Quaternary Science Reviews 30 (2011) 2948–2960

Contents lists available at ScienceDirect

Quaternary Science Reviews

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

Invited Review

A review of statistical chronology models for high-resolution, proxy-based Holocene palaeoenvironmental reconstruction

Andrew C. Parnell[a,*], Caitlin E. Buck[b], Thinh K. Doan[c]

[a] School of Mathematical Sciences, University College Dublin, Ireland
[b] School of Mathematics and Statistics, The University of Sheffield, UK
[c] School of Computer Science and Statistics, Trinity College Dublin, Ireland

ARTICLE INFO

Article history:
Received 19 April 2011
Received in revised form
28 July 2011
Accepted 29 July 2011
Available online 8 September 2011

Keywords:
Statistical chronology construction
Age–depth modelling
Proxy-based environmental reconstruction
Palaeoclimate reconstruction
Bpeat
OxCal
Bchron

ABSTRACT

In this paper we explain the background, workings, and results obtained from three recently developed statistical age-depth models implemented in freely available, general purpose software packages (Bpeat, OxCal and Bchron). These models aim to reconstruct the sedimentation rate in a single core (typically lake, peat or ocean) given a limited number of scientific date estimates (usually radiocarbon) and fixed depths. Most importantly, they provide a suitably qualified estimate of the uncertainty in the age-depth chronology and thus can be used in a variety of applications. We perform a large data-driven study of the three models and discuss their general utility in chronology construction for palaeoenvironmental research.

1. Introduction

Chronology construction is key to understanding the timing of past environmental change. Statistically estimated chronologies have been used, amongst many other uses, in creating estimates of palaeoclimate (e.g. Blaauw et al., 2010), determining past sea level change (e.g. Kemp et al., 2009), and calculating the dates of various related environmental events (e.g. Parnell et al., 2008). In the last few years there have been some major advances in the tools used to construct such chronologies, all of which adopt modern Bayesian statistical techniques. We give a brief introduction to the use of Bayesian statistics in this context and review the statistical methods behind Bayesian tools for chronology construction. Our focus is on developing age-depth models for cores extracted from sedimentary deposits (e.g. lakes, peat and oceans) from which a selection of dateable material has been recovered (typically organic material that can be radiocarbon dated) along with a much larger quantity of information relating to palaeoenvironment (typically in the form of indirect proxy observations such as pollen counts).

The chronological models described in this paper use the (relatively small number of) absolute scientific date estimates for samples extracted from a core and combine them with stratigraphic or event-based data taken from other remains within the same core. The construction of a chronology is thus most often a study in the sedimentation process at a particular site. This paper discusses various new statistical methods to reconstruct the sedimentation rate at an individual site with due respect to the various sources of uncertainty.

Chronology construction would be a reasonably simple task were it not for the numerous uncertainties that are present in the data. The most obvious source of uncertainty is that relating to the scientific dating method itself. Such errors are never negligible and, in the case of radiocarbon dating, can be hard to quantify, manage and describe. The specific nature of the material sent for dating or how it has been handled may also lead to erroneous date estimates (herein *outliers*) which will need to be removed or adjusted. At most sites, further uncertainty arises from our lack of knowledge about the specific nature of sedimentation. At such sites we need a sedimentation model which captures our *a priori* knowledge about sedimentation in general, but also accounts for the uncertainty inherent in using a general purpose model.

[*] Corresponding author. Tel.: +353 17167105.
E-mail address: andrew.parnell@ucd.ie (A.C. Parnell).

0277-3791/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.quascirev.2011.07.024

Thus, put succinctly, our task is as follows: use the scientific dating and depth information to produce a chronology which can be used to estimate the sedimentation rate and thus offer an estimate for the date at any given depth in the core along with a carefully quantified statement of uncertainty. In practice, the depths for which we are most interested in obtaining age estimates will be those at which proxy data have been collected and, hence, where we have palaeoenvironmental information.

In this paper, we describe three different models for constructing chronologies for use in high-resolution palaeoenvironmental reconstructions. We then describe a large-scale study of age-depth modelling for 111 cores taken from the European Pollen Database (http://www.europeanpollendatabase.net/). For each core, we run each age-depth model numerous times (as part of leave-one-out experiments) to determine the behaviour of the different models under a spectrum of different real-world situations. We use the results we obtain to offer guidance to users about when and why to adopt each method. Such advice is key to the successful use of such methods, because the three models are similar, but produce markedly different results depending on the user's input, and the nature of the scientific dating evidence available from the core (see Fig. 3 for an illustrative example).

In the early sections of the paper, we illustrate our observations and findings using a running example which relates to data derived from a single core from Słopiec, Kielce County, Poland (Szczepanek, 1992). The core has ten radiocarbon determinations on samples deposited unevenly in depth throughout the core. We have removed the most recent radiocarbon determination due to the fact that its estimated calibrated date lies outside the range of the radiocarbon calibration curve (which extends only as far as 1950 AD). (More recent dates can be handled by some of the packages we introduce, but their treatment is inconsistent and so we have removed it for comparison purposes.) Table 1 shows the data for the Słopiec core.

The paper is structured as follows. Sections 2 and 3 briefly introduce the topics of Bayesian statistics and statistical radiocarbon calibration respectively, with references therein for readers requiring more detail. Section 4 contains the main statistical arguments required to create a chronology from the radiocarbon determinations and an appropriate model for sedimentation. Section 5 reviews some of the previous models used to create chronologies. Section 6 contains the descriptions of the three new models, and gives brief details as to their use. Section 7 contains a broad survey of the different models and gives hints as to which models are appropriate for different situations. We discuss the implications and future directions of these approaches in Section 8.

2. Bayesian statistics

In this section we briefly introduce some of the ideas behind Bayesian statistical modelling, leaving their particular relevance to chronology building until later sections to avoid confusion between the more general ideas presented here and the specifics relevant to chronology construction. The text below is meant to be a very basic introduction to Bayesian ideas. Those interested in more detailed information may refer to one of the many excellent introductory Bayesian textbooks, including Gelman et al. (2003), Lee (2004), O'Hagan and Forster (2004).

Table 1
Radiocarbon and depth data for the Słopiec core, Poland (Szczepanek, 1992).

Reference	^{14}C determination	^{14}C error	Depth (cm)	Thickness (cm)
Gd-1157	1090	95	140	5
Gd-1241	2710	55	195	5
Gd-775	3450	75	220	5
Gd-1158	3650	50	240	5
Gd-776	9090	100	345	5
Gd-703	9330	145	395	5
Gd-700	9620	120	400	5
Gd-702	10080	160	420	5
Gd-704	10280	210	513.5	2.5

Most Bayesian statistics is concerned with *parameters* and *data*. The former represent quantities that we do not know but may wish to estimate, and the latter some fixed observations which we may like to use to estimate the former. Bayes' formula can be most simply stated in words as:

posterior is proportional to likelihood times prior

where the *posterior* represents the probability distribution of the parameters given the data, the *likelihood* represents the probability distribution of the data given the parameters, and the *prior* represents external information (not obtained from the data) about the parameters. The formula can thus be written more eloquently as:

$$p(\theta|x) \propto p(x|\theta) \times p(\theta)$$

where θ is a parameter (or a set of parameters) and x is data. The sign '|' is used to represent a conditional probability, thus $p(\theta|x)$ can be read as "*the probability distribution of theta given x*", or "*the probability distribution of our unknown parameters given our data*".

The greatest controversies associated with the use of Bayesian statistics typically arise from the inclusion of subjective prior information, since this gives rise to the placing of (typically) subjective restrictions on the values that the parameters can take prior to seeing the data. We do not concern ourselves with philosophical criticisms of the use of priors here (see the references above for more discussion). However, we make three points which are directly relevant for the construction of chronologies:

- Prior distributions need not be a subjective choice if there is good reason to include external information on the likely values. Many prior distributions can be based on other 'objective' scientific evidence; the Bayesian framework provides a neat way to synthesise such information.
- The likelihood itself may be a subjective choice if little is known about the data generating process. Normally-distributed likelihoods are very common in many areas of statistics (not just Bayesian) even though there may be little or no scientific reasoning behind their use.
- Bayes' formula allows for simple inversion of complicated statistical models. From the formula above, it is possible to obtain a probability distribution of parameters given data from its conditional inverse: the probability distribution of data given parameters, thus avoiding the error of the transconditional (Ambaum, 2010). This is especially useful in, for example, palaeoclimate reconstruction (where the parameters are unknown climates and the data are proxy samples) because it is more reasonable to create a probability distribution for proxy data given climate than vice versa.

We do not discuss here the computational and mathematical challenges of assuming a likelihood and prior probability distribution and obtaining a posterior. We note, however, that Bayes' theorem is commonly written with a proportional symbol ($\propto$) rather than an equality because the constant of proportionality is usually extremely difficult to calculate—involving high-dimensional integration of the likelihood and prior. In modern Bayesian inference, the need to calculate this integral exactly is avoided by estimating the posterior via *sampling*. Thus most modern Bayesian computer software simulates many values from the posterior distribution for θ rather than producing an exact probability density function. The resulting samples are all treated as equi-probable and, when summarised by calculating means, standard deviations, etc, provide estimates of the same summary for the probability distribution from which the samples are drawn. An estimate of the true posterior probability distribution can thus be provided by drawing a histogram of the sampled values.

