Radiocarbon (2025), **67**, pp. 646–655
doi:10.1017/RDC.2025.9

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Investigating classical χ^2 test methods for high-precision calibration of radiocarbon measurements

Andrea Scifo[1] and Michael Dee[1]

[1]University of Groningen, Centre for Isotope Research, Nijenborgh 6, 9747 AG Groningen, Netherlands
Corresponding author: Andrea Scifo; Email: a.scifo@rug.nl

Received: 22 May 2024; **Revised:** 19 November 2024; **Accepted:** 20 November 2024; **First published online:** 27 March 2025

Keywords: age estimations; Bayesian analysis; calibration; radiocarbon dating

Abstract

Sudden annual rises in radiocarbon concentration have proven to be valuable assets for achieving exact-year calibration of radiocarbon measurements. These extremely precise calibrations have usually been obtained through the use of classical χ^2 tests in conjunction with a local calibration curve of single-year resolution encompassing a rapid change in radiocarbon levels. As the latest Northern Hemisphere calibration curve, IntCal20, exhibits single-year resolution over the last 5000 years, in this study we investigate the possibility of performing calibration of radiocarbon dates using the classical χ^2 test and achieving high-precision dating more extensively, examining scenarios without the aid of such abrupt changes in radiocarbon concentration. In order to perform a broad analysis, we simulated 171 sets of radiocarbon measurements over the last two millennia, with different set lengths and sample spacings, and tested the effectiveness of the χ^2 test compared to the most commonly used Bayesian wiggle-matching technique for temporally ordered sequences of samples such as tree-rings sequences, the OxCal D_Sequence. The D_Sequence always produces a date range, albeit in certain cases very narrow; the χ^2 test proves to be a viable alternative to Bayesian wiggle-matching, as it achieves calibrations of comparable precision, providing also a highest-likelihood estimate within the uncertainty range.

Introduction

The process of calibrating radiocarbon measurements has undergone several developments over the history of the dating method. It was introduced when the assumption that the atmospheric $^{14}C/^{12}C$ ratio had been historically invariant was proven to be incorrect. Several discrepancies between tree rings of known age and dates determined by radiocarbon arose in the first years of the method, highlighting the need for a calibration process (de Vries 1958; Willis et al. 1960). A correlation between solar activity and the ^{14}C concentration of such tree rings was observed soon after (Stuiver 1961, 1965; Suess 1965). Using long continuous dendrochronological sequences, a calibration curve that expressed radiocarbon dates as a function of calendar time was devised by Minze Stuiver and Hans Suess in 1966 (Stuiver and Suess 1966) and was soon expanded considerably with further measurements on other wood species from newly established chronologies (Ferguson 1968; Suess 1967, 1970).

The first approach to radiocarbon calibration was the so-called intercept method. This simple geometric approach began the convention of plotting the radiocarbon measurement on the vertical axis and absolute time on the horizontal axis. The calibration of a new measurement was determined only by the intercept between the stated limits of probability of the measurement and the calibration curve. The projection of said intercepts onto the horizontal axis gave the calibrated range for the absolute date (Ralph and Klein 1979). This rudimentary method granted only centennial-scale precision, because contemporary measurement uncertainties were broad, and the shape of the calibration curve unrefined. More importantly, it did not even take into consideration the fact that the probability ranges of

radiocarbon measurements are Normally distributed (Telford et al. 2004). In order to solve this problem, computer programs which implemented numerical and probabilistic methods to perform the necessary calculations began to become the norm (Stuiver and Reimer 1986; van der Plicht and Mook 1989). This step, by taking into account the probability density function (PDF) associated with the measurement and searching for the values with the best agreement between the PDF and the calibration curve, considerably improved the precision of the calibrated probability ranges. Soon afterwards, an approach was developed for calibrating sets of samples with a known order or separation in time, such as stratigraphic sequences or tree rings. This process, a simple form of pattern matching, became known as wiggle-matching and produced the first radiocarbon results of sub-centennial resolution (van der Plicht and McCormac 1995). Around the same time, with the introduction of modelling based on Bayesian statistics, a greater variety of calibrations were able to achieve decadal resolution. Nowadays, the consensus of the radiocarbon community is to use probabilistic methods for all radiocarbon calibration, and Bayesian modelling for sequences of dates with additional temporal information. Some of the most common software packages used within the community to perform radiocarbon calibration are OxCal (Bronk Ramsey 1995, 2001, 2017), BCal (Buck et al. 1999), CLAM (Blaauw 2010); Bacon (Blaauw and Christen 2011); and CALIB (Stuiver and Reimer 1986, 1993). For a more complete and comprehensive history of radiocarbon calibration, see Reimer (2022).

