DEVELOPMENTS IN THE CALIBRATION AND MODELING OF RADIOCARBON DATES

Christopher Bronk Ramsey[1] • Michael Dee[1] • Sharen Lee[1] • Takeshi Nakagawa[2] • Richard A Staff[1]

ABSTRACT. Calibration is a core element of radiocarbon dating and is undergoing rapid development on a number of different fronts. This is most obvious in the area of ^{14}C archives suitable for calibration purposes, which are now demonstrating much greater coherence over the earlier age range of the technique. Of particular significance to this end is the development of purely terrestrial archives such as those from the Lake Suigetsu sedimentary profile and Kauri tree rings from New Zealand, in addition to the groundwater records from speleothems. Equally important, however, is the development of statistical tools that can be used with, and help develop, such calibration data. In the context of sedimentary deposition, age-depth modeling provides a very useful way to analyze series of measurements from cores, with or without the presence of additional varve information. New methods are under development, making use of model averaging, that generate more robust age models. In addition, all calibration requires a coherent approach to outliers, for both single samples and where entire data sets might be offset relative to the calibration curve. This paper looks at current developments in these areas.

INTRODUCTION

Calibration is a fundamental stage in the radiocarbon dating process. At present, the main focus in ^{14}C calibration studies is, rightly, the collection of new high-quality data sets such as those from the purely terrestrial archives of the Lake Suigetsu (central Japan) sedimentary profile (Suigetsu 2006 Project, http://www.suigetsu.org/) and Kauri tree rings from New Zealand suitable for integration into subsequent versions of the consensus calibration curve (Reimer et al. 2009). For the periods and regions where we have good dendrochronologically dated wood samples, the result of this exercise is a calibration curve (Reimer et al. 2004) that can be used both for the calibration of single samples, as well as allowing high-precision modeling of larger data sets, which can now achieve dating precisions on the subcentennial level (see e.g. Bayliss et al. 2007; Bayliss 2009; Bronk Ramsey 2009a). Dating precision at the decadal level is even possible in the case of wiggle-matching (Christen and Litton 1995; Bronk Ramsey et al. 2001; Galimberti et al. 2004) and is widely used for the comparison of calibration data sets. These levels of precision rely on the close attention to detail in the samples chosen for dating (Bayliss 2009), on the accuracy of the measurements themselves, and on the accuracy and applicability of the calibration curve applied.

In many cases, however, we cannot be so confident about either the information relating to our samples, or in the calibration curve itself, particularly so in those time periods for which we do not yet have a comprehensive atmospheric calibration data set. There are a number of different contributing reasons for this and the types of uncertainty involved have different origins:

- *Uncertainty in reservoir ^{14}C concentration*—such uncertainty might be either systematic (e.g. in the case of Southern Hemisphere calibration for time periods where there is no independent Southern Hemisphere curve, see McCormac et al. 2004) or sporadic (e.g. in the case of short-term atmospheric fluctuations that may not show up in a marine-derived calibration curve). Such uncertainties will result in samples exhibiting ^{14}C measurement offsets, either systematic or individual, relative to the calibration curve.

[1]Research Laboratory for Archaeology and the History of Art, Dyson Perrins Building, South Parks Road, Oxford OX1 3QY, United Kingdom.
[2]Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne, United Kingdom.

Proceedings of the 20th International Radiocarbon Conference, edited by A J T Jull
RADIOCARBON, Vol 52, Nr 2–3, 2010, p 953–961

- *Problems of contamination*—these are usually sample-specific and therefore show up as individual samples with offset ^{14}C measurement values, though it is also possible to have consistent offsets or biases in certain circumstances.
- *Measurement problems and biases*—these again show up as aberrant ^{14}C measurements, even though the reasons are different. Again, these can either affect individual samples randomly or result in a systematic bias, though the latter should be better constrained by laboratory (and interlaboratory) quality assurance.
- Finally, we come to *uncertainties in the chronological models applied*. These can be sample-specific, as in the case where the ^{14}C measurement is correct and the reservoir is well known but where the sample might be either too old for its context (residual) or too young (intrusive). Every effort should be made to eliminate such sample-specific problems, but this will always remain a possibility that has to be considered. We can also have more generic uncertainties in the model parameters (e.g. in the case of deposition models where more parameters need to be specified).

