for accumulation autocorrelation using Bacon, the posterior distribution of the accumulation autocorrelation was influenced by the segment length. Accumulation autocorrelation was low with low segment length and increased with increasing segment length.

Discussion

The performance of age–depth modelling routines has improved since Telford et al. (2004a) assessed age–depth models. Most importantly, the uncertainty estimation is now clearly superior.

This is mainly due to the introduction of age–depth models using Bayesian inference. With Bayesian age–depth models, the uncertainty increases between dated levels, which is a realistic feature (e.g. Parnell et al., 2011), indicative of our knowledge, or lack thereof, at each depth.

OxCal and Bacon users have to define parameters and/or prior distributions. According to Parnell et al. (2011), the need to supply parameter k is ‘one of the main advantages and disadvantages of the OxCal P-sequence model’, and the same is true for parameters in Bacon. While supplying parameters and priors enables users to influence age–depth models, users do not usually have enough information to justify their choice of parameters and priors. For instance, in our experiment to determine the influence of accumulation shape and segment length on Bacon age–depth models, changing the accumulation shape from 1.4 to 3.5 caused the 2.5% and 97.5% quantiles of the prior distribution for the accumulation rate to change substantially. However, we are unable to make a decision about the more appropriate accumulation rate prior. Additionally, no objective information for choosing the segment length is available. Still, changing these parameters and priors affects uncertainty estimates, so finding a way to make such choices more objective might be one way to improve the Bacon age–depth modelling routine.

For OxCal, Bronk Ramsey and Lee (2013) introduced the possibility to treat k as a variable, which alleviates the problem described by Parnell et al. (2011).

In contrast to OxCal and Bacon, users are not able to set any parameters manually in Bchron, which is recognised as one of the main disadvantages of Bchron by Parnell et al. (2011). Interestingly, Bchron is equivalent to letting parameter k in OxCal have a gamma distribution with unknown variance (Parnell et al., 2011). This might be the reason for the visual similarity between OxCal with variable k and Bchron age–depth models (Figure 2b). With classical age–depth modelling approaches, uncertainties are now routinely evaluated as R-code has been made available by Blaauw (2010). Uncertainty estimations, however, have not fundamentally changed since the assessment by Telford et al. (2004a), resulting in underestimated uncertainties unless many radiocarbon ages are used.

The number of radiocarbon dates and the uncertainty assigned to radiocarbon dates here is different from the study by Telford et al. (2004a) and thereby precludes a direct, quantitative

comparison of the results. Still, a qualitative comparison with results from a decade ago is possible and suggests that uncertainty estimation in particular has improved with the introduction of age–depth modelling routines using Bayesian inference.

From a statistical viewpoint, the uncertainties given by CLAM and the uncertainties given by the models using Bayesian inference are not directly comparable. The bootstrapping uncertainties given by CLAM are uncertainties of the mean response at a certain depth (i.e. the mean of a multitude of points) and not uncertainties of one single point predicted at the same depth (e.g. Draper and Smith, 1998). In contrast, the uncertainties given by the Bayesian models are true uncertainties of single point predictions at a certain depth. In this study, we do not account for this difference, as both types of uncertainties are presented to users as uncertainties of the age–depth model and we are interested in assessing the uncertainties users are provided with.

Accumulation rates modelled by CLAM, Bacon and Bchron mainly depend on the individual setup of the age–depth models. CLAM does not explicitly model the sediment accumulation process but uses different interpolation and regression methods to establish an age–depth relationship. CLAM with a smoothing spline is designed to draw a smooth line taking into account the dating information; hence, no large changes in accumulation rates occur, and the variability of accumulation through time is clearly underestimated. Increased uncertainty of accumulation rates when more radiocarbon dates are available is an unexpected and undesired behaviour, as using more radiocarbon dates should result in reduced uncertainties. Although Bayesian age–depth modelling routines simulate changing accumulation rates, the mean accumulation rate (that is mostly used by scientists) is fairly constant between two dated levels, and therefore, the age–depth model is a straight line between two dated levels. Still, the median accumulation rates produced by Bacon and especially Bchron using many radiocarbon dates seem realistic.