The discovery of rapid increases in atmospheric radiocarbon concentration (Miyake et al. 2012, 2013), colloquially known as Miyake Events, but more formally as radiocarbon production events, has the potential to revolutionize radiocarbon dating, as these anomalies can be used as fixed time-anchors for exact-year dating (Dee and Pope 2016). Indeed, they have already been used to pinpoint the dating of both environmental and archaeological sites (Büntgen et al. 2017; Kuitems et al. 2020, 2022; Oppenheimer et al. 2017; Wacker et al. 2014). These annual matches all utilized the Classical χ^2 test method. The process involved comparing the sample data with self-made high-resolution calibration curves constructed on local dendrochronologically dated tree-rings. The discovery of these sudden radiocarbon increases also implied that the 5-to-10-year resolution of many previous international calibration records was now something of a limitation, as the use of such blocks of tree rings may have allowed important features to be averaged out and hidden. It was thus immediately evident that enhancing the temporal resolution of calibration curves would prove beneficial for calibration purposes. The newest version of the curve (IntCal20, Reimer et al. (2020)) comprises annual resolution over the last 5000 years. Given these exciting new prospects, it is informative to examine whether high-precision pattern matching is only possible with such abrupt change points, like a radiocarbon production event and a bespoke calibration curve, or if it might be achievable more commonly.

In this study, we primarily investigate the potential of the Classical χ^2 as a method for high-precision radiocarbon dating, where a sequence of measurements of known time separation is available. We do so utilizing IntCal20 as the reference curve, and simulating a multitude of datasets via the OxCal software, mimicking results obtained on dendrochronologically ordered samples. We also compare the results of the calibrations with outcomes achieved using the D_Sequence (or wiggle-match) function in OxCal. As mentioned above, wiggle-matching in Bayesian software like OxCal is a common method used by the radiocarbon community (Bronk Ramsey 2001). The results of these different approaches for high-precision calibration will be compared with the latest calendar date of each simulated set, in order to assess their accuracy and consistency.

Methods

In this study, we simulated 171 different radiocarbon datasets for each century from the 1st century to the 19th century of the Current Era (CE), with different combinations of dataset span and sampling frequency. Each simulated dataset mimicked a possible set of radiocarbon measurements obtained on a series of temporally ordered organic samples, where the most common case would be a tree-ring sequence. During the design phase of the study, it was expected that the results would depend both on

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

Table 1. *Span, spacing and number of simulated dates for each simulation type*

Name	Code	Span (calendar years)	Spacing (calendar years between samples)	No. dates in sequence
Short annual	Sa	10	1	10
Short biennial	Sb	10	2	5
Short quintennial	Sq	10	5	2
Medium annual	Ma	25	1	25
Medium biennial	Mb	25	2	12
Medium quintennial	Mq	25	5	5
Long annual	La	50	1	50
Long biennial	Lb	50	2	25
Long quintennial	Lq	50	5	10

the distribution of data in the sample set, and on the shape of the calibration curve over the relevant period. Therefore, in order to cover as many realistic scenarios as possible, we decided to simulate, for every century analyzed, 3 different spans and 3 different internal spacings for the synthetic datasets. The 9 combinations of chosen spans and spacings are summarized in Table 1. Each combination is nicknamed according to its span (short, medium, long [S-M-L]) and its sample spacing (annual, biannual, quintennial [A-B-Q]). The number of simulated radiocarbon dates per dataset ranges from 2 to 50.

The whole simulation and analysis have been developed as a Python script. The script first randomly generates a set of calendar dates for each of the 9 simulations in each century, according to the respective parameters shown in Table 1. These dates will function as the "true dates" of the hypothetical sequence of samples; such dates, in a real-world scenario, would be unknown and therefore the goal of the calibration of the radiocarbon results. In this case study, the "true dates" form the basis from which radiocarbon dates are then simulated, as well as a final test of the accuracy of the methods we have investigated. We also set a fixed value for the assumed uncertainty associated with each radiocarbon measurement within this simulation, namely 20 ^{14}C years. The simulated radiocarbon dates were generated using an R script, run within the Python environment through the use of the rpy2 package, with dependencies on the R library oxcAAR (Hinz et al. 2018), a set of tools that enables a local installation of the OxCal software to be used remotely from within R. By feeding the sets of calendar dates and the assumed radiocarbon uncertainty into the function oxcalSimulate in the oxcAAR library (Hinz et al. 2018), the equivalent of R_Simulate in OxCal, a corresponding set of radiocarbon dates (in ^{14}C yr BP) is returned that would typically be expected for samples of those calendar years. Finally, all the simulated sets of radiocarbon measurements were exported and then calibrated against IntCal20 using two different methods: Bayesian wiggle-matching (in OxCal), and the Classical χ^2 test.

Given the following notations:

θ = assumed calendar year of latest sample in the sequence

r_i = number of spacings to the -ith sample (starting with 0 as the most recent)

$R_i \pm \delta R_i$ = radiocarbon measurements for the sample

$C(\theta - r_i) \pm \delta C(\theta - r_i)$ = radiocarbon concentrations from IntCal20 for the year $(\theta - r_i)$

The technique of Bayesian wiggle-matching is based on Bayes theorem, which states that

$$p(\theta|R_1, ..., R_n) \propto p(R_1, ..., R_n|\theta)\pi(\theta) \quad (1)$$

Where $\pi(\theta)$ is the prior on the calendar date, and is most frequently assumed to be non-informative. Then, given the calibration curve, we can calculate that (Bronk Ramsey 2009)