The last important concept we discuss, which is key to the construction and use of chronologies, is that of joint and marginal uncertainty. In many cases, we will have multiple parameters that we wish to estimate simultaneously based on our data—for example the dates of several different depths in the core. The associated posterior samples, when taken together, lead to estimates of the *joint* posterior distribution and will reflect the relationship between the parameters. Consider a simple example where we have, say, three depths at which we wish to estimate dates. At each iteration we will sample three numbers; one per date parameter. Running the sampler many times leads to many thousands (or more) sets of size three, each of which is a sample from the joint posterior for the three date parameters. If the three parameters are positively correlated then, in any given sample, all three will tend to co-vary—being all larger in the same iteration or all smaller in the same iteration. Their *marginal* uncertainty can be investigated simply by looking at the behaviour of all of the samples of a single parameter whilst ignoring the other values. The concept of joint and marginal uncertainty is of key importance when using chronologies in palaeoenvironmental reconstruction and/or palaeoclimatic event identification. The concept of joint uncertainty is explored further in Section 3 and illustrated in Fig. 2.

3. Bayes and radiocarbon dating

Not all chronologies used in Holocene palaeoclimate reconstruction require radiocarbon determinations. Some sediments are laminated in which case the chronology is (at least superficially) simple to derive via layer counting (e.g. Hajdas et al., 1993; Rasmussen et al., 2006). In most other cases, radiometric dating techniques are used. Of these, radiocarbon dating is by far the most common, but uranium–thorium dating is also used, especially for older sediments. The ubiquity of radiocarbon dating creates particular problems in chronology construction because it gives rise to date estimates which do not take the form of standard probability distributions. For this reason our discussion below focusses on the use and implications of radiocarbon dating for chronology building.

The need to calibrate radiocarbon determinations via a data-derived calibration curve is well-studied (see e.g. Blackwell and Buck, 2008, for a review) and software for using such a calibration curve to calibrate radiocarbon determinations (in the form of uncalibrated date estimates and laboratory standard errors) is now common-place (e.g. CALIB Stuiver and Reimer, 1993; OxCal Bronk Ramsey, 1995; BCal Buck et al., 1999). The latter two software packages use Bayesian models in which the parameter of interest (θ) is the unknown calibrated date of one or more samples and the data are the uncalibrated radiocarbon determinations and the radiocarbon calibration curve (the most recent estimate of which is) (Reimer et al., 2009). The form of the calibration curve (which is non-monotonic or 'wiggly') makes the proportionality constant in Bayes' equation hard to calculate and so, other than for the interpretation of single radiocarbon determinations, simulation techniques (as outlined above) are typically used.

The wiggles in the calibration curve mean that posterior probability distributions for calibrated dates are often multi-modal and difficult to describe or summarise except via graphical representations of the sort used in Fig. 1.

Although all the Bayesian software packages discussed here use simulation-based methods to estimate calendar dates, most do not present the resulting samples to the user. Instead we are often given the calendar date estimates summarised as a histogram or a 95% highest posterior density region (often written as HPD region or HDR), representing the shortest interval of time that encompasses 95% of the samples. In Fig. 1, we show a simulation-based estimate of the posterior distribution of the calibrated date of radiocarbon determination Gd-1241 (which corresponds to depth 195 cm in the Słopiec core). The estimate is based on 10,000 samples from the posterior distribution and, in the caption, we list the first 10 values in that sample.

Bayes' equation is particularly powerful when used to help calibrate multiple radiocarbon determinations simultaneously. Stratigraphic information and other chronological knowledge can be used to help formulate the prior distributions for the calibrated dates and thus ensure that the joint posterior samples that we produce are consistent with all available information. Such techniques have been used to great effect in archaeology where the parameters of greatest interest are typically the calibrated dates of the start and end of periods of human activity during which multiple samples suitable for radiocarbon dating were deposited see (e.g. Buck et al., 1992; Buck et al., 1996).

In some situations, multiple radiocarbon determinations may relate to the same event or to several different, but inter-related, events. When interpreting such data, it is necessary to investigate whether the data are consistent with one another or whether any of them should be considered as outliers. Typically extra parameters are added to the model, one per radiocarbon determination, which are defined to take values of zero or one though other methods have been proposed (see Bronk Ramsey, 2009; Christen and Perez, 2009). Posterior samples for such parameters will consist of sets of zeros and ones, indicating which of the radiocarbon determinations are considered to be outliers (relative to the others) and which are not. For any given radiocarbon determination, the proportion of samples for the associated outlier parameter that are equal to one may be considered an estimate of the posterior probability that the determination is an outlier. Such probabilities are estimated from the marginal posterior distributions for each outlier parameter.

In chronology modelling, the most obvious external (prior) knowledge comes from the stratigraphic relationship of the sediment found in the core; that younger sediment must lay upon older sediment. It is relatively simple to formulate a prior distribution

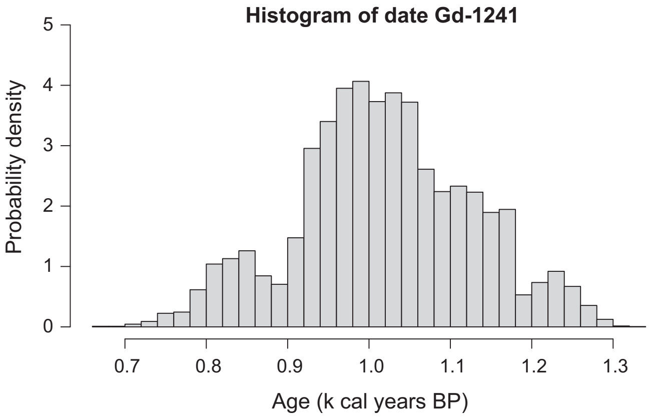

Fig. 1. Histogram of the estimated calibrated date of sample Gd-1241 from depth 195 cm in the Słopiec core. The first 10 posterior samples are 1.032, 1.127, 0.875, 0.985, 0.840, 0.997, 0.952, 0.946, 0.989, and 0.930. The histogram is created from 10,000 such samples. The mean is 1.019 k cal years BP, the mode 0.990 k cal years BP. The 95% HDR is from 0.795 to 1.243.

that forces such a relationship—in a simulation-based implementation, for example, we could simply reject all samples that violate the *a priori* ordering for the calibrated dates. The resulting joint posterior will contain only those samples in which older dates are associated with deeper sediments. For example, if our data consisted of just two radiocarbon determinations at depths 1 m and 1.5 m, we may expect them to be associated with a single joint sample of (calibrated) dates that included values such as 1.3 and 1.7 k cal years BP (respectively), or 1.57 and 1.68 k cal years BP, but we would never see 1.8 and 1.6 k cal years BP. Each individual sample would satisfy the joint rules of our prior distribution. However, if we were to plot a marginal histogram for each of the two calibrated dates, it is perfectly plausible for them to overlap, despite the fact that each joint sample satisfies our criteria. Fig. 2 illustrates such a scenario. If we were to calculate differences between the dates of these layers correctly, by subtracting values at each iteration of the sampler, we would never see any negative values because we would be using the joint samples (see Parnell et al., 2008, for more on this topic).

4. Bayesian chronology models

Using the tools of Bayesian statistics and radiocarbon calibration, we may now hope to build a chronology for a particular site. Recall that a chronology will state the relationship between depth and age such that an estimated age can be created for any given depth. This description of the problem may remind readers of (simple linear) regression where we are given the value of an explanatory variable and wish to predict the value of a response. In the chronology building example, our response variable is date and our explanatory variable is depth. The result of fitting such a model in the Bayesian way (via simulation) is that each sample we draw from the joint posterior will be a list of dates for a set of given depths, with the property that each sample is an equi-probable outcome given the regression model and the data. Collecting together all such samples will allow us to estimate the uncertainty in the chronology.

As with all Bayesian models, we need to define a likelihood and a prior distribution. In our situation, the likelihood we use is the same as that of the standard likelihood used for calibrating individual radiocarbon dates. Recall that the likelihood estimates the probability of observing the data given some parameter values. In chronology modelling the parameters relate to dates of the various layers in the sediment, thus the likelihood will estimate the probability of observing the radiocarbon determinations given the estimated calendar dates.

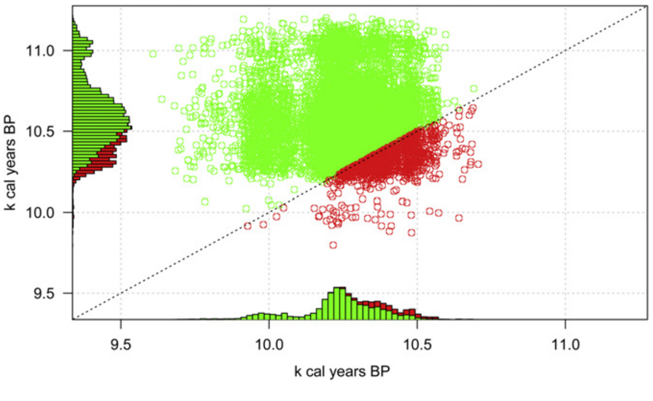

Fig. 2. A collection of joint posterior samples (scatter plot) and marginal histograms for the calibrated dates of two depths in the Słopiec core. The *y*-axis shows the calibrated date of Gd-700 (9620 ± 20; depth 400 cm), whilst the *x*-axis shows the calibrated date for Gd-703 (9330 ± 145; depth 395 cm). Using the stratigraphy to provide *a priori* relative chronological information allows us to remove samples (shown in red) for which the calibrated dates are not in the correct chronological order given the depths at which the radiocarbon samples were taken. Rejecting the red samples in the scatter plot leads to a reduction in uncertainty in our date estimates both in the joint distribution and in the marginal distributions. Note that the marginal distributions overlap despite every sample obeying our stratigraphic rule.

