**Nonlinear Processes
in Geophysics**

Calibration of a radiocarbon age

D. J. Keenan

The Limehouse Cut, London E14 6N, UK

Correspondence to: D. J. Keenan (doug.keenan@informath.org)

Abstract. The calibration of a radiocarbon age to a calendar date is reviewed. It is shown that the commonly-used programs for calibration sometimes give results that are significantly in error.

1 Introduction

Radiocarbon dating gives an estimate of the year in which an organism died. It involves measuring the concentration of ^{14}C (a radioactive isotope of carbon) in the carbon of the remains of an organism. Initially, it was believed that an accurate estimate of the year of death could be obtained via a simple calculation based on the rate of radioactive decay. During the 1950s, however, it was discovered that such a calculation could be very inaccurate. Instead, the year has to be estimated via a more complicated procedure known as "calibration".

The purpose of the present work is to examine the statistical underpinning of the calibration procedure. Even though calibration has been recognized for half a century, there are, it turns out, still some basic aspects to consider.

2 Overview of radiocarbon calibration

2.1 General principles

The term "radiocarbon" is commonly used to denote ^{14}C, an isotope of carbon which is radioactive with a half-life of about 5730 years. ^{14}C is produced by cosmic rays in the stratosphere and upper troposphere. It is then distributed throughout the rest of the troposphere, the oceans, and Earth's other exchangeable carbon reservoirs. In the surface atmosphere, about one part per trillion (ppt) of carbon is ^{14}C.

All organisms absorb carbon from their environment. Those that absorb their carbon directly or indirectly from the surface atmosphere have about 1 ppt of their carbon content as ^{14}C. Such organisms comprise almost all land-dwelling plants and animals. (Other organisms – e.g. fish – have slightly less of their carbon as ^{14}C; this affects how radiocarbon dating works, and there are methods of adjusting for it).

When an organism dies, carbon stops being absorbed. Hence after 5730 yr, about half of its ^{14}C will have radioactively decayed (to nitrogen): only about 0.5 ppt of the carbon of the organism's remains will be ^{14}C, and if the carbon of the remains is found to be 0.25 ppt ^{14}C, then the organism would be assumed to have died about 11 460 yr ago. Thus, a simple calculation can find the age, since death, from any ^{14}C concentration. (Remains older than about 50 000 yr, however, have a ^{14}C concentration that is in practice too small to measure; so they cannot be dated via ^{14}C).

Ages are conventionally reported together with the standard deviation of the laboratory ^{14}C measurement, e.g. 900 ± 25 ^{14}C BP (^{14}C-dated, years before present). The true standard deviation, though, will often be larger than what is reported, due to non-laboratory sources of error – e.g. the admixture of impurities with the remains.

Although a tree may live for hundreds, even thousands, of years, each ring of a tree absorbs carbon only during the year in which it grows. The year in which a ring was grown can be determined exactly (by counting); so radiocarbon dating can be tested by measuring the ^{14}C concentrations in old tree rings. Such testing found errors of up to several centuries. It turns out that the concentration of ^{14}C in the carbon of the surface atmosphere has not been a constant 1 ppt, but has varied with time. Thus, the simple calculation of age from ^{14}C concentration is unreliable.

Tree rings, though, also provide a solution to this problem. The concentration of ^{14}C in the carbon of an organism's remains can be compared with the concentrations in tree rings.

Tree rings that match, within confidence limits, give the years in which the organism could have plausibly died.

The matching procedure thus provides *calibration* of ^{14}C concentrations. (Calibration via tree rings, though, does not extend back 50 000 yr; other ways of calibrating are therefore being developed.) Ages that are estimated without calibration continue to be reported, and are called "uncalibrated ^{14}C ages", or simply "^{14}C ages".

2.2 An illustrative example

The two most commonly-used programs for calibration seem to be OxCal (Bronk Ramsey, 2001) and Calib (Stuiver and Reimer, 1993). Outputs from those programs are shown in Fig. 1, for a sample whose ^{14}C age is 2500 ± 100 ^{14}C BP.