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

$$p(R_1, \ldots, R_n|\theta) = \prod_{i=1}^{n} \phi(R_i, C(\theta - r_i), \delta R_i^2 + \delta C^2(\theta - r_i))$$

$$\propto \left(\prod_{i=1}^{n} \frac{1}{\sqrt{\delta R_i^2 + \delta C^2(\theta - r_i)}} \right) \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{(R_i - C(\theta - r_i))^2}{\delta R_i^2 + \delta C^2(\theta - r_i)} \right) \quad (2)$$

In OxCal, Bayesian wiggle-matching involves using a D_Sequence, an OxCal function that requires the temporal order and exact calendar year spacing between each sample. In our project, we have developed a Python function to automatically generate the OxCal code (in OxCal's own programming language, Chronological Query Language, CQL2) for defining an appropriate D_Sequence tailored to the specific dataset under analysis. Using the rpy2 package, we send this OxCal code to an R environment, where it's executed via the executeOxcalScript function of the oxcAAR library. This runs the OxCal code through a local installation of the OxCal software. Subsequently, the raw output is returned to the Python environment. We extract the relevant information for our study, specifically the 95.4% posterior range of the calibrated date of the latest sample in the sequence (i.e. generally the timber felling date) of each set of samples.

The Classical χ^2 test has already been used within the radiocarbon community to confirm the exact-year dating of buildings such as the chapel of St. Mustair (Wacker et al. 2014), to pinpoint the dating of environmental events (Büntgen et al. 2017; Oppenheimer et al. 2017), to date structures of unknown age like the monumental site of Por Bajin (Kuitems et al. 2020), and most recently the first European presence in the Americas by the Norse (Kuitems et al. 2022). In all these cases, the χ^2 test was performed using local sets of high-precision measurements on annual tree-rings over the period of the well attested rapid increases in atmospheric radiocarbon concentration in 775 CE and 993 CE as reference curves. Here, we take the simulated radiocarbon dates (sample) and use portions of the IntCal20 curve (reference), and investigate if an exact-year match is ever possible, even without the aid of such a distinct increase and, more generally, how this χ^2 test performs, compared to the dating precision returned by Bayesian wiggle-matching. The test involves matching samples to the calibration curve in such a way that the χ^2 function assumes the minimal value for an assumed age of the latest sample in the sequence (typically the bark edge), where the χ^2 function is defined as follows:

$$\chi^2_{(\theta)} = \sum_{i-1}^{n} \frac{(R_i - C(\theta - r_i))^2}{\delta R_i^2 + \delta C^2(\theta - r_i)} \quad (3)$$

Note that the aforementioned formula was reported in the original publication with an incorrect sign in the numerator (Wacker et al. 2014). In our study, the χ^2 test is performed to a 2σ (95.4%) level of confidence, or alternatively a 5% level of significance.

It is noteworthy that the term in the exponential in Equation (2) is almost identical to the χ^2 test statistic, as expressed in Equation (3), with the only difference being the (–1/2), which means that for the Bayesian wiggle-matching, low χ^2 values will give higher posterior probabilities, and vice versa. The other main difference between the two equations is in the product term related to the variance, which is likely to be almost constant for any θ on the scale of a wiggle match, as the variance of IntCal20 changes relatively slowly over the length of the possible range. All these details highlight the fact that the posterior probability of any specific calendar age in the Bayesian approach is very strongly linked to the χ^2 value, because of their similar mathematical foundation, and how the results of one may resemble the other, even though OxCal relies on MCMC sampling rather than on exact calculation.

Despite this, calibration of the simulated radiocarbon dates, using these different techniques, will return different outputs in each case. The OxCal-based Bayesian wiggle-matching will return a range of possible dates for the felling date of each given sequence, determined by summing up the probabilities of each calendar year, usually defined to be where the cumulative probability adds up to 95.4%. The χ^2 test will return both a highest-likelihood estimate for the felling date, namely the calendar