Methods for testing age–depth modelling routines

Telford et al. (2004a) introduced the method of using varved sediment sequences and simulating radiocarbon dates from known varve ages using the radiocarbon calibration curve to assess different regression-based age–depth modelling techniques. Blockley et al. (2007) used the same technique to assess the performance of wiggle match dating and of OxCal. In both studies, the method was recognised as a powerful tool to test age–depth modelling algorithms. Parnell et al. (2011) used a different approach for assessing the performance of age–depth modelling algorithms. They used many cores from the European Pollen Database and assessed the leave-one-out cross-validation performance of each age–depth model. The approach by Parnell et al. (2011) is applicable to many cores but has a limited number of test cases per core. In contrast, the approach presented by Telford et al. (2004a) is only applicable to a limited number of cores but has a large number of test cases per core, and the properties of the radiocarbon ages are exactly known as they are simulated.

As noted by Telford et al. (2004a), our procedure might be unduly pessimistic for low numbers of radiocarbon dates since we do not make use of observable lithological changes which might indicate changes in the accumulation rate (especially in Holzmaar), nor do we use iterative procedures (Christen and Buck, 1998) for selecting samples for dating. In other aspects, the setting we use is nearly optimal: all radiocarbon dates used are correct and have rather low uncertainties of 30 years, these radiocarbon dates come from one individual year and there are no hiatuses in the cores.

Outliers and hiatuses

In this study, we used simulated radiocarbon dates, and the cores analysed comprised complete sequences without hiatuses or

outliers, unlike other sequences that may span hiatuses and include outliers. Bronk Ramsey (2009) discusses three reasons why outliers may occur and develops different outlier models of varying complexity to detect outliers. These outlier models are all available in OxCal. Some of these models require the user to set a prior probability of a radiocarbon date being an outlier. Bacon uses a different approach to deal with outliers. Radiocarbon dates are, by default, assumed to come from a *t*-distribution. As *t*-distributions have heavier tails than normal distributions, the calibrated radiocarbon dates usually cover a wider range than when assuming a normal distribution. Thereby, it is often possible to retain outlying dates in the age–depth models. If dates that seem correct are treated as outliers (which may occur near a large change in the accumulation rate), it is possible to reduce the uncertainty assigned to radiocarbon dates by changing the distribution of the radiocarbon dates to normal distributions.

Bchron also has an outlier model that requires users to set a prior probability of radiocarbon dates being outliers. In Bchron, hiatuses occur naturally because of the minimum assumptions of smoothness (A.C. Parnell, 2016, personal communication). Accumulation rates above and below a hiatus are unrelated. It is possible to include hiatuses in the age model using OxCal and to estimate the duration of the hiatus, as well as including additional information such as layer counts (varves, annual layers in ice cores) to further constrain age–depth models.

The presence of outliers can alter the performance of age–depth modelling routines. For instance, Blaauw and Christen (2011) report an excellent performance of Bacon on a sediment core from the Round Loch of Glenhead, whereas it took them ‘much trial-and-error to find parameter distributions that resulted in successful, converged P sequence runs’.

With complicated sediment sequences, it might be worth following Blockley et al. (2007), who recommend using more than one age–depth modelling routine and comparing results of these algorithms.

Advantages and disadvantages of age–depth modelling routines

CLAM

The CLAM model tested (smoothing spline, smoothing parameter = 0.4) is designed to draw smooth lines between dated levels. Thus, the uncertainty between dated levels does not increase, which leads to overly narrow model ensembles and ignores the fact that we have less information on the age of the sediment between dated levels. Currently, CLAM does not have a monotonicity constraint which renders age–depth modelling difficult as soon as PDFs of calibrated ¹⁴C ages overlap. Age–depth models with age reversals are removed from the ensemble of age–depth models at the end of the age–depth modelling procedure. For Iceberg Lake, a very low number of age–depth models remained after this procedure. Imposing a monotonicity constraint seems feasible using monotonic splines (Wood, 1994).

OxCal

OxCal is the only method tested where changing parameters (*k*) had a systematic influence on the age–depth models and their uncertainty. In OxCal, changes in sedimentation rate are mainly controlled by parameter *k* that controls the number of sedimentation events per unit depth. Increasing the number of sedimentation events per unit depth reduces the flexibility of the accumulation rate (Bronk Ramsey, 2008). Therefore, changes of *k* have a systematic influence on model flexibility and thereby on confidence intervals (Bronk Ramsey, 2008), with overconfident models for high *k* and pessimistic models for low *k*. An overly high *k* also results in biased mean age–depth models. Choosing a

good k is not a trivial task, but setting a prior for k seems useful in the lakes studied, generating good results in terms of RMSEP, although model ensembles were wide. For the lakes studied, a moderate number of radiocarbon dates were needed to get meaningful posterior distributions for k .