Ideally, we need to deal with all of these uncertainties in our calibration and analysis, as we already deal with the simple measurement uncertainties quoted with any ^{14}C date. It is this issue of calibration in an imperfect world, where we have noisy data, which this paper seeks to discuss.

DEALING WITH NOISY DATA

There are a number of different ways of dealing with noisy data. The simplest is to consider the possible problems individually, and test the robustness of the calibration and model output to any such issues. One can examine, for example, the effect of removing individual sample measurements from a model, or offsetting all of the ^{14}C measurements by a small amount to see how this affects the conclusions drawn. Often, there are particular samples that appear aberrant, either by eye or through diagnostic statistics such as the Agreement index of OxCal (Bronk Ramsey 1995), and these can be targeted for testing accordingly.

The alternative approach to trialling many different models individually is to apply model averaging to achieve this in a more systematic manner. Here, we will consider a few fairly simple forms of model averaging, all of which can be expressed by adding variable parameters to the model. The mathematical details of these approaches, and how to implement them in OxCal, are covered in Bronk Ramsey 2009b and the same model names (*d-type*, *r-type*, *s-type* and *t-type*) defined therein are also given here for reference.

Systematic Radiocarbon Offsets (*d-type*)

The most obvious example of how a model averaging approach might work involves the introduction of a parameter to describe the systematic offset between a set of ^{14}C measurements and the calibration curve. We normally expect this parameter to be close to zero but can allow some scope for variability to cope with either systematic measurement biases or consistent local variation from the calibration curve due to geographical location (see e.g. Imamura et al. 2007), reservoir mixing, or growing season effects (Kromer et al. 2001). Such an approach is already used for local marine reservoir effects in the form of the ΔR value (Jones and Nicholls 2001). For atmospheric calibration, the same approach can be used but, for northern latitudes, with an expected mean of zero. For example, a ΔR of 0 ± 10 would allow for small measurement biases between the calibration curve and a specific set of samples. For the Southern Hemisphere, the same approach can be used to allow the Northern Hemisphere curve to be used to investigate the ^{14}C offset in the Southern Hemisphere (Hogg et al. 2009). This might well be a useful method for characterizing and investigating many

https://doi.org/10.1017/S0033822200046063 Published online by Cambridge University Press

primary calibration data sets. We can either see this as a single model with a new parameter, or an average over models associated with different offsets, with a prior weighting given by the ΔR value.

This approach actually deals with 2 different types of problems that are mathematically indistinguishable: measurement bias (in which the measurements are systematically offset from their true value) and curve offset (in which the true ^{14}C concentration of the reservoir involved is offset from the calibration data set employed). Figure 1 shows schematically how a whole set of measurements are allowed to shift together in ^{14}C measurement value.

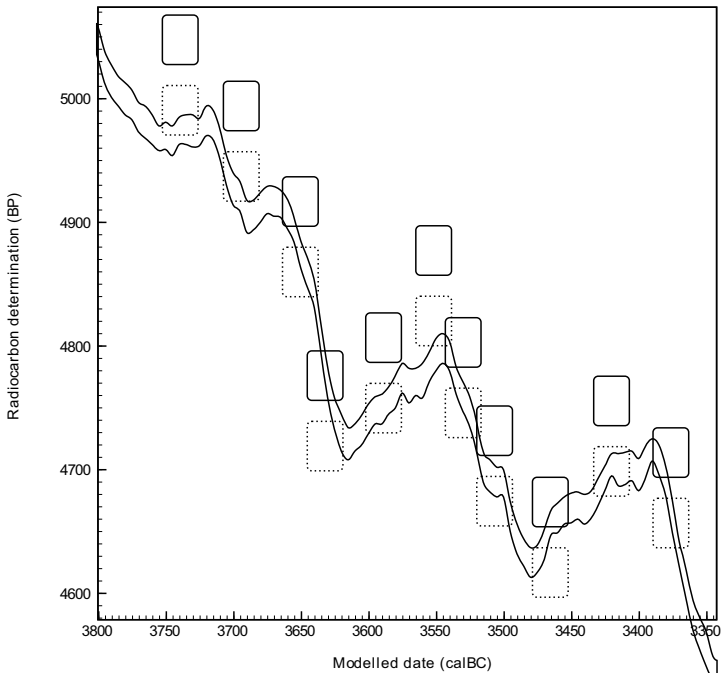

Figure 1 Schematic showing how the *d-type* offset is treated. A series of ^{14}C dates, which might, for example, be from a tree-ring sequence, are shown as solid rectangles, where the vertical extent of the rectangle shows ±1 s.d. in the measurement. A calendrical fit for the series might be found by allowing for an offset relative to the calibration curve. This can either be considered as a systematic offset in the measurements (as shown here in the dotted series of rectangles) or as an offset in the calibration curve. The horizontal extent of the rectangles show the 68% range for the fit to the calendar timescale.