The key task of Bayesian chronology building, and the main focus of this paper, is determination of an appropriate prior probability distribution for the chronology. Some desirable properties of this prior distribution might include the following.

1. *Monotonicity*: the idea that deeper sediments must be older.
2. *Flexibility*: that changes in sedimentation rate are allowed to occur as dictated by the data and any other external information input by the user. The sedimentation rate may even be allowed to decrease to zero, corresponding to a hiatus in sedimentation, or increase to an arbitrarily large value corresponding to a large instantaneous dump of sediment.
3. *Matching uncertainty at adjacent levels*: depths that lie close to radiocarbon-dated samples (which produce highly multimodal posterior probability distributions for their calibrated dates) should themselves have estimated calendar dates with multi-modal probability distributions. To ignore this structure in the uncertainty is to not use the available dating evidence as fully as possible.
4. *Variable uncertainty at undated levels*: that uncertainty in dates should be larger at depths that are far from layers for which absolute dating evidence is available. Conversely, uncertainty in dates should decrease at depths where many dated samples lie near to one another. Note that this property is fundamentally different to that of standard linear regression, where a simplistic and unrealistic assumption of constant variance is usually made.
5. *Outlier detection*: that dating evidence which conflicts with other evidence from the same sequence should be removed or adjusted to produce a robust chronology which satisfies all of the other desirable properties. The most obvious example of an outlier occurs when a single piece of dating evidence lies outside the age-depth ordering implied by the remainder of the evidence. Much more complicated situations exist, however, where groups of dating evidence may be selected together in favour of other groups. We discuss this further in Section 7.

More technically, the prior distribution on the chronology is an interpolation or smoothing model based on the calibrated radiocarbon determinations. The most advanced models perform the radiocarbon calibration step in conjunction with the interpolation/smoothing step to provide joint chronology estimates. The resulting estimates may then be compared with the calibrated radiocarbon dates to give a fuller picture of the effect of the prior on the age–depth relationship. Such comparisons are usually shown graphically on an age–depth plot.

In the following sections we discuss some of the currently popular tools for creating chronologies, with reference to the desirable properties stated above. We first discuss some traditional approaches which violate many of the key properties, and follow this with a more detailed look at three of the recently developed Bayesian packages. It is important to note that all of these latter packages use a near-identical likelihood term, but differ in the construction of their prior distribution. Solving the Bayesian equation in each will give sample-based joint estimates of dates at given depths. As with any other simulation-based Bayesian algorithm, there will be many thousands (or even millions) of these samples which are then summarised. As the end-user is usually not interested in the individual samples, they are often not output by the package. Instead posterior estimates are summarised, either as an envelope or a blurred error bar plot of date against depth, or simply by quoting tables of highest posterior density intervals for the individual depths (See the top left panel of Fig. 3 for an illustration). Although rarely used, the original sample chronologies can be of key utility when investigating specific aspects of palaeoenvironmental (in particular palaeoclimate) reconstruction and event identification. We discuss this further in Section 8.

5. Previous statistical models for chronology reconstruction

Perhaps the most simple method for creating a chronology is to join the means (or other summary statistic) of the calibrated dates via linear interpolation and thus allow estimation of dates at other depths via interpolation. Such an approach violates nearly all of the desirable properties mentioned in the previous section. Most obviously, the radiocarbon uncertainties are not included, and no

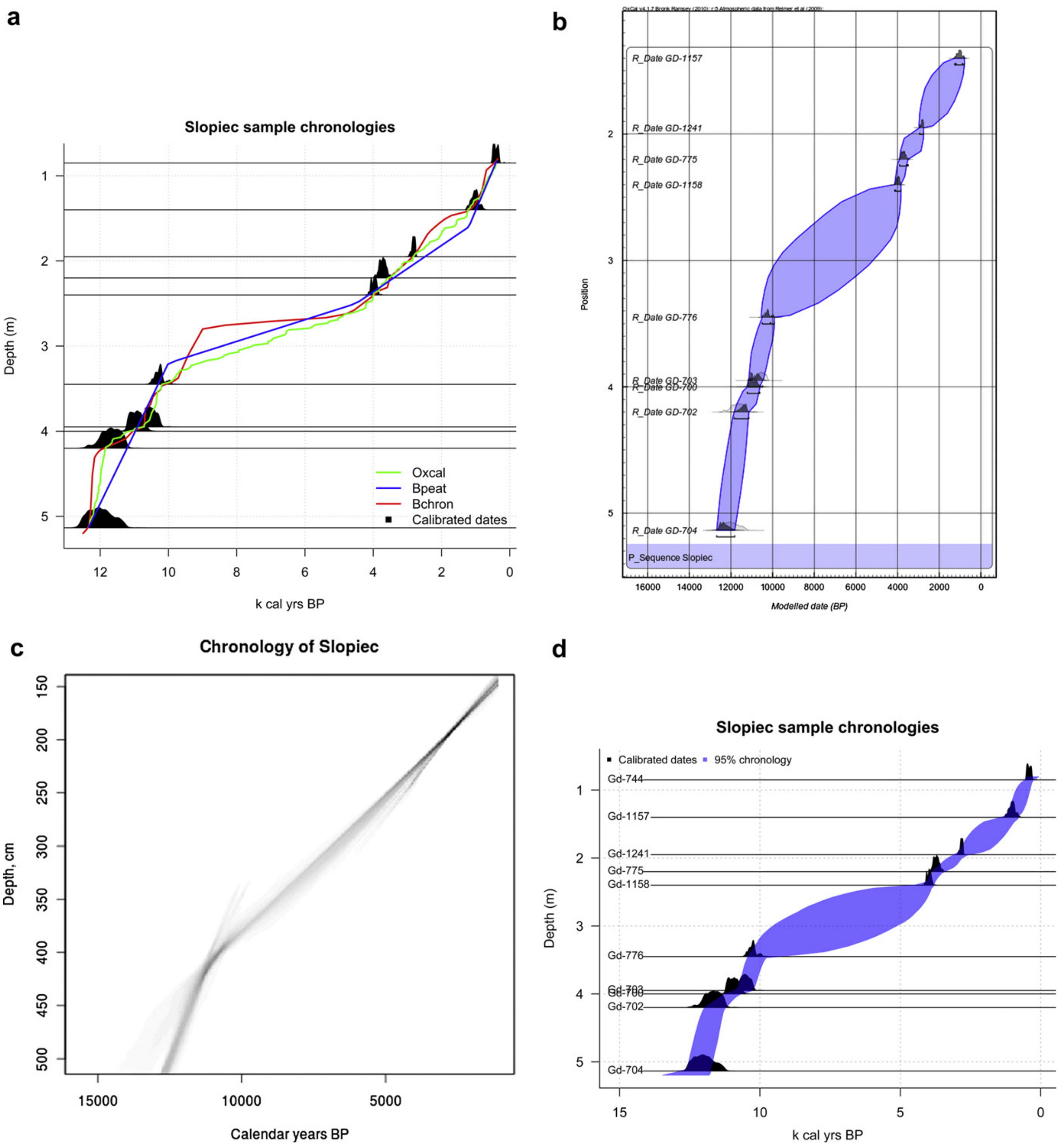

Fig. 3. Chronologies and associated summaries for the Słopiec core. Top left panel: individual sample posterior chronologies from each of the three models. Notice that OxCal chronologies are characterised by a step function, Bpeat by sections of long straight lines, and Bchron by stochastic linear segments. The top-right panel shows the 95% credibility intervals from running the OxCal model on the Słopiec data with a fixed *k* value of 3. The bottom left shows the Bpeat output with the number of sections set at 5 (the grey scale does not give an indication of the error probability range). The bottom-right panel shows the Bchron 95% posterior credibility intervals.

single summary measure will adequately describe a calibrated radiocarbon date (Telford et al., 2004). Second, no uncertainty will be included for other undated depths, and thus there is no flexibility in the behaviour of the chronology. Third, there is no obvious way to deal with outlying scientific dating evidence and so chronologies may be non-monotonic. An extension to this method could be to perform linear interpolation on the 95% confidence intervals for the calibrated dates. Alas, again, this will violate our desirable properties of variable uncertainty, flexibility, and still contains no method to deal with outliers.

A more reasonable method for creating chronologies is that known as ‘wiggle-match dating’ WMD (e.g. Blaauw et al., 2003). Here, a linear regression model is proposed which aims to match the calibrated radiocarbon dates from our core as closely as possible to the calibrated dates that make up the calibration curve. This method works by noticing that a proposed linear regression chronology will give an estimated calibrated date for every depth in the core, including those for which there are radiocarbon determinations. By plotting the calibrated dates implied by the regression model with the radiocarbon determinations, and overlaying the calibration curve, we can get an estimate of the strength of fit of the proposed chronology. Adjusting the linear regression, or optimising it by least squares or maximum likelihood, will provide a closer fit between the calibrated dates from a core and those in the calibration curve, and thus a more satisfactory chronology.