The plot from OxCal displays a bell curve in light red. A similar bell curve, in greys, is displayed in the plot from Calib. The bell curve represents the sample ^{14}C age, with the scale on the vertical axis pertaining. (Note that the OxCal and Calib vertical scales have different extents.)

The OxCal plot also displays a thick blue line. A similar thick line, in grey, is displayed in the Calib plot. The thick line is the "calibration curve". The calibration curve represents the tree-ring ^{14}C concentrations used in the calibration procedure. There are at least two issues here. First, the calibration "curve" is not a curve in the common sense; rather, each point on the curve has a potential error, which is usually specified by the standard deviation of the measurement: in Fig. 1, the top of the thick line indicates the upper 1σ bound, and the bottom of the thick line indicates the lower 1σ bound. Second, both OxCal and Calib treat the curve as continuous; doing so requires interpolation, because we do not have ^{14}C measurements in continuous time, only for each tree ring, i.e. for each calendar year. (We might not even have a direct measurement for each year, as in practice tree rings are often not measured individually, but in sequences of ten; that is not important, though, for the analysis here.)

The OxCal plot additionally displays a greyed area, along the horizontal axis. In the Calib plot, a similar greyed area is displayed. The greyed area represents the probability distribution of the calendar years for the sample. It is the main output of the calibration procedure. In this example, 95 % of the greyed area lies within the range 814–398 BC, which is displayed explicitly in the OxCal plot. That range is thus a 95 %-confidence interval for the date of the sample.

3 A discrete calibration method

3.1 A formal derivation

Our goal is to determine the probability that a given sample is from year y, for each possible calendar year y. We have two inputs: (1) a calibration curve; (2) the sample's radiocarbon measurement, i.e. a Gaussian distribution for the sample's ^{14}C age.

Choose a non-empty finite set $T \subset \mathbb{Z}$, to represent the possible calendar years, i.e. the years spanned by the calibration curve. Let G be the set of Gaussian probability density functions; choose a function $c : T \to G$, to represent the calibration curve. Radiocarbon ages are to be specified by integers, rounding ages as required. Represent the distribution of the sample's ^{14}C age by a probability mass function q on $\mathbb{Z}$.

By definition, for all $a \in \mathbb{Z}$ and all $y \in T$,

$$\Pr(\text{age} = a | \text{year} = y) = \int_{[a-1/2, a+1/2]} c(y)(x) \, \mathrm{d}x. \tag{1}$$

Denote this quantity by $p_y(a)$. By Bayes' Theorem, and assuming a uniform prior distribution on T (i.e. the calendar years are a priori equally probable),

$$\Pr(\text{year} = y | \text{age} = a) = p_y(a) / \sum_{t \in T} p_t(a). \tag{2}$$

Thus

$$\Pr(\text{year} = y | \text{age distribution } q)$$
$$= \sum_{a \in \mathbb{Z}} q(a) \cdot p_y(a) / \sum_{t \in T} p_t(a) p_t(a). \tag{3}$$

Equation (3) is the principal result of this paper.

In the outer summation, the range is effectively small, because the terms rapidly approach 0. Hence, computer implementation of Eq. (3) is straightforward. That makes Eq. (3) the basis of what might be called a "discrete calibration method".
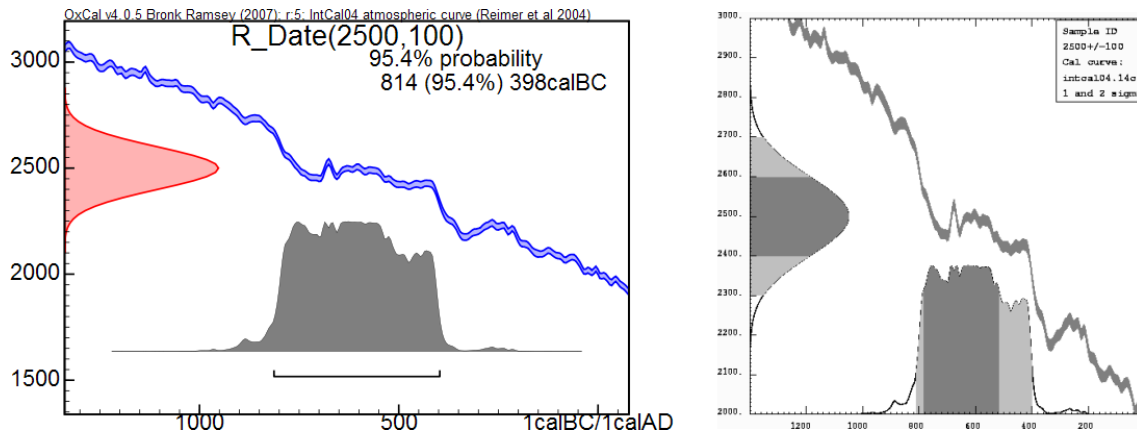
3.2 Implementation considerations