Individual Radiocarbon Offsets (*r-type*)

In many instances, offsets will only apply to single samples. To some extent, such offsets can be built into the uncertainty for the individual samples, by adding in an additional 8-yr uncertainty for short-lived material (Stuiver et al. 1998). However, in many cases the distribution of such offsets is not normal and only applies to a small proportion of samples, not to all. In such cases, it is helpful to use a form of model averaging based on outliers. This introduces 2 new parameters for each measurement, one defining whether the sample is an outlier (with an offset) and another that defines the degree of that offset. The modeling then averages over the case where the measurement is included in the model as normal, and the case where a significant offset between the measurement and the calibration curve is allowed. To use such a model, it is necessary to specify a prior probability for such a shift to be found (typically 0.05).

https://doi.org/10.1017/S0033822200046063 Published online by Cambridge University Press

This approach allows us to deal with both the problems of measurement offsets (due to contamination of the sample or the measurement itself) and transient offsets in the calibration curve, which might only affect one of a series of samples. The latter application makes this approach very useful in comparing calibration data sets against each other. For example, when assessing terrestrial measurements from Lake Suigetsu (Kitagawa and van der Plicht 1998) against data sets from marine archives (e.g. Fairbanks et al. 2005; Hughen et al. 2006), it is important to allow some measurements to deviate significantly because the marine data will not reflect short-term, large-scale deviations in the ^{14}C record (Staff et al. 2010). Failure to take account of this would make any such comparison very strongly dependent on a small number of data points that show discrepancies between the different records.

Figure 2 shows how single ^{14}C measurements are allowed to shift under this type of outlier model. If such a shift is not essential for a fit, then the results of the analysis will average over cases where the shift is allowed and where it is not.

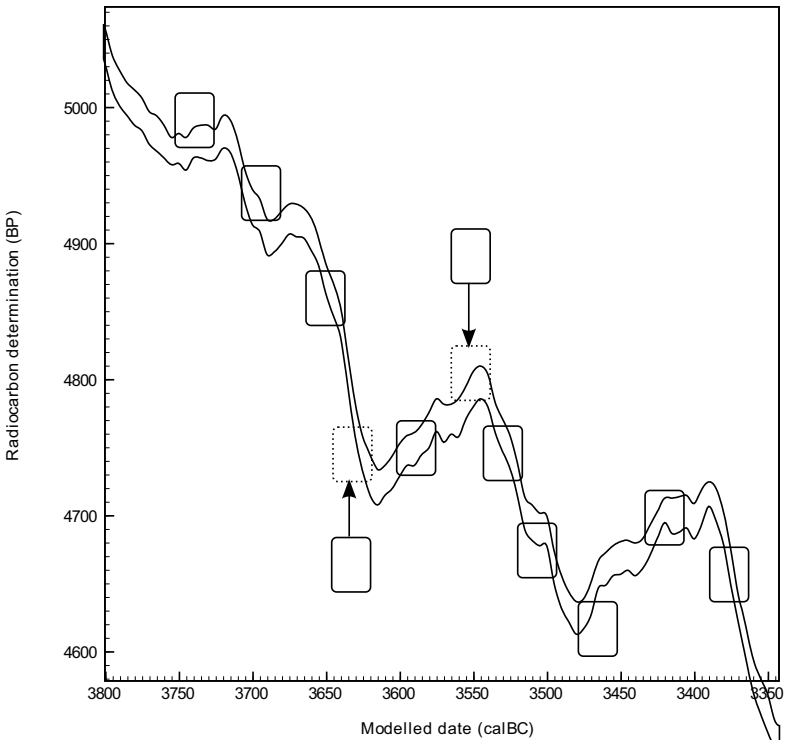

Figure 2 Schematic demonstrating the treatment of *r-type* outliers. Where a particular sample is labeled as an outlier, the ^{14}C measurement for the sample (shown as a solid rectangle) is allowed to deviate from that of the calibration curve (shown as a dotted rectangle). Such offsets are usually considered for all samples during outlier analysis, but the samples shown here with arrows are clearly more likely to be outliers than the others in the series and during model averaging would be treated as such for much more of the time.