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

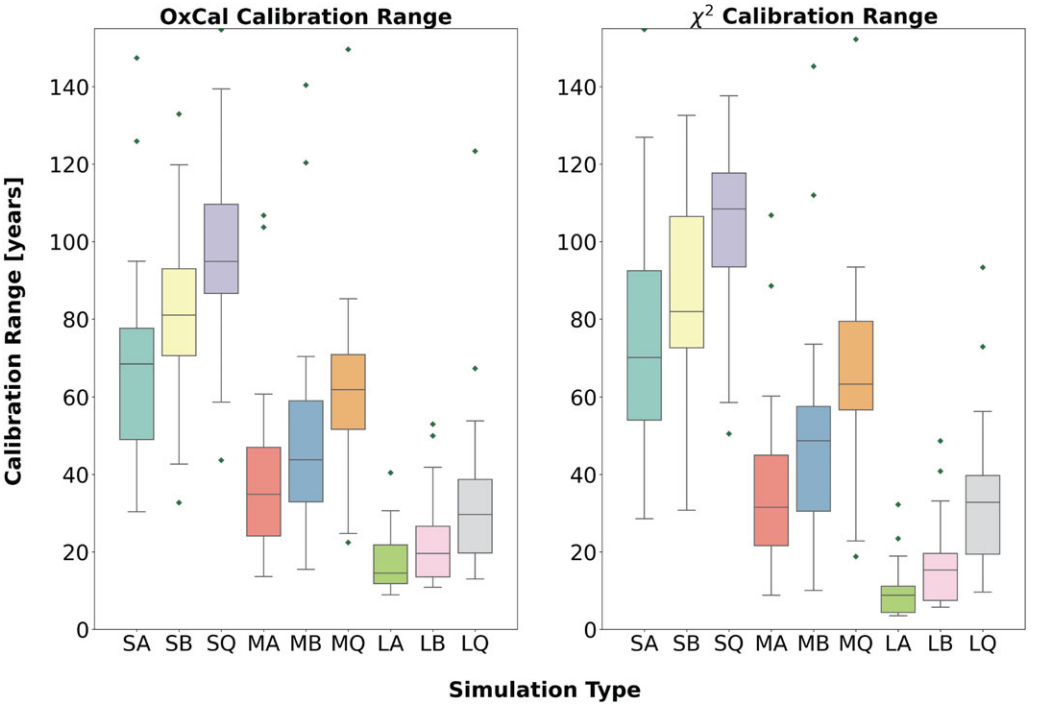

Figure 1. *Lengths of 95.4% probability calibration ranges per simulation type over the entire Common Era, from OxCal (left) and the χ^2 test (right).*

year for which the χ^2 test returns the minimal value, given that this value passes the test, and a range of dates that define the confidence interval for the highest-likelihood estimate. For this range, in order to make the comparison with OxCal more direct, we chose to perform the same sum of probabilities, rather than defining the range as a hard cutoff for those dates that pass the test. We first converted the χ^2 values into likelihood probabilities, then sorted those probabilities in descending order (with the most likely year first), and then performed the cumulative sum of the sorted probabilities until reaching 95.4% of the total probability. In a real-world scenario, users will usually have archaeological or environmental evidence which will point to an expected date range over which to perform the test. In this simulation, in order to mimic that scenario, we use the OxCal-derived date ranges as realistic intervals for the calibration, but then we extend those by 100 years on each side in order to avoid making the χ^2 test statistically dependent on the outcome of the Bayesian wiggle-matching, and inadvertently altering the coverage probability of the test. Lastly, we will be able to evaluate the accuracy of the results of these techniques by comparing them with the "true dates" of the originally generated calendar dates sequences.

It is important to clarify that the 95.4% probability density ranges returned by OxCal calibration incorporate years of differing probability, and even gaps, within the entire 95.4% range. These variations in probability density are influenced by the shape of the calibration curve over the period in question. For simplicity's sake, OxCal output tables summarize, and researchers frequently report, the full 95.4% range from the oldest to youngest possible calendar year. Given the intricacies of the calibration process and the diverse scenarios analyzed in this study, we also chose to report the entire 95.4% probability range to encompass the full range of potential calendar ages with confidence. Similarly, when reporting ranges of dates that pass the χ^2 test, it's important to note that there may be contiguous periods in which none of the years pass at the stated level of probability, interspersed by ones in which every year does. In these cases, for our study, a single range is also reported—from the oldest to the youngest passing date—to simplify the interpretation. This approach results in broader calibration