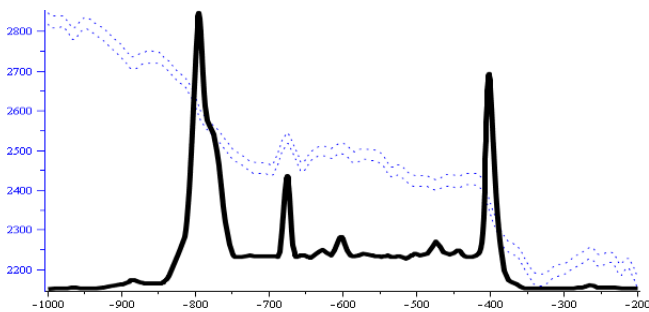In an implementation of Eq. (3), the inner sums can be precomputed, e.g. for all positive $a < 50\,000$.

For computational efficiency, $p_y(a)$ can be approximated. For instance, $p_y(a) \doteq c(y)(a)$. The accuracy of that approximation can be easily checked. Let $c(y)$ have mean μ and standard deviation σ. Numerical tests (not shown) demonstrate that for integer $\sigma \in [5, 500]$, which easily covers the practical range, the inaccuracy is less than the error due to changing μ by 1 – a change that is smaller than what can be feasibly measured. Other approximations for $p_y(a)$ could alternatively be used.

With a continuous variant of Eq. (3), it might be computationally feasible to evaluate the probabilities via numerical integration.

The probability distribution of a radiocarbon age is not exactly Gaussian. Rather, the D^{14}C observation is Gaussian (Stuiver and Polach, 1977), and so the radiocarbon age is actually log-Gaussian. It would be simple to redo the foregoing using a log-Gaussian distribution (or using the D^{14}C observation directly). For radiocarbon, the distinction between the two distributions is immaterial, unless the age is near the measurement limit, and so it is usually ignored in the literature.

Fig. 1. Calibration graph of 2500 ± 100 ^{14}C BP (via the IntCal04 calibration curve; Reimer et al., 2004), using OxCal 4.0 (left) and Calib 5.0 (right). The two programs give essentially the same result.

Fig. 2. Calibration graph of 2500 ± 100 ^{14}C BP (via the IntCal04 calibration curve; Reimer et al., 2004), using the discrete method. The height of the black line indicates the probability for the corresponding calendar year (scale not shown). The 1σ extents of the calibration curve are indicated by the dotted blue lines; the vertical axis, shown in blue, pertains to the calibration curve.

4 Discrepancy with other methods

4.1 Examples

The discrete calibration method described in Sect. 3.1 sometimes gives results that are significantly different than those produced by commonly-used calibration programs. As an example, consider again the age 2500 ± 100 ^{14}C BP. Using the discrete method gives the calibration graph illustrated in Fig. 2. In Fig. 2, the area under the solid black line should be compared with the greyed area along the horizontal axis in Fig. 1: the two are obviously very different.

The calibration graphs from the BCal program (Buck et al., 1999) and the program of Fairbanks et al. (2005) (figures not shown) are very similar to those from OxCal and Calib. Thus, the results from standard calibration programs are very similar to each other – but substantively differ from the result given by the discrete method.