Underestimated Measurement Uncertainty (*s-type*)

Outlier analysis was first proposed as a method for dealing with measurement problems in ^{14}C, and in particular for the very detection of outliers themselves (Christen 1994). In cases where outliers

https://doi.org/10.1017/S0033822200046063 Published online by Cambridge University Press

are due solely to the measurement process, it might be reasonable to assume that they are related to the quoted uncertainty in the measurement. In such cases, an outlier model can be used that tests the effect for each sample of increasing the uncertainty in the measurement (typically by just over 2). If the agreement with the other data is much better with such a change, then it is more likely that the sample is an outlier. This is the basis of this type of outlier detection. This approach also allows you to look at the average of model outputs for cases where such a multiplication has taken place and cases where it has not. However, in most circumstances, it is likely that any ^{14}C offsets are independent of the quoted precision of the measurement, and for model averaging purposes (as opposed to simply identifying outliers) the *r-type* offset model is likely to be more appropriate.

Figure 3 shows schematically the approach taken when this type of model is applied. Each sample is given a prior probability of being an outlier (typically 0.05). We then average over cases where the sample's error term is taken at face value and cases where it is increased. This allows us to deal with the possibility that some of the samples may have underestimated error terms.

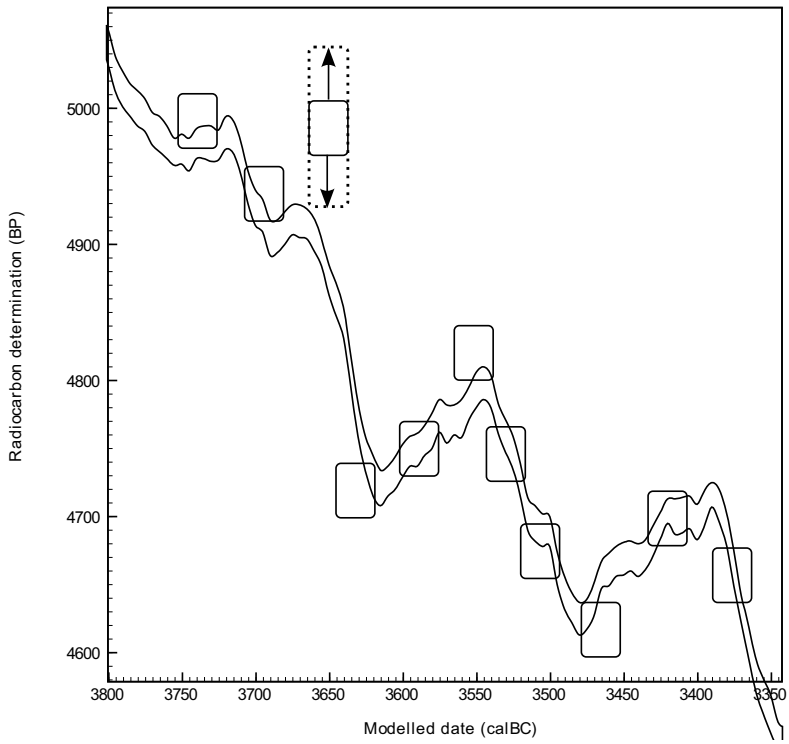

Figure 3 Schematic showing the approach taken with *s-type* outlier analysis. When treated as outliers, samples are considered with an increased error term. A sample, such as that shown here with 2 arrows, would clearly agree better with the series of dates if its error term were larger and so it can be identified as an outlier.

Offsets Due to Uncertainty in Context (*t-type*)

With some records, and especially with those most suitable for generating calibration data, there is little or no doubt about the temporal ordering of the samples (tree-ring sequences, speleothems, and laminated sediments, for example). However, in many sedimentary sequences and in most archaeological sites, there is the chance of either residuality of samples or the intrusion of younger material

https://doi.org/10.1017/S0033822200046063 Published online by Cambridge University Press

into older contexts. Such occurrences can and should be reduced by careful choice of sample material and context but can never be entirely eliminated.