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

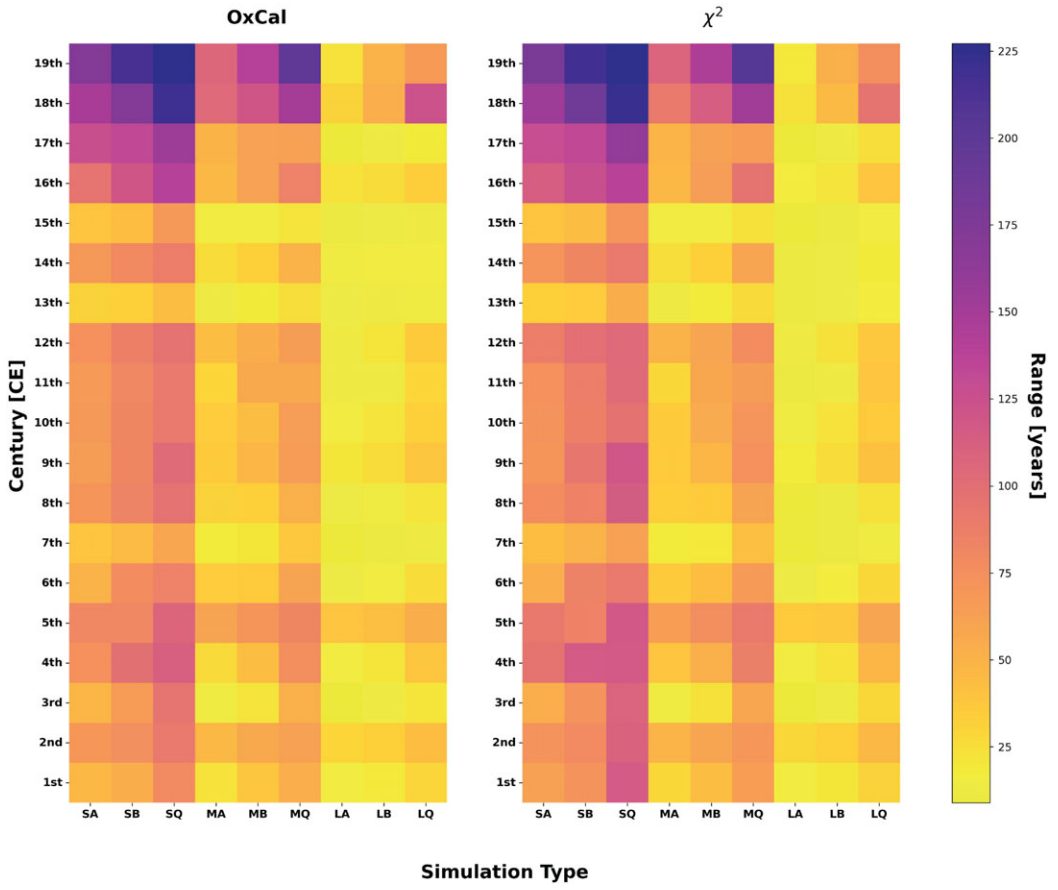

Figure 2. *Lengths of 95.4% probability calibration ranges per simulation type per each century of the Common Era, from OxCal (left) and the χ^2 test (right).*

ranges than those directly generated by the algorithm. However, it facilitates a fairer comparison between the two methods used in this study and ensures clarity in reporting of the results.

To mitigate the influence of randomness on our results, we adopt a Monte Carlo-like approach. This involves repeating the entire process, including the generation of new calendar dates, simulation of new radiocarbon dates, and then their calibration using the aforementioned methods, 40 times. Subsequently, we average the final results for each simulation scenario (refer to Table 1) to obtain comprehensive and robust findings.

All the aforementioned Python scripts are reported, free-to-use, together with the full results of each iteration, on GitHub at *github.com/scifondre/chisq_radiocarbon*. The script for performing the χ^2 test-based calibration, additionally to being part of the whole program, is also provided as a separate standalone script.

Results and discussion

As mentioned previously, Bayesian wiggle-matching in OxCal produces a PDF over calendar years, which shows the likelihood of each possible calendar age for the given radiocarbon measurement. The χ^2 test instead produces two outcomes: one is also a range of dates, corresponding to the suite of dates that pass the test, and the other is the highest-likelihood estimate, namely the single value within that given range which minimizes the χ^2 value; that is, the best candidate for an exact-year match. In this section, we compare the range of dates produced by wiggle-matching and χ^2 test calibration, and then

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

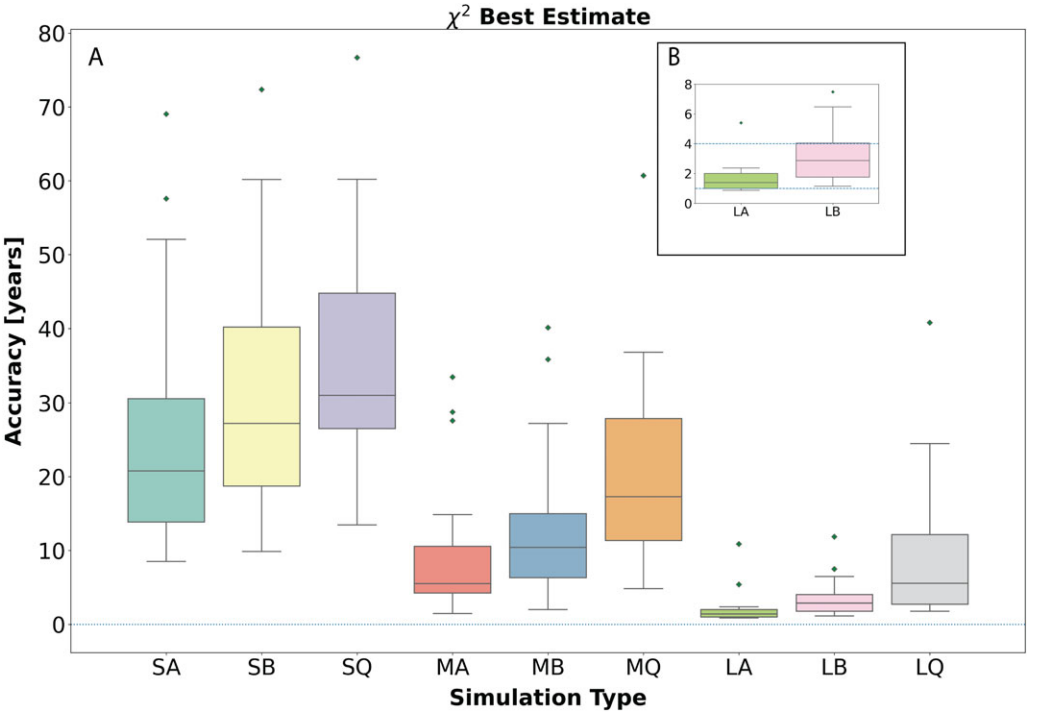

Figure 3. *Accuracy of the χ^2 test-based highest-likelihood estimate per simulation type over the entire Common Era (A). The inset (B) provides a magnified view of simulation types LA and LB.*

evaluate the accuracy of the χ^2 test highest-likelihood estimate by comparing it with the "true date" from the original simulation. For this analysis, we averaged the results of each Monte Carlo iteration. We firstly present the results of the comparison between date ranges produced by wiggle-matching and χ^2 test calibration in the form of boxplots (boxes reporting Interquartile Ranges, or middle 50%), where we summarize the results per simulation type for the entire Common Era (Figure 1).