The WMD approach highlights one of the key differences between the many chronology models. In many situations, the approach enforces the linearity constraint at the expense of a good fit with the radiocarbon date distributions, violating our matching uncertainty property. It imposes monotonicity by insisting that the linear regression slope coefficients be positive. Similarly, it will remove outliers simply by ignoring any date estimates that do not fit within the bounds of the best-fit linear regression. If a constant variance assumption is made (which it is in nearly all existing WMD methods) the uncertainty on dates at undated depths is constant (i.e. does not increase away from dated levels).

An extension to simple WMD adopted by Blaauw et al. (2003) is to use separate linear regressions for different parts of the core. This is most often implemented by choosing fixed depths where a change in sedimentation rate is hypothesised, and performing standard WMD in each section. This creates a far more flexible chronology structure, but does not yield sufficient variable uncertainty, and makes no attempt to match the uncertainty to the shape of calibrated radiocarbon date distributions. A more subtle problem is that, if the variance parameters of the different linear regressions are allowed to change between sections, the size of the uncertainty may jump unrealistically. Thus at depth d we may have variance v_1 but at $d+\delta$ (with δ a very small number) we have variance v_2 much bigger (or much smaller) than v_1. We urge caution in allowing such a jump in variability of the chronology in the space of just a few millimetres of depth without strong *a priori* evidence. A software version of this model has been implemented in a package known as CLAM (Blaauw, 2010).

Finally, an entirely separate model for creating chronologies is proposed by (Heegaard et al., 2005). They use a mixed-effects non-parametric regression to estimate the chronology. Their proposed model involves pre-calibrating the radiocarbon determinations, and using the upper and lower 95% HDR interval values to create a smoothed estimate of the mean chronology with uncertainty. They include an extra variance term to account for radiocarbon determinations that may lie away from the chronology which thus allows them to ignore outlying data automatically. The key advantage of this method is its flexibility, in that it can produce smoothly varying non-linear monotonic chronologies with smooth continuous variances. The main disadvantage, however, is the pre-calibration of radiocarbon determinations, and the use of only the upper and lower interval values to compute the chronology. This has the effect of ignoring the multi-modal structure of the calibrated dates, and will produce chronologies with uncertainties whose characteristics do not match those of the associated dates.

6. New Bayesian chronological models

We now turn to the current generation of chronology models. These can be characterised by the following common features.

- They are all implemented in freely available software or provided as pre-compiled software, are free of charge to the end-user and are packaged for use by non-specialists[1]. Those using them require no statistical expertise, but they do need to read carefully the papers that launched the software in order to be sure that they understand the assumptions on which the various models are based.
- They all use a specific, well-defined monotonic stochastic process as a prior distribution.
- They all share the standard likelihood function used for calibrating individual determinations and include prior knowledge in the calibration process.
- They all provide methods for dealing with outlying date estimates, either by suggesting which should be removed or by automatically removing them as part of the model estimation stage.
- They produce posterior samples, each of which is a joint estimate of a chronology, i.e. a set of dates for given depths. Each of these sets is treated as equi-probable, and thus can be summarised to produce probability distributions of date for a given depth, or age–depth plots with uncertainty estimates.
- Finally, they each use an algorithm known as Markov chain Monte Carlo (MCMC) to estimate the model parameters (or equivalently solve Bayes’ equation). The version most often used is a rejection algorithm which proposes parameter values and accepts or rejects them depending on how well they match the likelihood and the prior. In order to create an adequate number of accepted samples, the algorithm will be run for many thousands or millions of iterations.

Readers will note that one of the models on which we are focussing (Bchron) was developed by the lead author. Given this, we should make it clear that this paper is not written in an attempt to endorse one particular model over the others. Rather our motivation is to ensure that users of Bayesian age–depth modelling software are well informed about the key differences between the models that they implement and are thus better able to make informed choices about which model to use for any given core on which they are working. We note that all three models exhibit more of our desirable properties than the more traditional approaches to chronology construction, but observe that they all still have their limitations. Fig. 3 shows each of the models as applied to the Słopiec core in the format presented by the computer package (i.e. a summarised set of individually sampled chronologies). The top left panel shows a possible sample joint chronology from each.

In the following sections we highlight the main features of each of the three models, and discuss when each may be most appropriate for use. A more technical discussion can be found in (Haslett

[1] Two of the models (Bpeat and Bchron) use the open-source, free statistical package R (R Development Core Team, 2010), which was also adopted by Heegaard et al. (2005). For those looking to improve their chronological and statistical modelling skills, a course in R is highly recommended.

and Parnell, 2008). Each piece of software has other features that we do not discuss; individual users may find these useful, but we are not attempting an exhaustive study. Rather, we focus on the particulars which affect the fundamental shape and uncertainty of the resulting chronologies.

6.1. Bpeat – (Blaauw and Christen, 2005)

The Bpeat model runs in the statistical package R though is based on C^{++} code, and outputs most of the results in html format for easy viewing in a web browser. The model can be thought of as a fully Bayesian, advanced and automated version of the WMD model as described in Section 5. Users are required to know a small amount of R to run the program, though there is also a menu system for first-time users. The input files are in tab-separated, text format and contain the (uncalibrated) radiocarbon ages, the associated laboratory errors, and the depth at which the relevant samples were obtained.

The prior model implemented in Bpeat is a linear regression model with a fixed number of changes in sedimentation rate. The important advance over previous WMD models (aside from the Bayesian modelling and specific software implementation) is that the depths at which changes in the sedimentation rate occur are automatically identified by the model. Thus the user can, for example, specify that there are three sections in the model, and the Bpeat program will find the depths at which the two sedimentation rate changes occur. This type of model is often known as change-point linear regression (Carlin et al., 1992).

The key advantage of such an approach over previous WMD approaches is that the data drive the identification of the depths at which sedimentation rate changes occur; i.e. such information is not seen as prior knowledge that must be known separately from the radiocarbon data (although external knowledge can also be added if it is available). The locations of such changes are thus estimated with appropriate uncertainty.

The main disadvantage of the approach is that it is still WMD and thus the posterior chronological uncertainty for the depths at which we do not have radiocarbon determinations do not have the properties we desire; in particular, depths close to radiocarbon samples do not have the multi-modal posterior probability density that we would expect. That said, the non-matching variance problem is partially solved by using the same linear regression variance throughout the core, which makes the model simpler, if not more realistic. Another key point is that, if the number of sections is under-estimated by the user it cannot be changed by the algorithm and so it will tend to ignore many of the radiocarbon determinations from the core in order to fit as well as it can to the data. Conversely, we can get a more flexible chronology (one that goes closer to the radiocarbon data) by increasing the number of sections, even well beyond the number of dated levels. However, this leads to model estimation problems due to the large number of parameters which now need to be estimated (for example a slope parameter for every section).

An important additional feature of the Bpeat model is the ability to detect possible hiatuses in the sedimentation rate. This is accomplished by allowing gaps in the dates between the sections. Thus one section with one sedimentation rate may end at a depth d and with date θ, but the next section may not start until $\theta - \gamma$, where γ is the length of the hiatus. A prior distribution is given for the value of γ, and this may be changed by the user.

Outliers are automatically removed by the Bpeat model according to the outlier detection algorithm given by (Christen, 1994). The method for doing so, often known as 'flag and shift', uses a binary 'flag' parameter and a normally-distributed 'shift' parameter to down-weight the impact of radiocarbon determinations which are considered outliers. The model creates an associated flag and shift parameter for each of the radiocarbon determinations. The posterior proportion of flag parameters equal to one for each determination is often used as an estimate for the probability that it is an outlier. An *a priori* probability distribution is given for the proportion of non-zero flag parameters which can be set by the user. The default is 0.05, corresponding to the findings of the radiocarbon fourth international assessment report (Scott, 2003).

Once the model is fitted, posterior plots and summaries of the data are presented in html format so that they can be viewed in a web browser. Users are offered grey scale plots of the fitted chronologies and estimates for the calibrated ages at individual depths. An additional feature of the software allows for plotting of a proxy variable alongside the estimated chronology to allow for comparison of different time series. This may be used for judging the temporal uncertainty in a proxy signal from the same core.

The usual procedure for fitting the Bpeat model involves choosing a range of values for the number of sections (usually starting at 1, 2, 3, etc) and fitting the model for each value. Bpeat provides a model summary statistic F which approximately corresponds to the proportion of determinations that are actually used in producing the final chronology. In general this will always increase with the number of sections, but can give a useful guide to the removal of poor models. Fig. 3 (bottom left) shows a Bpeat model applied to the Słopiec data.

6.2. OxCal sequence models – (Bronk Ramsey, 2008)

OxCal is a large and full-featured suite of radiocarbon dating tools (Bronk Ramsey, 2001; Bronk Ramsey, 2008) offered as pre-compiled code to run on a number of operating systems, or in a web browser. Users input age/depth data into OxCal via the standard OxCal syntax language. The models that we will review here are the Bayesian chronological models implemented through commands suffixed with the word *sequence* which are used to represent various forms of prior distribution in the context of sedimentation deposition. The simplest of these (called simply *sequence*) does no more than restrict the calibrated dates for the dated layers to lie in depth order and so performs no interpolation for undated depths. Other commands (*D_sequence*, *U_sequence*, *V_sequence*) impose various assumptions on the sedimentation rate. We confine our discussion to the most general of these, known as *P_sequence*.