As a second example, consider 4530 ± 50 ^{14}C BP (this example is used by Telford et al., 2004). Calibrating via OxCal and Calib gives the graphs shown in Fig. 3. The calibration graphs from the BCal program and the program of Fairbanks et al. (2005) are very similar to those from the OxCal and Calib programs (figures not shown).

Calibrating 4530 ± 50 ^{14}C BP via the discrete method gives the graph in Fig. 4. As in the first example, the results from standard calibration programs substantively differ from the result given by the discrete method.
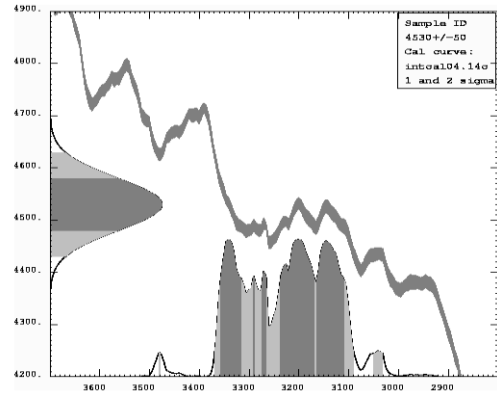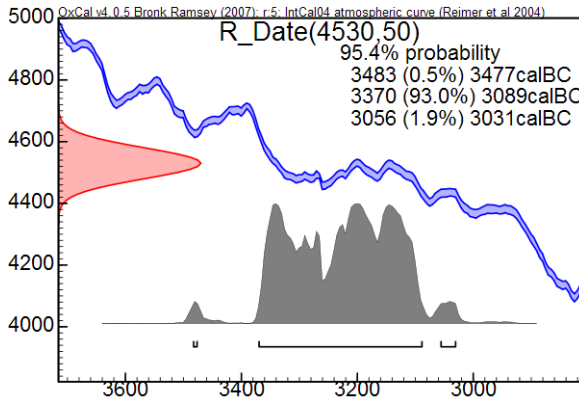
As third example, consider 4300 ± 70 ^{14}C BP. Using the discrete method, the 95.4 %-confidence interval for the calibrated date ends in 2709 BC. Using OxCal 4.0, the confidence interval ends in 2668 BC, and using Calib 5.0 the confidence interval ends in 2666 BC. (All calibrations are via the IntCal04 calibration curve; Reimer et al., 2004.)

Sometimes the calibration graphs produced by the standard programs are essentially the same as those derived by the discrete method. In general, the differences between the standard programs and the discrete method will be immaterial if the calibration curve is steeply decreasing within the plausible age range of the sample. For example, the calibration curve is steep across 2400–2300 ^{14}C BP (see Figs. 1 and 2); so the calibration graphs produced by the standard programs for 2350 ± 15 ^{14}C BP are essentially the same as the graph produced by the discrete method (figures not shown).
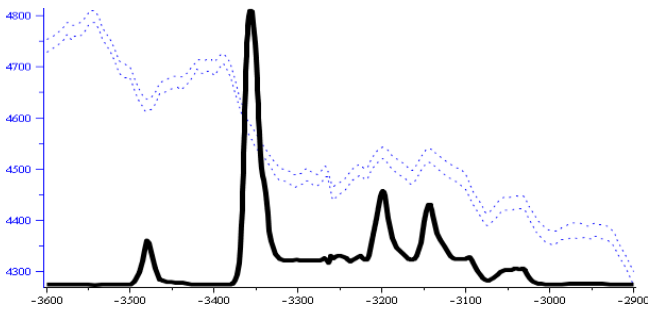
4.2 An intuitive explanation

This subsection presents an intuitive explanation for the discrepancy between the standard programs and the discrete method.
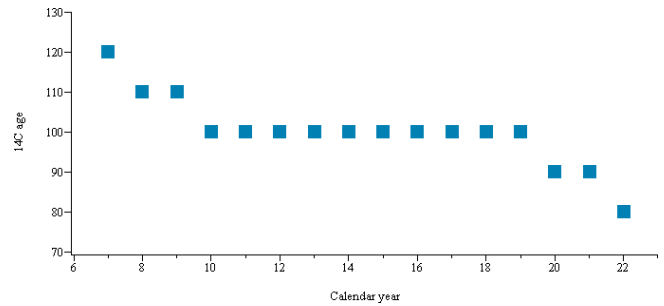
For a first example, suppose that we have both a sample's radiocarbon measurement and a calibration curve with no errors, i.e. all the standard deviations are zero. Assume that the calibration curve looks like that shown in Fig. 5.