Upon initial examination, it is noteworthy how similar the ranges are between the two methods. Looking more closely at the values of mean, standard deviation, minimum and maximum calibration range size for each dataset type, we can see that the χ^2 test performs better than OxCal in the case of the "Long" datasets, OxCal performs slightly better than the χ^2 test for the "Short" datasets, and for the "Medium" datasets the difference between the two methods is negligible.

Looking at the results for each dataset type, it is not surprising that on average the longer the simulated dataset, the shorter the calibrated date ranges. That is, the "Short" datasets return the least precise and the "Long" datasets return the most precise results. It is also not surprising that, for datasets of the same span, annual sampling leads to the more favorable results, while quintennial sampling tends to generate the least favorable results, as fewer samples are measured. Furthermore, it is highly informative to compare the results from the datasets that contain similar number of dates. For instance, datasets SA, MB and LQ contain 10, 12 and 10 radiocarbon dates respectively, making them comparable from a financial and labor point-of-view. From the outputs of this study, it is clear that dataset SA is the worst option, while LQ is likely to be the best, although somewhat comparable to MB. Therefore, where longer sample sequences are possible, a choice of biennial sampling over 25 years or quintennial sampling over 50 years is likely to produce markedly higher precision for the date being sought (tree felling) than annually sampled sets over 10 years. A similar comparison can be carried out by analyzing the datasets SB and MQ, as both comprise 5 dates, and datasets MA and LB, which contain 25 dates each. Again, for both pairs of dataset types, the choice of dataset that spans the longer period of time, with the more widely spaced sampling frequency, leads on average to the most precise results.

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

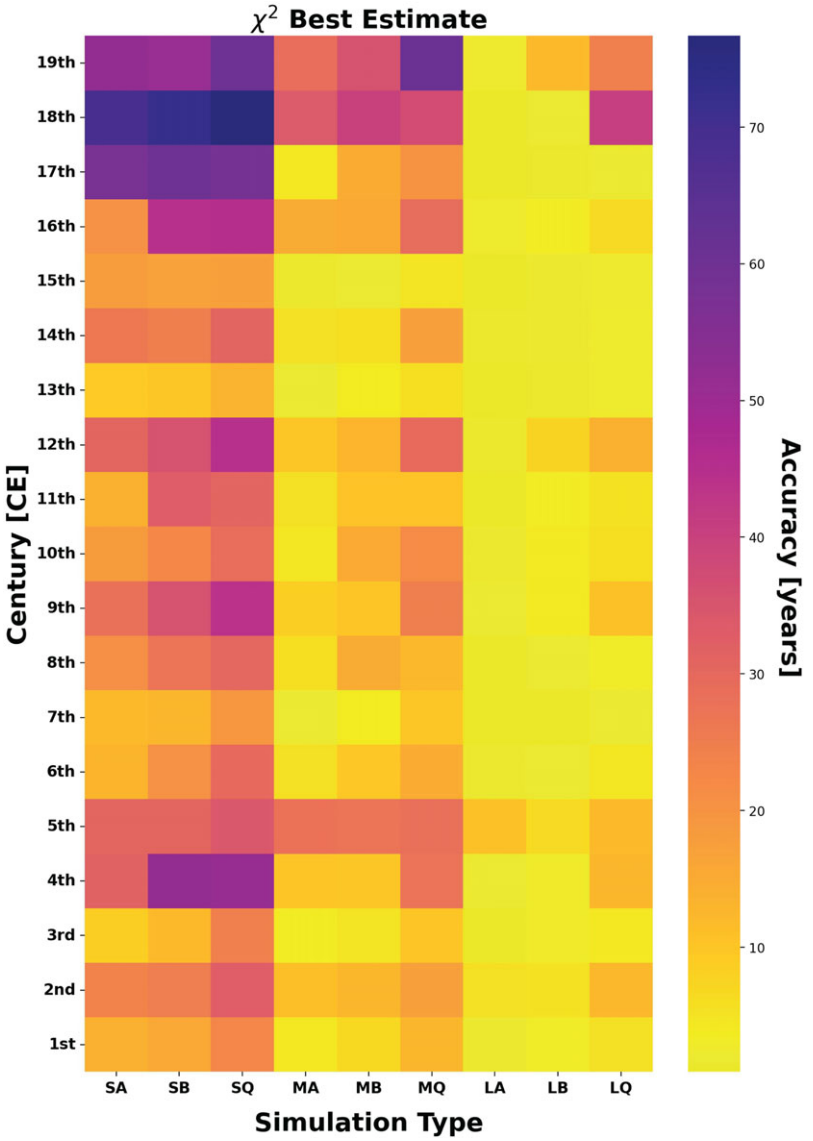

Figure 4. *Accuracy of the χ^2 test-based highest-likelihood estimate per simulation type per each century of the Common Era.*

In Figure 2 we present the same comparison as above, but in more detail as we employ a heatmap where the outputs of each simulation type are specified for each century of the Current Era by way of a color map. This heatmap works as an effective representation of calibration precision over each century of the Current Era as a function of dataset length and sampling frequency. Certain centuries conspicuously lead to a more precise calibrated ranges, regardless of the chosen dataset type; for instance, the 15th, 13th, 11th, 7th and 3rd centuries CE show better results than average.