The prior model used in *P_sequence* is fundamentally different to that used in Bpeat. Instead of fitting a linear regression-type model over many dated layers, the OxCal prior distribution is based on *differences* in the dates of layers. For example, if θ_1 represents the calibrated date at depth d_1 and θ_2 represents the calibrated date at a lower depth d_2, then θ_2 must be older than depth θ_1 and the difference $\theta_2 - \theta_1$ may only take positive values. The *P_sequence* prior model uses a probability distribution which only allows the differences in dates between the dated depths to take positive values. We call this the *increment probability distribution*. Readers may be aware of many standard probability distributions that only take positive values, for example the Exponential, the Poisson or the Gamma distributions.

There are several advantages to using such an approach which include: a large range of very flexible probability distributions, the guarantee of a monotonic function, and the fact that the prior model makes no statements about individual dates, only the differences between them. Furthermore, the probability distribution can be adjusted so that its mean, variance or other components depend on the depth differences between the layers ($d_2 - d_1$ in the above example), creating an enormous variety of different chronology shapes. An estimated calibrated date for a depth at which we do not have radiocarbon data can be created simply by knowing the difference between the non-dated depth and the dated depth, and then simulating from the required probability distribution.

In practice, simulating dates for non-dated depths is slightly more complicated, as we will usually have data on the dates above and below the depth we are interested in. The problem then becomes one of *bridge* sampling (i.e. drawing sample paths between two end-points). In many simple statistical cases we can do this exactly. For example if the increment probability distribution is Gamma, the bridge sampling distribution is Dirichlet. Similarly, if the increment probability distribution is Poisson, the bridge sampling distribution is Binomial. The OxCal *P_sequence* model uses the latter.

Given an increment probability distribution, OxCal will draw joint posterior samples to form an estimated chronology. The nature of the increment probability distribution is such that the resultant chronologies look like step functions, or staircases, where each step is of an equal size. This size is controlled by a parameter k, setting k large will give a less variable step function (recall that the variance of a Poisson distribution is equal to its mean, here related to $1/k$), whereas setting k small will produce a function with much more variability. Bronk Ramsey (2008) gives details as to how the value of k might be determined.

The parameter k is fixed by the user before the model is run. This act proves to be one of the main advantages *and* disadvantages of the OxCal *P_sequence* model. The flexibility allows users to set directly the degree of uncertainty in the resulting chronologies. Setting k small produces a very variable chronology with clear increasing (or ‘bowing’) of the uncertainty between the dated depths. This may be desirable in cores which have very variable sedimentation rates. Setting k large gives low uncertainty and thus little ‘bowing’ effect, and is possibly best applied to cores with few changes in sedimentation rate. Unfortunately, the fact that the user has near-total control on this variability allows for the uncertainty to be decided upon without any input from the given data.

The OxCal package treats outliers in a different way to Bpeat. Instead of associating flag and shift parameters with the radiocarbon determinations, an *agreement index* is produced, which measures the degree of overlap between the posterior distribution and that of the likelihood for the calendar date at each depth for which a radiocarbon determination exists. A user can make a decision based on this index as to whether a rerun of the model is needed in which some of the data is treated as outlying and hence removed. The default agreement index threshold is set at 60%. Unlike the flag and shift methodology described above, the 60% threshold is not informed or updated by data from the core. Alternative options for outlier removal are available (see Bronk Ramsey, 2009).

The OxCal sequence commands allow for a number of extra enhancements to chronological modelling which may be of use to many. The sequences can be separated into boundaries where clear changes in sedimentation rate occur, and different k parameters can be used in each. However, the depth of these boundaries is set by the user, rather than being chosen by the model as in Bpeat. Fixed calibrated dates for certain layers (perhaps known from external prior information) can be included with the *date* command and scientific dating evidence from methods other than radiocarbon can be accommodated too. A number of other useful tools for specific types of sequence are also available, but we do not discuss them here; (see Bronk Ramsey, 2001) and (Bronk Ramsey, 2008) for details.

In summary, the usual procedure for running an OxCal model is to decide upon a k value and any appropriate boundary changes, and then run the *P_sequence* model. Runs may be repeated with different k-values or with some data removed if the agreement indices suggest that one or more radiocarbon determination is an outlier. Summary plots of the posterior chronologies are produced, though the individual joint samples from individual steps of the MCMC chain are difficult to obtain. A final *P_sequence* model for the Słopiec data is shown in the top-right panel of Fig. 3.

6.3. Bchron – (Haslett and Parnell, 2008)

The Bchron model is implemented as a downloadable R package. Like the other models, it takes as input the radiocarbon determinations, associated laboratory errors and depths for the samples in a single core, and outputs joint chronological samples which can then be summarised in an age–depth plot. The standard Bchron input file also allows for uncertainties associated with use of bulk samples to be included by incorporating information about the thickness of each sample, and for chronological information from different types of dating evidence to be given.

Like OxCal, the Bchron model centres on estimating probability distributions for the date increments between the depths in the sediment. As before, the Bayesian method allows for the estimation of the parameters of the chosen increment probability distribution. The distribution used by Bchron is the Compound Poisson-Gamma also known as the Tweedie distribution; (Kaas, 2001), and is equivalent to letting the k parameter defined by OxCal have a Gamma distribution with unknown variance. The resulting joint posterior chronology samples no longer appear like step functions but give instead linear piece-wise sedimentation episodes.

The linear piece-wise chronologies produced by Bchron give the superficial appearance of a cross between the OxCal and Bpeat models. Similar to OxCal, the chronologies will have increased uncertainty (and hence bow) away from the dated levels, yet like Bpeat there will be random change-points in the sedimentation rate, determined by the model and the data in a Bayesian fashion. Unlike Bpeat, however, the user is not required to have any input into the number of changes in sedimentation or their locations in the core.

Bchron provides for a rich family of possible chronologies. However, one key disadvantage of its approach (and that of increment models in general) is that a change in sedimentation rate is assumed to occur at each depth for which a radiocarbon sample is available (though that change may be small). There is, of course, no reason to believe that a change in sedimentation will really have occurred at the exact depth of every dated level. Another disadvantage is that, unlike OxCal, the sedimentation parameters are shared across the entire core. Thus, although the Bchron model may be very useful for producing flexible chronologies, it does not allow much opportunity (especially compared to OxCal) for users to individually influence the chronology behaviour.

Outlier handling, as implemented by Bchron, allows for two different types of flag and shift. The first type permits small shifts for determinations that need a small adjustment, perhaps as a result of (small) intrusions or some residual material in the sample. The second type allows for the complete removal of determinations which lead to calibrated date estimates that lie well beyond the standard age–depth relationship. When such data are identified, Bchron completely ignores both their approximate location and the shape of their calibrated uncertainty. By default, all radiocarbon determinations are assumed to have *a priori* probabilities of 0.05 and 0.001 (respectively) of being each type or outlier. The values can be changed by the user and are updated by the model to provide posterior estimates in the standard Bayesian way.

The usual method for running Bchron is to create an input file, and set appropriate outlier probabilities. The software then runs in two stages. At the first stage it simply produces estimated values for the parameters of the increment probability distribution and the outlier probabilities. At the second stage it creates chronologies and thus produces suitable joint samples and age–depth plots. A picture of the standard Bchron output for the Słopiec core is shown in the bottom-right panel of Fig. 3.

7. A large data survey of Bayesian chronology models

In this section we perform a large-scale comparison of the three models outlined in the previous section. This comparison is not meant to be exhaustive, but will establish the kinds of cores and sedimentation rates which are best suited to each model, and offer a general performance indicator. Most importantly we use the uncertainty in the chronologies rather than single measures of best fit (e.g. the mean or the mode, (see Telford et al., 2004 for criticism) to guide our model performance assessments.

Many statistical models are tested by investigating how well they fit to simulated data. Such tests reveal any problems in estimating parameters (especially in the case where the data are simulated from the model under test) and will show how robust the model is to violations of its assumptions (see Haslett and Parnell (2008) Section 3.5 for an example). However, if we only fit the model to simulated data we can only give an indication of model performance in idealised circumstances. For present purposes, we prefer to use a large amount of real data, from the European Pollen Database (EPD). Our comparison of the three models uses 111 cores taken from Austria, France, Germany, Greece, Ireland, Italy, Sweden and the UK. The list comprises almost every core that the EPD holds for these countries; we have removed those with less than 5 radiocarbon determinations lying in the range of the IntCal09 calibration curve (Reimer et al., 2009). Further restrictions on the cores used are given below. The key advantage of using real data is that the models must try to replicate real-world sedimentation rates in a variety of situations, rather than convenient and artificially-simulated values.

The tests we perform use a technique known as leave-one-out cross validation (often written 1-CV). The technique involves systematically leaving out each individual radiocarbon determination in each core, fitting our chosen model, and investigating the impact that this has on our chronological estimate. The investigations involve comparing appropriate probability distributions. In our tests, the distributions that we compare are probability density functions (pdfs) for calibrated date estimates at particular depths in the core. For each core, we first calibrate all of the radiocarbon determinations individually and thus obtain individual (independent) estimates of the calibrated radiocarbon dates at each depth in the core (henceforth we will call these *individually calibrated pdfs*). We then create *model-derived pdfs* from fitting the chosen model with this date removed and creating the age pdf at the depth of the removed date. We compare the individually calibrated pdf with the model-derived pdf.