To deal with such possibilities, another form of outlier analysis can be applied. In this again, we introduce 2 new parameters for each sample. The first of these determines whether or not the sample is an outlier and the second determines the temporal offset between the formation of the sample and the age of the context within which it is found. This is very similar to the *r-type* offset described above except that in this instance the offset is temporal, and the ^{14}C measurement itself is taken at face value.

As with the other methods, a prior probability of samples being outliers must be assigned (typically 0.05), and this underlines the point that much work still needs to be done in advance to ensure that the vast majority of samples are of the correct age for their context. This approach cannot be used to deal with widespread or systematic offsets through choice of poor sample types and contexts.

Figure 4 shows how this form of outlier analysis operates. The result of the model output is again an average of cases where samples are assumed to date to their context and cases where they are not. As with the other outlier methods, however, other information (likelihoods from ^{14}C measurements, and ordering information) is taken into account in determining how likely it is that the sample is inconsistent with its context and a revised estimate of this is given as the posterior outlier probability.

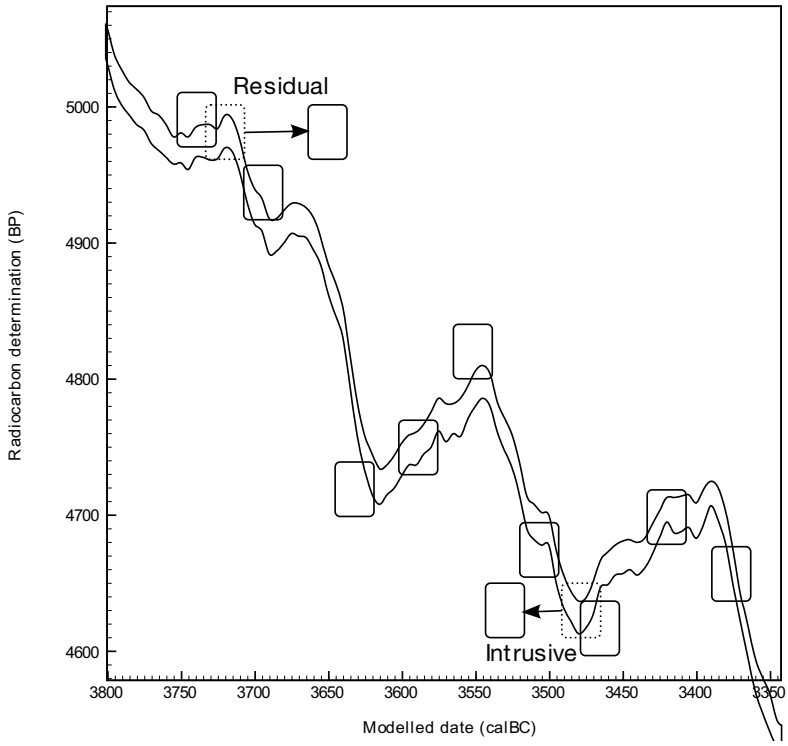

Figure 4 Schematic showing how *t-type* outliers are managed. In this model, samples that are treated as outliers are allowed to have a different calendar date for their context (shown as a solid rectangle) from those for the samples themselves (shown as dotted rectangles). Such a model would not make sense for a tree-ring sequence, but might well be appropriate in a sedimentary sequence or an archaeological site.

https://doi.org/10.1017/S0033822200046063 Published online by Cambridge University Press

Other Uses of Model Averaging

All of the cases of model averaging considered thus far in this paper are simple cases of model sets that allow either individual samples to be offset or a range of systematic biases from the calibration curve to be considered. Model averaging might also be used in a number of other ways, such as looking at the effect of different model groupings or a range of different model parameters. Here, we will consider just one further example where model averaging clearly might be very useful.