One of the aims of this study was to investigate the levels of accuracy for the highest-likelihood estimate obtainable by the χ^2 test, without the presence of an annual radiocarbon production increase, like the 775 CE event. In order to interpret the results clearly and consistently, we define the accuracy of the approach to be the calendrical difference between the highest-likelihood estimate for the dating of the final sample in the sequence (tree felling) determined by the χ^2 test and the known "true date" for the

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

same sample. We show this result in the form of boxplots per simulation type over the entire simulation period (Figure 3).

The same pattern is observed as above; that is, between datasets with similar number of dates, the ones with the longer time span and more dispersed sampling provide the best results. Additionally, Figure 3B highlights the remarkable fact that, without any sharp radiocarbon feature, datasets LA and LB return final date estimates (minimum of the χ^2 test analysis) between 1 and 4 years of the "true date," regardless of the century, consistently over each iteration of the Monte Carlo analysis. For individual iterations of the Monte Carlo analysis, an exact-year match is often achieved in the LA category, as it is visible in Figure 3B where the lower whisker of the averaged results extends below 1 year. However, exact year matches cannot be guaranteed by this approach, as their occurrence is a function of the variability of the calibration record for the specific years in question. It is paramount to stress the fact that when reporting highest-likelihood estimate felling dates through this method, regardless of the level of accuracy reported in this study, such a date must always be given with the appropriate confidence level, as defined by the calibration range. The script provided for performing the χ^2 test-based calibration has this feature embedded.

In Figure 4, the accuracy of the χ^2 test-based highest-likelihood estimate is given again in form of a heatmap, showing the results divided per simulation type and per century. As well as observing the same pattern as Figure 2, namely that the 15th, 13th, 11th, 7th and 3rd centuries achieve the best precisions, it provides insight on some interesting cases where shorter datasets can achieve a similar degree of accuracy to longer datasets and can be used as a map of likely high-precision accuracy in the Common Era.

Conclusions

OxCal-based Bayesian wiggle-matching is a reliable technique for the calibration of radiocarbon dates. Its output is a range of calendar dates, the breath of which depends on the chosen degree of probability and the shape of the calibration curve over the time frame of interest. The Classical χ^2 test has till now been used as a tool to achieve exact-year dating when certain circumstances are met. These include the presence of a sharp increase in radiocarbon production, like the 775 CE event, very high precision radiocarbon measurements, and a high-precision bespoke calibration curve obtained from local dendrochronological samples.

In this study, we investigated the use of the Classical χ^2 test as an alternative to OxCal-based Bayesian wiggle-matching for calibrating time-ordered sequences of radiocarbon dates, without any production anomalies, at standard measurement precision (typically around ±20 years BP), and employing IntCal20. Our results show how this method determines accurate calibration ranges, and that in general it performs comparably to OxCal; in certain cases achieving slightly shorter ranges and in other slightly longer ranges, depending on the features of the simulated dataset. In general, it presents itself as a viable alternative to Bayesian wiggle-matching. Additionally, the Classical χ^2 test provides a reliable estimate for the latest sample in a sequence. The extent of the uncertainty in this estimate varies based on the length and sampling frequency of the dataset, and the timeframe being investigated. It is noteworthy to point out that, for any period of the Common Era, when calibrating a dataset spanning 50 years with annual or biennial sampling, the highest-likelihood estimate for the last date in the sequence is almost certainly within 4 years of the correct value. However, this estimate must always be reported with the broader uncertainty range given by the method.

In summary, although an exact-year match cannot be guaranteed without the aid of a sudden radiocarbon increase, we suggest that χ^2 test calibration can be considered a viable alternative to Bayesian wiggle-matching, achieving comparable results while having the benefit of being a simpler procedure, in terms of computational power required and mathematical complexity. Additionally, being a Frequentist approach relying on exact calculations rather than on sampling, it can provide a highest-likelihood estimate date that can be interpreted as the most likely true date within its uncertainty range.

Acknowledgments. This work was supported by a European Research Council grant (ECHOES, 714679).

https://doi.org/10.1017/RDC.2025.9 Published online by Cambridge University Press

References

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

Blaauw M and Christen JA (2011) Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Analysis* **6**(3), 457–474, 418.

Bronk Ramsey C (1995) Radiocarbon calibration and analysis of stratigraphy: The OxCal program. *Radiocarbon* **37**(2), 425–430.

Bronk Ramsey C (2001) Development of the radiocarbon calibration program. *Radiocarbon* **43**(2A), 355–363.

Bronk Ramsey C (2009) Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**(1), 337–360.

Bronk Ramsey C (2017) Methods for summarizing radiocarbon datasets. *Radiocarbon* **59**(6), 1809–1833.

Buck CE, Christen JA and James GN (1999) BCal: An on-line Bayesian radiocarbon calibration tool. *Internet Archaeology* **7**.