More technically, for each model and each core, we systematically leave out one radiocarbon data determination at a time (here called the 'left out' determination or data point). We then construct the chronology from the remaining data points and compute the (marginal) posterior probability density function for the calibrated date estimate at the depth where the left out data point is known to lie. These probabilistic date estimates, derived from the results of a single 1-CV experiment, are the model-derived pdfs. A schematic diagram for the whole 1-CV process is shown in Fig. 4. Note that since we repeat the 1-CV method for every radiocarbon determination in every core, it is a very large computational and time-consuming task, since some of the models are not built for batch use.

An alternative comparison may be proposed in which the calibrated date estimates obtained using the three chronological models are compared against calibrated date estimates obtained from a model that simply imposes the stratigraphic order implied by the known depths. Intuitively, such an approach seems superior because it seems inappropriate for us to work with any date estimates that do not conform to the *a priori* age-depth order. However, the radiocarbon determinations in real cores suffer from numerous outliers and so no depth constraint could be used without also using appropriate outlier removal. Since the method of outlier removal is one of the many key differences between the models that we wish to compare, we have decided not to use any depth information in the calibration of individual determinations.

Having established appropriate methods for deriving pdfs, we now need a way to compare them. Unfortunately, there is no single right way to undertake such comparisons. We use three different measures.

Modal distance the number of calibrated years that separate the modes of the individually calibrated and model-derived pdfs. We choose this measure because it offers a simple way to tell how far apart the two pdfs are. The further the modal value of the pdfs are apart the worse the model could be considered to be performing.

Proportion of 'contained' pdfs the proportion (or percentage) of occasions on which the 95% credible interval for the pdf of the individually calibrated ('left out') determination is contained entirely within the model-derived pdf for the calibrated date estimate at the relevant depth. Intuitively, for a good chronological model, we would expect this proportion to be high since leaving out a data point should lead to increased uncertainty in the calibrated date at the depth from which the ('left out') data point was taken and thus to a model-derived pdf that has a wider 95% credible interval than that for the individually calibrated pdf. Note that we should not expect 95% of dates' ranges to lie within the 95% interval, largely because of the presence of outliers. However, we may expect a good model to contain a higher proportion of these intervals. Of course, it would be possible to create a model for which the model-derived pdf was so large as to always contain the 95% interval for the radiocarbon date. Thus this statistic should not be taken in isolation.

Kullback-Leibler divergence measure (Kullback and Leibler, 1951) quantifies distance between two pdfs (as opposed to a measure of distance between summary statistics such as modes). It is defined as:

$$KL(p,q) = \int p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x$$

where p and q are the model-derived and individually calibrated pdfs respectively. It is not a symmetric distance measure (reversing p and q in the above equation will yield a different KL value), but will provide a more complete measure of the performance of the chronological model. Under this measure, a model which produces a similar mode will be penalised if higher moments (e.g. the variance) do not also match between the two pdfs.

In summary, a good model will have (1) small, (2) large and, most importantly, (3) small. We create these summary statistics for a complete set of leave-one-out experiments for every core, for each of the three chronological models.

We note from the previous sections that the Bpeat and OxCal models require various parameters to be set *a priori* which are then not updated during the model run. For Bpeat we need to set the number of sections, and for OxCal we need to set the k parameter as well as any boundary points. Running the cross-validation exercise for every possible combination of these is computationally and practically infeasible, so we run these models with set values. For the Bpeat model, we set the numbers of sections to be 5, and for the OxCal model we use $k = 3$ with no boundaries (see Blaauw and Christen, 2005; Bronk Ramsey, 2008, for further guidance). This k

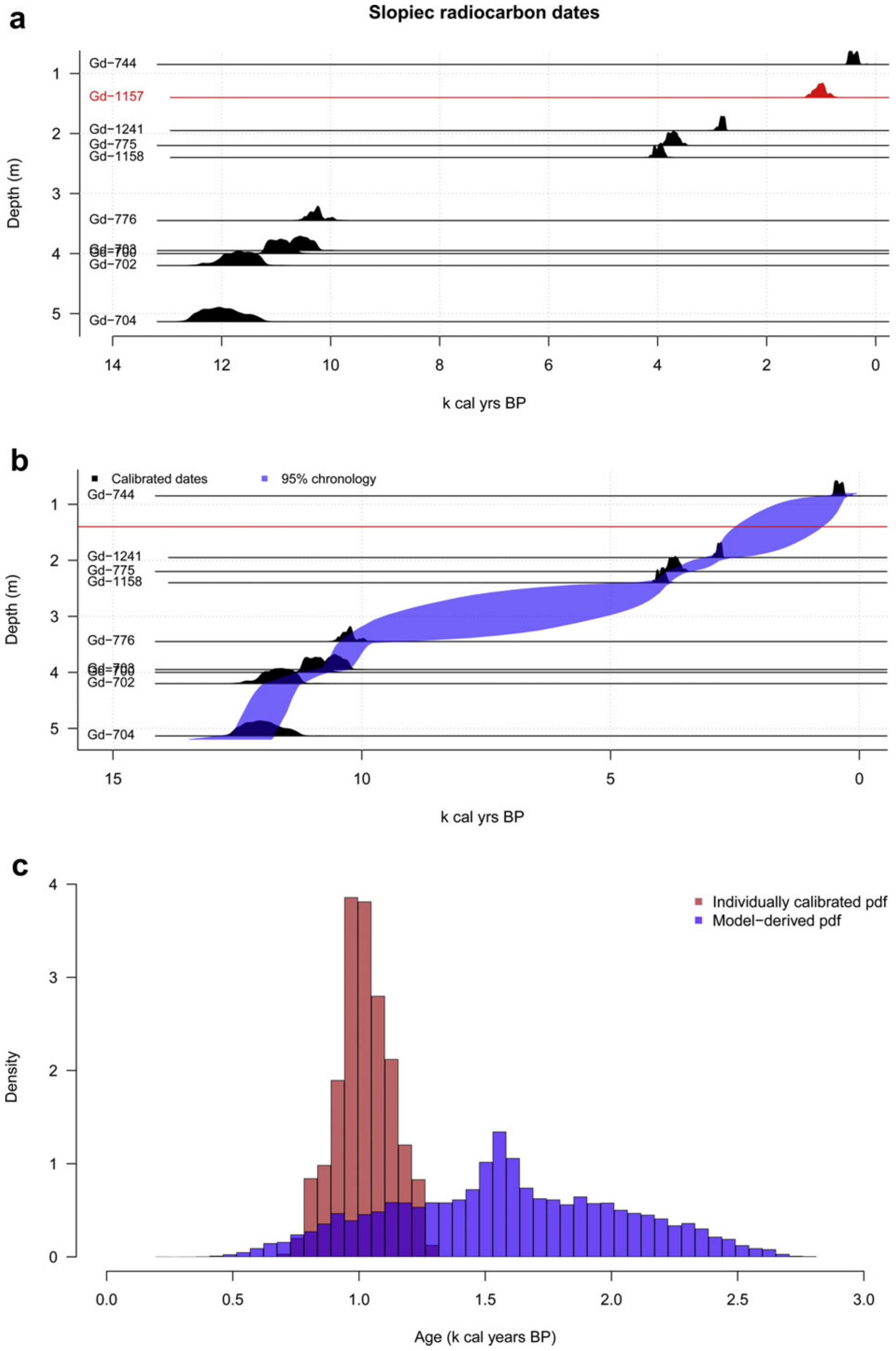

Fig. 4. Schematic diagram of the leave-one-out cross validation (1-CV) process for Słopiec using Bchron. In the top panel are the results of individually calibrating each radiocarbon determination in the core (without imposing any *a priori* chronological information) and then plotting them in stratigraphic order (these are what we call 'individually calibrated pdfs'). Determination GD-1157, from depth 140 cm, (whose individually calibrated pdf is shown in red) is then removed and the chronological model fitted (using either Bpeat, Bchron, or OxCal). The middle panel summarises the resulting estimated chronology, together with a red line at depth 140 cm (where GD-1157 was removed). The lower panel shows an enlarged version of the individually calibrated pdf (red histogram) from the top panel overlain with the model-derived pdf at depth 140 cm extracted from the output used to produce the middle panel. The information used to plot these two pdfs can be used to: calculate the difference in modes, assess whether the 95% HDR for one estimate is encompassed by that for the other, and compute the Kullback-Leibler divergence (see the main text for explanation of each of these).

value is used in (Bronk Ramsey, 2008) for a lake sediment model so seems appropriate for use with the majority of EPD cores. We note that it is considerably larger than that used in other published works (e.g. Staff et al., 2010), where $k = 0.4$. Of course, in the analysis of some individual cores there may be considerable local knowledge which can guide the fitting of the models. We provide the code used to run these comparisons so that readers can expand the data used and improve upon our tests, either *en masse* or for individual cores, should they wish (for more information see the Supplementary material).

Before moving to discuss results, we need to highlight some practical issues and note how we address them.

Convergence problems: the MCMC algorithms that Bpeat, OxCal and Bchron use require a large number of iterations to produce reliable (i.e. converged and well mixed) results. In our experiments, we have used the default values set by the individual authors and have not run any convergence checking ourselves. OxCal provides automatic convergence checking and diagnosis. However, Bpeat and Bchron run for a set number of iterations before reporting results. It is possible that some of the poor results shown by all three software packages (especially Bpeat) are due to a lack of convergence and/or poor mixing.