Fig. 3. Calibration graph of 4530 ± 50 ^{14}C BP (via the IntCal04 calibration curve; Reimer et al., 2004), using OxCal 4.0 (left) and Calib 5.0 (right). The two programs give essentially the same result.

Fig. 4. Calibration graph of 4530 ± 50 ^{14}C BP (via the IntCal04 calibration curve; Reimer et al., 2004), using the discrete method. The result substantively differs from the result illustrated in Fig. 3.

Fig. 5. An idealized calibration curve, with one measurement for each calendar year and ages exactly 120, 110, 100, 90, and 80 ^{14}C BP.

If the sample's measurement were exactly 100 ^{14}C BP, then the calibration graph would look like Fig. 6 (assuming each calendar year is a priori equally probable).

Notice that the non-zero probabilities in Fig. 6 depend on the number of calendar years that have an age of 100 ^{14}C BP. For instance, if the calibration curve had twenty calendar years with age 100 ^{14}C BP, instead of ten, then the probability that the sample was from any given one of those years would decrease from 0.10 to 0.05. More generally, the formula for the probability of the sample being from calendar year y is

$$\frac{\text{(probability that calendar year } y \text{ has radiocarbon age 100 BP)}}{\text{(number of calendar years that have radiocarbon age 100 BP)}} \quad (4)$$

where the probabilities in the numerator are all either 0 or 1 (in this example).

Next, make the example somewhat more realistic and suppose that the sample's measurement is not known exactly. Instead, suppose that the measurement has a discrete probability distribution. Then the probability of the sample being from calendar year y is

$$\Sigma_a (\text{probability sample has age } a) \cdot \frac{(\text{probability year } y \text{ has age } a)}{(\text{\# of years that have age } a)} \quad (5)$$

where the sum is taken over all possible radiocarbon ages a.

As an illustration, consider the sample distribution given by Table 1. The calibration graph for the sample is easily calculated (even by hand). It is shown in Fig. 7.

It is perhaps worth doing partial visual checks on Fig. 7. For instance, the probability that the sample has ^{14}C age 90 is 1/4 (by Table 1), and there are two calendar years whose ^{14}C age is 90 (by Fig. 5: years 20 and 21); so those two years should have total probability, after calibration, of 1/4, and indeed Fig. 7 displays this. Additionally, the sum of all the probabilities is 1.

Figures 7 and 2 have some similarity because they each illustrate how a broad plain in the calibration curve can lead to a low plain in the calibration graph, with peaks on either side. The standard calibration programs do not produce that shape (see Fig. 1).

The standard calibration programs produce a result that is different from the discrete method because they do not consider how the probability of a given calendar year is lessened if the year has the same, or nearly the same, radiocarbon age as other calendar years. The issue is illustrated in Figs. 6
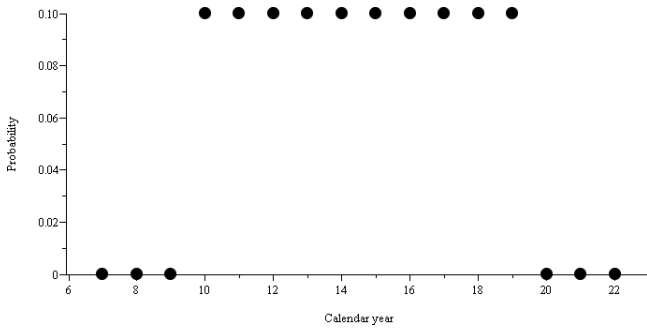
Fig. 6. Calibration graph of a sample whose radiocarbon age is exactly 100 ^{14}C BP (via the calibration curve in Fig. 5).
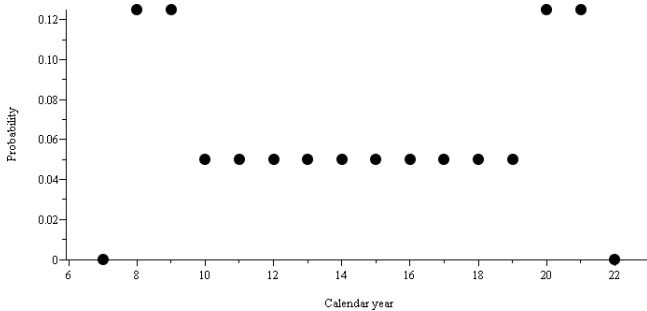