OxCal by default only produces ensembles of age–depth models at dated depths, but returns a PDF of probable ages at each depth. Additional depths can be specified at which ensembles of age–depth models can be formed. OxCal failed to converge when specifying additional depths at 0.5 cm increments throughout a core. This prevents the use of novel methods for estimating similarities between age–uncertain time series (e.g. Rehfeld and Kurths, 2014).

It is possible to use OxCal by mainly using a graphical user interface where users are allowed to choose models, import data and change settings. The user is then able to see the code produced and make subsequent changes directly in the code. It is also possible to call OxCal from, and to handle all the results in, external software such as R. OxCal is quite fast. OxCal only returns a result when automatic convergence checking shows successful convergence, but users are not provided with trace plots or Gelman–Rubin diagnostics (e.g. Gelman et al., 2013).

Bacon

Users have to define two prior distributions and the segment length over which accumulation is constant using Bacon. Users are provided with a trace plot of the MCMC algorithm whereby they can partly assess convergence and mixing of the MCMC algorithm. Goring et al. (2012) have calculated priors for accumulation rate and accumulation shape from sediments in North America that are currently used as default settings for priors in Bacon. If the default mean accumulation rate seems unrealistic, Bacon suggests changing the prior for mean accumulation rate according to information available from ^{14}C dates. Changing the shape of the accumulation rate prior influences age–depth models (see Blaauw and Christen, 2011). For the sediment sequence of Holzmaar with changing accumulation rates, the MCMC algorithm was slow to converge and did not mix well using default settings for the shape of the accumulation rate prior. This is in line with Goring et al. (2012) who found temporally changing parameters for gamma distributions fitted to temporally binned mean accumulations rates of sediments in North America.

For the sediment sequence of Holzmaar, the changing real accumulation rates led Bacon to consider some (in our case correct) radiocarbon dates as outliers, which was avoided by changing the distribution of the radiocarbon dates from a t -distribution to a normal distribution. The segment length has an unpredictable influence on the spread of the ensemble. In the experiment testing the influence of changing accumulation shape and segment length, the large interaction between accumulation shape and segment length is especially worrying as the effect of changing segment length depends on the accumulation shape chosen. When using long segment lengths, the algorithm did not converge rapidly and the mixing was poor. Unfortunately, the manual accompanying Bacon does not offer a solution to the problem of choosing a segment length as it suggests decreasing the segment length ‘until the model appears sufficiently smooth’. The only clear recommendation we can give is to choose segment lengths shorter than the mean distance between dated levels and to choose segment lengths allowing for rapid convergence. Running Bacon is rather time-consuming but is very user-friendly for users not experienced in R.

Bchron

Bchron does not allow users to change settings (Haslett and Parnell, 2008; Parnell et al., 2011). In Bchron, accumulation rates in consecutive segments are unrelated, hence abrupt changes in

accumulation rates are possible (‘minimal assumptions on smoothness’; Haslett and Parnell, 2008). The only restriction is given by the likelihood term in the Bayesian model. Therefore, especially for low numbers of radiocarbon dates, the confidence intervals given by Bchron are too wide, except when the probability densities of calibrated radiocarbon dates overlap. For Holzmaar, the median accumulation seemed realistic when many radiocarbon dates were used, although the 97.5 percentile of the accumulation rate was very spiky. Bchron is available as an R-package and on GitHub (<https://github.com/andrewcparnell/Bchron>). Running Bchron is, like OxCal, relatively fast (under default settings with 10,000 iterations).

Conclusion

We re-applied the powerful method for testing age–depth modelling routines developed by Telford et al. (2004a) which uses simulated radiocarbon dates derived from varve chronologies. The age–depth modelling routines and especially their ability to realistically assess uncertainties of age–depth models have been greatly improved in the last 10 years. All the Bayesian age–depth modelling routines tested in this study perform well in terms of precision and uncertainty estimation when a moderate number of radiocarbon dates are used for Holocene length sequences. However, all the age–depth modelling routines struggle to produce reliable sediment accumulation rates. Bayesian models are slightly more difficult to use than regression-based models, but investing time into constructing such age–depth models is justified given the fundamental role age–depth models have when interpreting sediment sequences. The radiocarbon dates used in this study were nearly perfect, yet the optimal parameter settings in Bayesian age–depth modelling routines are different for each lake considered and even differ with the number of radiocarbon dates used.