Differential outlier handling: in certain circumstances, for cores with strongly outlying data points and when using the default agreement index for handling outliers, OxCal will not produce results. In order to be able to include OxCal in all of our comparisons, we chose to remove all such cores from our analysis (thus reducing the number of core in our study from 138 to the final 111). Readers should note, however, that in doing this we are almost certainly skewing our performance checking in favour of Oxcal (not least because there are similarities between our performance measures 2 and 3 and the agreement index that Oxcal uses for outlier detection). Bchron and Bpeat handle outlying data points automatically and will usually produce valid results even when Oxcal fails to do so. As such, it could be argued that the cores left in our experiment favour the OxCal method.

Within-core averaging Each core will lead to several modal distances, proportions of 'contained' pdfs, and Kullback-Leibler divergence measures (one per radiocarbon determination). To summarise these, we have averaged the values within each core. We do this in an attempt to prevent cores with a large number of radiocarbon determinations from biasing the results. However, it also has the effect of smoothing out individual large or small distances. Since *a priori*, we feel that each of the three chronology models is equally likely to lead to large and small distances, averaging in this way seems unlikely to bias results in favour of one model over the others.

Extrapolation issues In certain circumstances we were unable to produce extrapolated date estimates using Bpeat and OxCal, for example when the 'left out' radiocarbon determination was the one at the top or bottom of the core. In such circumstances we have not calculated distance measures for these particular pdfs.

Fig. 5 shows the results of our comparison. The "violin plots" (top and bottom panels) show both a traditional boxplot and a kernel density estimate, with the median value shown as a white spot. The top panel shows the $\log_{10}$ difference between the modes; both Bchron and OxCal seem to perform better than Bpeat here, with Bchron having a slightly lower median and smaller variance. In the second panel, it is clear that a slightly higher proportion of OxCal derived 95% credible intervals completely contain the 95% credible interval of the relevant individually calibrated pdf (note here that we would expect this, since OxCal does not produce output when it identifies strongly outlying data points and so cores on which OxCal performs worst under this measure have been removed from our study). The third panel shows Bchron to have slightly lower Kullback-Leibler divergences, the variance from OxCal being raised slightly due to a few cores with higher scores.

Upon further inspection, the cores on which OxCal performed worst were those where there was a significant gap (i.e. at least 50 cm) between the depths of two radiocarbon samples. It may be that, for such cores, the k value of 3 produces too extreme an increase in uncertainty on date estimates for depths that are far from depths with data and thus induces a higher mean KL divergence and associated higher 95% CI. Unfortunately, although changing the k value will likely produce a lower KL divergence measure for this subset of cores, it is also likely to produce worse measures for others cores.

8. Discussion: chronology construction in the future

In this paper we have reviewed a selection of models available for constructing age–depth chronologies from radiocarbon determinations derived from a single palaeoenvironmental record (typically a core). There are many other (older) methods for constructing such chronologies, but the three we reviewed are the most statistically sophisticated. All use the Bayesian approach and produce joint samples of chronologies which can then be summarised. The three models differ in the prior structure used to obtain a reconstruction of the sedimentation rate for a single core, and in the way they identify (and possibly remove) the effect of outlying data points. We have conducted a large-sample, real-data comparison between the models to identify the range of performances that may be obtained.

Our research suggests that each of the chronology programmes is useful under different circumstances. Bpeat will be most useful when there are large numbers of outlying data points, as it tends to enforce simple linearity when OxCal and Bchron may struggle to fit. OxCal seems ideally suited to situations where there is strong prior, individual knowledge about the nature of sedimentation in a particular core, though we caution users not to use k-values which produce spuriously precise chronologies unless they have good reason. Bchron is most suited to batch-processing of cores where standard outlier handling, and flexible forms for the chronologies has precedence over the ability to hand-tune results for particular cores.

There are many possibilities for the expansion of these models. Obviously useful advances would include the following.

- Improving the stochastic processes upon which the chronological models are based. Models that can take account of extra information (for example, dating quality), or give more flexibility to the stochastic process, will likely improve performance scores. We hope that such models will be compared to those already in existence using our real-world-data comparison approach.
- Tying together date information from multiple cores where layers are known to be of similar or identical ages. For example, when a clear identical change in proxy signal is observed at nearby locations or when ash from the same volcanic eruption is identified in several cores.
- Using physics-based models of sedimentation or deposition to guide the construction of chronologies (e.g. Merritt et al., 2003), and Klauenberg et al., this volume.
- The use of extra information (such as climate reconstructions) to guide inference about the formation of sediment. For example, using pollen grain size and abundance as a covariate in estimating the sedimentation rate. Such information might, for example, be used to guide selection of the k parameter in the OxCal model.
- Increasing computational efficiency by avoiding slow Monte Carlo simulation techniques.
- Other small improvements relevant to individual applications. For example, the ability to use different types of dating evidence, to vary calibration curves, or to specify prior knowledge in a range of different formats. Some of these are already available in some chronological models.

Finally, we make some remarks on how chronological models might be used in palaeoclimate reconstruction. One method is to take an undated depth, produce the pdf of the age of that particular layer from a chronology model, and then plot it along with a proxy-based palaeoclimate estimate for the same layer. This will produce individual 'blobs'; a climate reconstruction with uncertainty for each individual layer (with the size of the 'blob' indicating

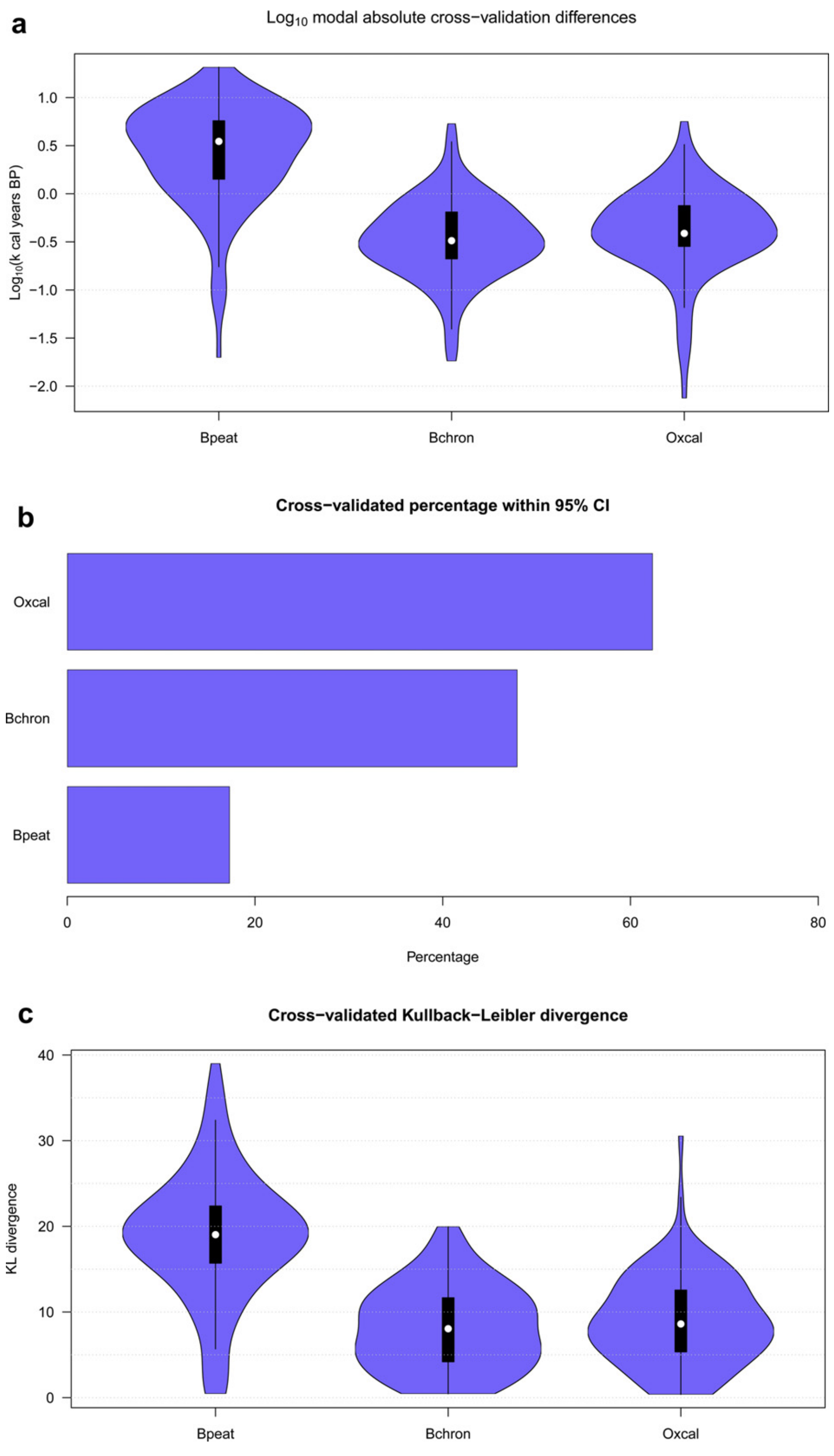

Fig. 5. Summaries of the cross-validated results for Bpeat, Bchron, and OxCal, The top panel shows a "violin plot" (Hintze and Nelson, 1998) for cross-validated modal differences of the relevant pdfs, note the log scale (lower = better). The middle panel shows the percentage of 95% credible intervals for the individually calibrated pdfs that are totally contained within the 1-CV model-derived pdf for the relevant depth in the core (higher = better). The bottom panel shows a violin plot of the Kullback-Leibler divergence between individually calibrated and 1-CV model-derived pdfs (lower = better). The boxplots show the median (white spot), quartiles (black box) and maximum and minimum (whiskers).

uncertainty in both climate and age). For example, at Słopiec, we can produce, using a chronology model, a pdf of the estimated age at 305 cm depth. From the pollen spectra at that depth, we may also be able to obtain a pdf of a climate variable, for example the mean temperature of the coldest month. Combining the two pdfs gives a bivariate density (a 'blob') of the age/climate uncertainty for that particular data point. Bayesian methods could be used to combine many such blobs (i.e. for every depth layer in the core) and thus take account of both chronometric and climatological uncertainty. Unfortunately, such a method will ignore the joint uncertainty in the chronology (as well as any correlation between the climate variable and the dating).