Büntgen U, Eggertsson Ó, Wacker L, Sigl M, Ljungqvist FC, Di Cosmo N, Plunkett G, Krusic PJ, Newfield TP, Esper J et al. (2017) Multi-proxy dating of Iceland's major pre-settlement Katla eruption to 822–823 CE. *Geology* **45**(9), 783–786.

de Vries H (1958) Variation in concentration of radiocarbon with time and location on Earth. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series B6L*, 94–102.

Dee MW and Pope BJS (2016) Anchoring historical sequences using a new source of astro-chronological tie-points. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **472**(2192), 20160263.

Ferguson CW (1968) Bristlecone pine: science and esthetics: A 7100-year tree-ring chronology aids scientists; old trees draw visitors to California mountains. *Science* **159**(3817), 839–846.

Hinz M, Schmid C, Knitter D and Tietze C (2018) oxcAAR: Interface to "OxCal" radiocarbon calibration.

Kuitems M, Panin A, Scifo A, Arzhantseva I, Kononov Y, Doeve P, Neocleous A and Dee M (2020) Radiocarbon-based approach capable of subannual precision resolves the origins of the site of Por-Bajin. *Proc Natl Acad Sci USA* **117**(25), 14038–14041.

Kuitems M, Wallace BL, Lindsay C, Scifo A, Doeve P, Jenkins K, Lindauer S, Erdil P, Ledger PM, Forbes V et al. (2022) Evidence for European presence in the Americas in AD 1021. *Nature* **601**(7893), 388–391.

Miyake F, Masuda K and Nakamura T (2013) Another rapid event in the carbon-14 content of tree rings. *Nat Commun* **4**:1748.

Miyake F, Nagaya K, Masuda K and Nakamura T (2012) A signature of cosmic-ray increase in AD 774–775 from tree rings in Japan. *Nature* **486**(7402), 240–242.

Oppenheimer C, Wacker L, Xu J, Galván JD, Stoffel M, Guillet S, Corona C, Sigl M, Di Cosmo N, Hajdas I et al. (2017) Multi-proxy dating the "Millennium Eruption" of Changbaishan to late 946 CE. *Quaternary Science Reviews* **158**, 164–171.

Ralph EK and Klein J (1979) Composite computer plots of 14C dates for tree-ring-dated bristlecone pines and sequoias. *Radiocarbon dating: proceedings of the Ninth International Conference Los Angeles and La Jolla, 1976*. Berkeley (CA), University of California Press, 545–553.

Reimer PJ (2022) Evolution of radiocarbon calibration. *Radiocarbon* **64**(3), 523–539.

Reimer PJ, Austin WEN, Bard E, Bayliss A, Blackwell PG, Bronk Ramsey C, Butzin M, Cheng H, Edwards RL, Friedrich M et al. (2020) The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 cal kBP). *Radiocarbon* **62**(4), 725–757. doi: 10.1017/RDC.2020.41.

Stuiver M (1961) Variations in radiocarbon concentration and sunspot activity. *Journal of Geophysical Research (1896–1977)* **66**(1), 273–276.

Stuiver M (1965) Carbon-14 content of 18th- and 19th-century wood: variations correlated with sunspot activity. *Science* **149**(3683), 533–535.

Stuiver M and Reimer PJ (1986) A computer program for radiocarbon age calibration. *Radiocarbon* **28**(2B), 1022–1030.

Stuiver M and Reimer PJ (1993) Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program. *Radiocarbon* **35**(1), 215–230.

Stuiver M and Suess HE (1966) On the relationship between radiocarbon dates and true sample ages. *Radiocarbon* **8**, 534–540.

Suess HE (1965) Secular variations of the cosmic-ray-produced carbon 14 in the atmosphere and their interpretations. *Journal of Geophysical Research (1896–1977)* **70**(23), 5937–5952.

Suess HE (1967) Bristlecone pine calibration of the radiocarbon time scale from 4100 BC to 1500 BC. *Symposium on radioactive dating and methods of low-level counting*. IAEA, 143–151.

Suess HE (1970) Bristlecone-Pine calibration of the radiocarbon time-scale 5200 BC to the present. In Olsson IU (ed), *Radiocarbon Variations and Absolute Chronology. Proceedings of the 12th Nobel Symposium*. Stockholm: Almqvist and Wiksell, 303–312.

Telford RJ, Heegaard E and Birks HJB (2004) The intercept is a poor estimate of a calibrated radiocarbon age. *The Holocene* **14**(2), 296–298.

van der Plicht J and McCormac FG (1995) A note on calibration curves. *Radiocarbon* **37**(3), 963–964.

van der Plicht J and Mook WG (1989) Calibration of radiocarbon ages by computer. *Radiocarbon* **31**(3), 805–816.

Wacker L, Güttler D, Synal H-A, Walti N, Goll J and Hurni JP (2014) Radiocarbon dating to a single year by means of rapid atmospheric ^{14}C changes. *Radiocarbon* **56**(2), 573–579.

Willis EH, Tauber H and Münnich KO (1960) Variations in the atmospheric radiocarbon concentration over the past 1300 years. *Radiocarbon* **2**, 1–4.

Cite this article: Scifo A and Dee M (2025). Investigating classical χ^2 test methods for high-precision calibration of radiocarbon measurements. *Radiocarbon* **67**, 646–655. https://doi.org/10.1017/RDC.2025.9