Fig. 7. Calibration graph of the sample whose age distribution is given by Table 1 (via the calibration curve in Fig. 5).

and 7: the more calendar years that lie on the central plain, the lower the probability each year on the plain should have.

The underlying issue can perhaps be seen more easily with a simpler example. Assume that the calibration curve has only three years: 9, 10, 11. Additionally, assume that those years have ^{14}C ages 110, 100, 100, with the standard deviations being zero (as in Fig. 5). Suppose that the sample's measurement has a probability distribution with $\Pr(\text{age} = 110) = 1/2$ and $\Pr(\text{age} = 100) = 1/2$. What is $\Pr(\text{year} = 9)$? To answer the question, note that "year = 9" is true if and only if "age = 110" is true, and thus $\Pr(\text{year} = 9) = \Pr(\text{age} = 110)$; so the answer is $1/2$. The method used by the standard calibration programs, however, would not give that answer, but instead give $1/3$.

The standard calibration programs give an incorrect answer because of the way they treat the sample's measurement. Specifically, if the sample's ^{14}C age is $m \pm s$ ^{14}C BP, the standard programs do not (generally) consider the probability distribution of the age to be Gaussian (or log-Gaussian) with mean m and standard deviation s. For instance, the programs might effectively presume that $\Pr(\text{age} < m)$ is greater than $\Pr(\text{age} > m)$, or they might presume that $\Pr(\text{age} < m)$ is less than $\Pr(\text{age} > m)$ – depending on the calibration curve. That is incorrect, because the sample's radiocarbon age is derived from a lab measurement that has a known Gaussian distribution.

Table 1. Sample's measurement.

^{14}C age	Probability
90	1/4
100	1/2
110	1/4

Finally, the line of reasoning that lead to Eq. (5) can be extended, by supposing that the calibration curve has uncertainty. That is, we take each point on the curve to be a (discrete) distribution. Then the probability of the sample being from year y is similar to what is given by Eq. (5), except that the denominator (# of years that have age a) changes to Σ_t (probability that year t has age a), with t ranging over all calendar years. That leads to Eq. (3) again, i.e. this line of reasoning gives an alternative derivation of the principal result.

5 Combining ages

Occasionally, repeated radiocarbon measurements are made on a single sample, in order to improve precision. Ward and Wilson (1978) describe a statistical method for combining repeated measurements, and their method has become standard in the literature. The same method has also been used when measurements are made on different samples; such use is considered in this section.

Ward and Wilson (1978) consider two distinct cases. Case I is where all the measurements are made on the same sample, which is believed to be homogeneous. In this case, the authors advise first doing a simple statistical test, based on the chi-squared distribution, as a partial check that there was no measurement, or other, error. Assuming that the test is passed, the radiocarbon measurements are combined via a simple weighted average (with the weights being the inverse variances).

Case II is for when "one does not *know* whether all determinations are estimating the same date (or effectively indistinguishable different dates)"; the emphasis is theirs (Ward and Wilson, 1978, p. 21). Case II is thus for measurements of different samples. In Case II, the same chi-squared statistical test is used as in Case I. The authors say that if, as a result of the chi-squared test, "the estimates of the real dates are judged not to be significantly different and, if from archaeological considerations, it is deemed appropriate, then the radiocarbon determinations can be combined" (p. 23). As the authors state, though, there is a "fundamental difference" between Case I and Case II: the simple weighted average that is used for combining measurements in Case I should not be used in Case II. Instead, a more complicated method must be used. An approximating method is described by the authors.