In sedimentary sequences, deposition models can be used to deal with uncertainties and fluctuations in deposition rate (see e.g. Blaauw et al. 2003; Blaauw and Christen 2005; Bronk Ramsey 2008). Similar approaches can also be used for dealing with varve-count uncertainties or those from other forms of layer counting (Bronk Ramsey 2008). In all of these models, there are some hidden or explicit parameters that determine the rigidity in the model. In the *P_Sequence* method of OxCal (Bronk Ramsey 2008) this is the *k* value, which is normally given a single value. For example, in modeling the uncertainty in varve counts for a lake record, such as that of Lake Suigetsu, using a *k* value of 0.25 means that if we have 1000 yr of varve-counted sediment and, if we knew the age at the start and the end of the segment, we would have an uncertainty of 3.2% (32 yr) in the center of the segment (see equation A17 of Bronk Ramsey 2008). A *k* value of 0.1, in the same example, would give an uncertainty of 5% (50 yr). In the case of varves, we might well be able to estimate these uncertainties, but it can be more difficult in other forms of sedimentary deposit. A useful form of model averaging might be to average over different *k* values, allowing the model to find those values most consistent with the data.

Limitations

It is important to also understand the limitations of such approaches. The use of outlier modeling, in particular, is no substitute for good quality control of samples for dating. It is also important not to use an over-complicated model for a simple situation. Where the number of samples is small and there may, for example, only be 1 outlier, it is probably better to consider the samples individually and use these methods solely for outlier detection. The outlier analysis methods become most useful with large data sets, where there are likely to be many outliers, and consideration of the implications of removing single samples becomes impossible.

The *r-type* and *t-type* outlier models are conceptually different and it will almost always make sense only to apply only one of these to a particular situation. Nonetheless, the results from using either method may be very similar, although this is less likely to be the case, for example, near plateaus in the ^{14}C curve. The *s-type* model (Christen 1994) does give significantly different results when used for model averaging with extreme outliers (Bronk Ramsey 2009b) and may be best reserved for outlier detection in such circumstances.

In all cases, it is important to think carefully about the implications of the models being implemented and, as with other developments of Bayesian statistical analysis, averaging over the inclusion and exclusion of outliers should not be seen as a black box to cover up poorly understood problems in the data.

CONCLUSIONS

Bayesian modeling is playing an increasingly important role in the calibration and interpretation of ^{14}C dates. It has also been used extensively for comparison of calibration data sets, particularly by use of wiggle-matching methods (Christen and Litton 1995; Bronk Ramsey et al. 2001).

https://doi.org/10.1017/S0033822200046063 Published online by Cambridge University Press

An important development of such modeling practices over the next few years is likely to be the use of model averaging approaches to make the procedure more robust and to provide a more representative summary of the different possible interpretations of the data available, in light of the context from which the ^{14}C samples originate.

ACKNOWLEDGMENTS

We would like to acknowledge all of the researchers who have contributed to this field of research. In particular, the work of Martin Jones and Geoff Nicholls who first presented the correct treatment of ΔR corrections and Andres Christen who was the first to look at the Bayesian treatment of outliers in ^{14}C dates (Christen 1994).

The research behind this paper was conducted in support of a number of research projects including the NERC funded Suigetsu 2006 Project (NE/F004400/1), the NERC funded consortium project on the "Response of Humans to Abrupt Environmental Transitions" (NE/E015670/1) and the Leverhulme-funded project on "Synchronising Absolute Scientific Dating and the Egyptian Historical Chronology" (F/08 662/A). It also relies on work to develop OxCal funded by English Heritage (3164 MAIN).

REFERENCES

Bayliss A, Bronk Ramsey C, van der Plicht J, Whittle A. 2007. Bradshaw and Bayes: towards a timetable for the Neolithic. *Cambridge Archaeological Journal* 17(S1):1–28.

Bayliss A. 2009. Rolling out revolution: using radiocarbon dating in archaeology. *Radiocarbon* 51(1):123–47.

Blaauw M, Heuvelink GBM, 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(14):1485–500.

Blaauw M, Christen JA. 2005. Radiocarbon peat chronologies and environmental change. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 54(4):805–16.

Bronk Ramsey C. 1995. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. *Radiocarbon* 37(2):425–30.

Bronk Ramsey C. 2008. Deposition models for chronological records. *Quaternary Science Reviews* 27(1–2): 42–60.

Bronk Ramsey C. 2009a. Bayesian analysis of radiocarbon dates. *Radiocarbon* 51(1):337–60.

Bronk Ramsey C. 2009b. Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon* 51(3):1023–45.

Bronk Ramsey C, van der Plicht J, Weninger B. 2001. 'Wiggle matching' radiocarbon dates. *Radiocarbon* 43(2A):381–9.