Acknowledgements

We thank Christopher Bronk Ramsey, Andrew C Parnell, Maarten Blaauw and an anonymous reviewer for comments on an earlier version of this manuscript. We also thank Gavin L Simpson, Joseph D Chipperfield and Johannes Werner for useful comments on preliminary results. We thank Cathy Jenks for language editing this manuscript.

Funding

Norwegian Research Council FriMedBio project palaeoDrivers (213607) helped support this work.

References

- Bennett KD (1994) Confidence intervals for age estimates and deposition times in late-Quaternary sediment sequences. *The Holocene* 4(4): 337–348.
- Blaauw M (2010) Methods and code for ‘classical’ age-modelling of radiocarbon sequences. *Quaternary Geochronology* 5(5): 512–518.
- Blaauw M and Christen JA (2011) Flexible paleoclimate age–depth models using an autoregressive gamma process. *Bayesian Analysis* 6(3): 457–474.
- Blockley SPE, Blaauw M, Bronk Ramsey C et al. (2007) Building and testing age models for radiocarbon dates in Lateglacial and early Holocene sediments. *Quaternary Science Reviews* 26(15–16): 1915–1926.
- Bradbury JP, Dean WE and Anderson RY (1993) Holocene climatic and limnologic history of the north-central United States as recorded in the varved sediments of Elk Lake, Minnesota: A synthesis. In: Bradbury JP and Dean WE (eds) *Elk Lake, Minnesota: Evidence for Rapid Climate Change in the*

- North-Central United States (Special Paper), Boulder, CO: Geological Society of America, pp. 309–328.
- Bronk Ramsey C (1995) Radiocarbon calibration and analysis of stratigraphy: The OxCal program. *Radiocarbon* 37(2): 425–430.
- Bronk Ramsey C (2008) Deposition models for chronological records. *Quaternary Science Reviews* 27(1–2): 42–60.
- Bronk Ramsey C (2009) Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon* 51(3): 1023–1045.
- Bronk Ramsey C and Lee S (2013) Recent and planned developments of the program Oxcal. *Radiocarbon* 55(2–3): 720–730.
- Bronk Ramsey C, Dee M, Lee S et al. (2010) Developments in the calibration and modeling of radiocarbon dates. *Radiocarbon* 52(3): 953–961.
- Christen JA and Buck CE (1998) Sample selection in radiocarbon dating. *Journal of the Royal Statistical Society Series C: Applied Statistics* 47: 543–557.
- Draper NR and Smith H (1998) *Applied Regression Analysis*. New York: John Wiley & Sons.
- Gelman A, Carlin JB, Stern HS et al. (2013) *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall.
- Gneiting T and Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102: 359–378.
- Goring S, Williams JW, Blois JL et al. (2012) Deposition times in the northeastern United States during the Holocene: Establishing valid priors for Bayesian age models. *Quaternary Science Reviews* 48: 54–60.
- Haslett J and Parnell AC (2008) A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society Series C: Applied Statistics* 57: 399–418.
- Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast* 15: 559–570.
- Jeffers ES, Bonsall MB, Froyd CA et al. (2015) The relative importance of biotic and abiotic processes for structuring plant communities through time. *Journal of Ecology* 103(2): 459–472.
- Loso MG (2009) Summer temperatures during the Medieval Warm Period and Little Ice Age inferred from varved proglacial lake sediments in southern Alaska. *Journal of Paleolimnology* 41(1): 117–128.
- Parnell AC, Buck CE and Doan TK (2011) A review of statistical chronology models for high-resolution, proxy-based Holocene palaeoenvironmental reconstruction. *Quaternary Science Reviews* 30(21–22): 2948–2960.
- Rehfeld K and Kurths J (2014) Similarity estimators for irregular and age-uncertain time series. *Climate of the Past* 10(1): 107–122.
- Reimer PJ, Bard E, Bayliss A et al. (2013) IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* 55(4): 1869–1887.
- Telford RJ, Heegaard E and Birks HJB (2004a) All age-depth models are wrong: But how badly? *Quaternary Science Reviews* 23(1–2): 1–5.
- Telford RJ, Heegaard E and Birks HJB (2004b) The intercept is a poor estimate of a calibrated radiocarbon age. *The Holocene* 14(2): 296–298.
- Wood SN (1994) Monotonic smoothing splines fitted by cross-validation. *Siam Journal on Scientific Computing* 15(5): 1126–1133.
- Zolitschka B, Brauer A, Stockhausen H et al. (1998) An annually dated Late Weichselian continental paleoclimate record from the Eifel, Germany. *Geology* 28: 783–786.