The distinction between Case I and Case II has not always been considered in subsequent literature. In particular, we

sometimes read of different samples, taken from some paleoenvironmental horizon, whose radiocarbon measurements are combined using the simple weighted average of Case I. Such combining is inappropriate, unless it is known a priori and with certainty that all the measurements present the same radiocarbon age. Ward and Wilson were clear and correct in stating this.

6 Discussion and conclusion

A method to calibrate a single radiocarbon age has been described. The method is simple and easy to implement – especially because it is essentially discrete.

The commonly-used programs for calibration have been shown to sometimes give results that are significantly in error. In most cases, the resulting inaccuracy in the bounds of confidence intervals will be very small, but cases where that inaccuracy is sizeable have been presented.

The error in commonly-used calibration programs has been shown to substantially change the shapes of calibration graphs. That change will tend to have some effect on Bayesian analysis of (multiple) radiocarbon ages – a type of analysis that has become common in recent years: see e.g. Bronk Ramsey (2009). Determining when the effect is substantive would require further study. (Additionally, it can be noted that Bronk Ramsey (2009) also presented a derivation of the statistical method employed by the calibration programs; by the results herein, that derivation must be mistaken.)

Finally, the simple combining of multiple radiocarbon ages has been reviewed. In particular, the combining of ages from different samples, which are not known, a priori, to be from the same date, is generally proscribed.

Supplementary material related to this article is available online at:
http://www.nonlin-processes-geophys.net/19/345/2012/
npg-19-345-2012-supplement.zip.

References

Bronk Ramsey, C.: Development of the radiocarbon calibration program, Radiocarbon, 43, 355–363, program available at: http://c14.arch.ox.ac.uk/oxcal.html (last access: 11 June 2012), 2001.

Bronk Ramsey, C.: Bayesian analysis of radiocarbon dates, Radiocarbon, 51, 337–360, 2009.

Buck, C. E., Christen, J. A., and James, G. N.: BCal: an online Bayesian radiocarbon calibration tool, Internet Archaeology, 7, program available at: http://bcal.shef.ac.uk/ (last access: 11 June 2012), 1999.

Fairbanks, R. G., Mortlock, R. A., Chiu, T.-C., Cao, L., Kaplan, A., Guilderson, T. P., Fairbanks, T. W., and Bloom, A. L.: Marine radiocarbon calibration curve spanning 0 to 50,000 years BP based on paired ^{230}Th/^{234}U/^{238}U and ^{14}C dates on pristine corals, Quaternary Sci. Rev., 24, 1781–1796, program available at: http://radiocarbon.ldeo.columbia.edu/research/radcarbcal.htm (last access: 11 June 2012), 2005.

Reimer, P. J., Baillie, M. G. L., Bard, E., Bayliss, A., Beck, J. W., Bertrand, C. J. H., Blackwell, P. G., Buck, C. E., Burr, G. S., Cutler, K. B., Damon, P. E., Edwards, R. L., Fairbanks, R. G., Friedrich, M., Guilderson, T. P., Hogg, A. G., Hughen, K. A., Kromer, B., McCormac, G., Manning, S., Bronk Ramsey, C., Reimer, R. W., Remmele, S., Southon, J. R., Stuiver, M., Talamo, S., Taylor, F. W., van der Plicht, J., and Weyhenmeyer, C. E.: IntCal04 terrestrial radiocarbon age calibration, Radiocarbon, 46, 1029–1058, 2004.

Stuiver, M. and Polach, H. A.: Discussion: reporting of ^{14}C data, Radiocarbon, 19, 355–363, 1977.

Stuiver, M. and Reimer, P. J.: Extended ^{14}C database and revised CALIB radiocarbon calibration program, Radiocarbon, 35, 215–230, program available at: http://calib.qub.ac.uk/calib/ (last access: 11 June 2012), 1993.

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

Ward, G. K. and Wilson, S. R.: Procedures for comparing and combining radiocarbon age determinations, Archaeometry, 20, 19–31, 1978.