A more advanced approach would be to sample from the joint posterior set of chronologies and perform a climate reconstruction for every depth, drawing one of the joint chronological samples as we do so. The resulting joint climate and chronology reconstruction would have reduced uncertainty as the monotonic information is no longer being thrown away during climate reconstruction. Thus in our example we would now sample an individual complete chronology from Słopiec (like those found in the top left panel of Fig. 3), and produce a reconstruction of the mean temperature of the coldest month for every depth simultaneously. An even more advanced version still (mentioned in the list of desirable extensions above) will reconstruct both the chronology and the climate simultaneously. The current generation of chronology models (i.e. those compared above) do not allow such sophisticated reconstruction, but recent discussions with colleagues suggest that such a methodology may well arrive in the near future.

Acknowledgements

The authors would like to thank the University of Cambridge, the Newton Institute and the organisers of the 'Mathematical and Statistical Approaches to Climate Modelling and Prediction' programme for visiting fellowships. We would also like to thank other members of the SUPRAnet group (Studying Uncertainty in Palaeoclimate Reconstructions), funded by the Leverhulme Trust, with whom they have interacted over the last three years and whose members inspired the work for this paper. Finally, we would like to thank Alex Bayliss of English Heritage for providing some guidance on the manuscript.

Appendix. Supplementary data

Supplementary data related to this article can be found online at doi:10.1016/j.quascirev.2011.07.024.

References

Ambaum, M., 2010. Significance tests in climate science. Journal of Climate 23 (22), 5927–5932.

Blaauw, M., 2010. Methods and code for 'classical' age-modelling of radiocarbon sequences. Quaternary Geochronology 5 (5), 512–518.

Blaauw, M., Christen, J.A., 2005. Radiocarbon peat chronologies and environmental change. Applied Statistics 54, 805–816.

Blaauw, M., Heuvelink, G.B.M., Mauquoy, D., van der Plicht, J., van Geel, B., 2003. A numerical approach to ^{14}C wiggle-match dating of organic deposits: best fits and confidence intervals. Quaternary Science Reviews 22, 1485–1500.

Blaauw, M., Wohlfarth, B., Christen, J.A., Ampel, L., Veres, D., Hughen, K.A., Preusser, F., Svensson, A., 2010. Were last glacial climate events simultaneous between greenland and france? A quantitative comparison using non-tuned chronologies. Journal of Quaternary Science 25 (3), 387–394.

Blackwell, P., Buck, C., 2008. Estimating radiocarbon calibration curves. Bayesian Analysis 3, 225–248.

Bronk Ramsey, C., 1995. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. Radiocarbon 37, 425–430.

Bronk Ramsey, C., 2001. Development of the radiocarbon calibration program OxCal. Radiocarbon 43, 355–363.

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.

Buck, C.E., Cavanagh, W.G., Litton, C.D., 1996. Bayesian Approach to Interpreting Archaeological Data. John Wiley and Sons Ltd, Chichester.

Buck, C.E., Christen, J.A., James, G.N., 1999. BCal: an on-line Bayesian radiocarbon calibration tool. Internet Archaeology 7.

Buck, C.E., Litton, C.D., Smith, A.F.M., 1992. Calibration of radiocarbon results pertaining to related archaeological events. Journal of Archaeological Science 19, 497–512.

Carlin, B.P., Gelfand, A.E., Smith, A.F.M., 1992. Hierarchical bayesian analysis of change point problems. Applied Statistics 41, 389–405.

Christen, J., Perez, E., 2009. A new robust statistical model for radiocarbon data. Radiocarbon 51 (3), 1047–1059.

Christen, J.A., 1994. Summarizing a set of radiocarbon determinations: a robust approach. Applied Statistics 43, 489–503.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, second ed. (Chapman & Hall/CRC Texts in Statistical Science) (2 ed.). Chapman and Hall/CRC.

Hajdas, I., Ivy, S.D., Beer, J., Bonani, G., Imboden, D., Lotted, A.F., Sturm, M., Suter, M., 1993. Ams radiocarbon dating and varve chronology of lake soppensee: 6000 to 12000 ^{14}c years bp. Climate Dynamics 9, 107–116. 10.1007/BF00209748.

Haslett, J., Parnell, A.C., 2008. A simple monotone process with application to radiocarbon-dated depth chronologies. Journal of the Royal Statistical Society, Series C 57, 399–418.

Heegaard, E., Birks, H.J.B., Telford, R.J., 2005. Relationships between calibrated ages and depth in stratigraphical sequences: an estimation procedure by mixed-effect regression. The Holocene 15 (4), 612–618.

Hintze, J., Nelson, R., 1998. Violin plots: a box plot-density trace synergism. The American Statistician 52 (2), 181–184.

Kaas, R., 2001. Compound Poisson distribution and GLMs – Tweedie's Distribution Handelingen van het contactforum, 3–43.

Kemp, A.C., Horton, B.P., van de Plassche, O., Culver, S.J., Parnell, A.C., Corbett, D.R., Gehrels, W.R., Douglas, B.C., 2009. Timing and magnitude of recent accelerated sea-level rise (North Carolina, United States). Geology 37 (11), 1035–1038.

Klauenberg, K., Blackwell, P., Buck, C., Mulvaney, R., Röthlisberger, R., Wolff, E.W. Bayesian glaciological modelling to quantify uncertainties in ice core chronologies. Quaternary Science Reviews, in this volume.

Kullback, S., Leibler, R.A., 1951. On information and Sufficiency. The Annals of Mathematical Statistics 22 (1), 79–86.

Lee, P.M., 2004. Bayesian Statistics: An Introduction. Oxford, Oxford, England.

Merritt, W.S., Letcherb, R.W., Jakemanb, A.J., 2003. A review of erosion and sediment transport models. Environmental Modelling and Software 18, 761–799.

O'Hagan, A., Forster, J., 2004. The Advanced Theory of Statistics, Vol. 2B: Bayesian Inference, second ed. Wiley.

Parnell, A.C., Haslett, J., Allen, J.R.M., Buck, C.E., Huntley, B., 2008. A new approach to assessing synchroneity of past events using Bayesian reconstructions of sedimentation history. Quaternary Science Reviews 27 (19–20), 1872–1885.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasmussen, S.O., Andersen, K.K., Svensson, A.M., Steffensen, J.P., Vinther, B.M., Clausen, H.B., Siggaard-Andersen, M.L., Johnsen, S.J., Larsen, L.B., Dahl-Jensen, D., Bigler, M., Rothlisberger, R., FIscher, H., Goto-Azuma, K., Hannson, M.E., Ruth, U., 2006. A new greenland ice core chronology for the last glacial termination. Journal of Geophysical Research 111 (D6).

Reimer, P.J., Baillie, M.G., Bard, E., Bayliss, A., Beck, J.W., Blackwell, P.G., Ramsey, C.B., Buck, C.E., Burr, G.S., Edwards, R.L., Friedrich, M., Grootes, P.M., Guilderson, T.P., Hajdas, I., Heaton, T.J., Hogg, A.G., Hughen, K.A., Kaiser, K.F., Kromer, B., McCormac, F.G., Manning, S.W., Reimer, R.W., Richards, D.A., Southon, J.R., Talamo, S., Turney, C.S.M., van der Plicht, J., Weyhenmeyer, C.E., 2009. IntCAL09 and Marine09 Radiocarbon age calibration curves, 0–50,000 years cal BP. Radiocarbon 51 (4), 1111–1150.

Scott, E.M., 2003. The fourth international radiocarbon inter-comparison. Radiocarbon 45, 135–408.

Staff, R., Bronk Ramsey, C., Nakagawa, T., 2010. A re-analysis of the lake suigetsu terrestrial radiocarbon calibration dataset. Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms 268 (7–8), 960–965. Proceedings of the Eleventh International Conference on Accelerator Mass Spectrometry.

Stuiver, M., Reimer, P.J., 1993. Extended 14C database and revised CALIB radiocarbon calibration program. Radiocarbon 35, 215–230.

Szczepanek, K., 1992. The peat-bog at Słopiec and the history of the vegetation of the Gory Swietokrzyskie Mountains (Central Poland) in the past 10,000 years. Veroff. Geobot. Inst. ETH, Stiftung Rnbel, Znrich 107, 365–368.

Telford, R.J., Heegaard, E., Birks, H.J.B., 2004. The intercept is a poor estimate of a calibrated radiocarbon age. The Holocene 14, 296–298.