Christen JA. 1994. Summarizing a set of radiocarbon determinations: a robust approach. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 43(3): 489–503.

Christen JA, Litton CD. 1995. A Bayesian approach to wiggle-matching. *Journal of Archaeological Science* 22(6):719–25.

Fairbanks RG, Mortlock RA, Chiu T-C, Cao L, Kaplan A, Guilderson TP, Fairbanks TW, Bloom AL, Grootes PM, Nadeau M-J. 2005. Marine radiocarbon calibration curve spanning 0 to 50,000 years B.P. based on paired $^{230}Th/^{234}U/^{238}U$ and ^{14}C dates on pristine corals. *Quaternary Science Reviews* 24(16–17):1781–96.

Galimberti M, Bronk Ramsey C, Manning SW. 2004. Wiggle-match dating of tree-ring sequences. *Radiocarbon* 46(2):917–24.

Hogg A, Bronk Ramsey C, Turney CSM, Palmer J. 2009. Bayesian evaluation of the Southern Hemisphere radiocarbon offset during the Holocene. *Radiocarbon* 51(4):1165–76.

Hughen K, Southon J, Lehman S, Bertrand C, Turnbull J. 2006. Marine-derived ^{14}C calibration and activity record for the past 50,000 years updated from the cariaco basin. *Quaternary Science Reviews* 25(23–24): 3216–227.

Imamura M, Ozaki H, Mitsutani T, Niu E, Itoh S. 2007. Radiocarbon wiggle-matching of Japanese historical materials with a possible systematic age offset. *Radiocarbon* 49(2):331–7.

Jones M, Nicholls G. 2001. Reservoir offset models for radiocarbon calibration. *Radiocarbon* 43(1):119–24.

Kitagawa H, van der Plicht J. 1998. Atmospheric radiocarbon calibration to 45,000 yr BP: Late Glacial fluctuations and cosmogenic isotope production. *Science* 279(5354):1187–90.

Kromer B, Manning SW, Kuniholm PI, Newton MW, Spurk M, Levin I. 2001. Regional $^{14}CO_2$ offsets in the

troposphere: magnitude, mechanisms, and consequences. *Science* 294(5551):2529–32.

McCormac FG, Hogg AG, Blackwell PG, Buck CE, Higham TFG, Reimer PJ. 2004. SHCal04 Southern Hemisphere calibration, 0–11.0 cal kyr BP. *Radiocarbon* 46(3):1087–92.

Reimer PJ, Baillie MGL, Bard E, Bayliss A, Beck JW, Bertrand CJH, Blackwell PG, Buck CE, Burr GS, Cutler KB, Damon PE, Edwards RL, Fairbanks RG, Friedrich M, Guilderson TP, Hogg AG, Hughen KA, Kromer B, McCormac G, Manning S, Bronk Ramsey C, Reimer RW, Remmele S, Southon JR, Stuiver M, Talamo S, Taylor FW, van der Plicht J, Weyhenmeyer CE. 2004. IntCal04 terrestrial radiocarbon age calibration, 0–26 cal kyr BP. *Radiocarbon* 46(3):1029–58.

Reimer PJ, Baillie MGL, Bard E, Bayliss A, Beck JW, Blackwell PG, Bronk Ramsey C, Buck CE, Burr GS, Edwards RL, Friedrich M, Grootes PM, Guilderson TP, Hajdas I, Heaton TJ, Hogg AG, Hughen KA, Kaiser KF, Kromer B, McCormac FG, Manning SW, Reimer RW, Richards DA, Southon JR, Talamo S, Turney CSM, van der Plicht J, Weyhenmeyer CE. 2009. IntCal09 and Marine09 radiocarbon age calibration curves, 0–50,000 years cal BP. *Radiocarbon* 51(4): 1111–50.

Staff RA, Bronk Ramsey C, Nakagawa T, Suigetsu 2006 Project Members. 2010. A re-analysis of the Lake Suigetsu terrestrial radiocarbon calibration dataset. *Nuclear Instruments and Methods in Physics Research B* 268(7–8):960–5.

Stuiver M, Reimer PJ, Bard E, Beck JW, Burr GS, Hughen KA, Kromer B, McCormac G, van der Plicht J, Spurk M. 1998. INTCAL98 radiocarbon age calibration, 24,000–0 cal BP. *Radiocarbon* 40(3):1041–83.