Quantification and propagation of errors when converting vertebrate biomineral oxygen isotope data to temperature for palaeoclimate reconstruction

Alexander J. E. Pryor[a*], Rhiannon E. Stevens[b], Tamsin C. O'Connell [a,b], John R. Lister[c]

* Corresponding author: Alex_Pryor@cantab.net. +44 7737 114165

RES - res57@cam.ac.uk

TO'C - tco21@cam.ac.uk

JRL - lister@damtp.cam.ac.uk

[a] *Department of Archaeology & Anthropology, University of Cambridge, Downing Street, Cambridge, CB2 3DZ, United Kingdom.*

[b] *McDonald Institute for Archaeological Research, University of Cambridge, Downing Street, Cambridge, CB2 3ER, United Kingdom.*

[c] *Institute of Theoretical Geophysics, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WA, United Kingdom*

Abstract

Oxygen isotope analysis of bioapatite in vertebrate remains (bones and teeth) is commonly used to address questions on palaeoclimate from the Eocene to the recent past. Researchers currently use a range of methods to calibrate their data, enabling the isotopic composition of precipitation and the air temperature to be estimated. In some situations the regression method used can significantly affect the resulting palaeoclimatic interpretations. Furthermore, to understand the uncertainties in the results, it is necessary to quantify the errors involved in calibration. Studies in which isotopic data are converted rarely address these points, and a better understanding of the calibration process is needed. This paper compares regression methods employed in recent publications to calibrate isotopic data for palaeoclimatic interpretation and determines that least-squares regression inverted to $x = (y - b)/a$ is the most appropriate method to use for calibrating causal isotopic relationships. We also identify the main sources of error introduced at each conversion stage, and investigate ways to minimise this error. We demonstrate that larger sample sizes substantially reduce the uncertainties inherent within the calibration process: typical uncertainty in temperature inferred from a single sample is at least ±4ºC, which multiple samples can reduce to ±1–2ºC. Moreover, the gain even from one to four samples is greater than the gain from any further increases. We also show that when converting $\delta^{18}O_{precipitation}$ to temperature, use of annually averaged data can give significantly less uncertainty in inferred temperatures than use of monthly rainfall data. Equations and an online spreadsheet for the quantification of errors are provided for general use, and could be extended to contexts beyond the specific application of this paper.

Palaeotemperature estimation from isotopic data can be highly informative for our understanding of past climates and their impact on humans and animals. However, for such estimates to be useful, there must be confidence in their accuracy, and this includes an assessment of calibration error. We give a series of recommendations for assessing uncertainty when making calibrations of $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–Temperature. Use of these guidelines will provide a more solid foundation for palaeoclimate inferences made from vertebrate isotopic data.

Key words (6)

phosphate, enamel, regression, calibration, temperature, paleoclimate

1. Introduction

Oxygen isotope analysis of bioapatite in vertebrate remains (bones and teeth) and shell carbonates in terrestrial and marine invertebrates are commonly used to address questions on palaeoclimate, palaeoecology and palaeotemperature from the Eocene to the recent past (e.g. FRICKE et al., 1995; LÉCOLLE, 1985; VAN DAM and REICHART, 2009; ZANAZZI et al., 2007; ZANCHETTA et al., 2005). It is sometimes possible to use $\delta^{18}O_{bioapatite}$ values to address the questions of interest directly, without requiring the data to be converted/calibrated to other forms (e.g. FORBES et al., 2010; HALLIN et al., 2012). In many isotopic studies, however, the data are converted to quantitative estimates of the oxygen isotopic value of precipitation and thence to temperature (ARPPE and KARHU, 2010; NAVARRO et al., 2004; SKRZYPEK et al., 2011; TÜTKEN et al., 2007). These investigations require two data conversions that are based on well demonstrated correlations:

Z1 A species-specific conversion, using $\delta^{18}O_{bioapatite}$ to estimate the mean isotopic composition of ingested water ($\delta^{18}O_{drinking\ water}$)(KOHN, 1996; LONGINELLI, 1984; LUZ et al., 1984; LUZ and KOLODNY, 1985). For the purposes of palaeoclimatic reconstruction $\delta^{18}O_{drinking\ water}$ is typically assumed to be equivalent to local mean $\delta^{18}O_{precipitation}$;

Z2 A regionally-specific conversion, using the estimated value of mean $\delta^{18}O_{precipitation}$ to estimate mean air temperature T (ROZANSKI et al., 1992), which relates to the period the bioapatite was growing.

These correlations exist because of physical laws that govern the movement of isotopes through the biological and hydrological systems, and they remain consistently statistically significant across geographical regions and species (DANSGAARD, 1964; LONGINELLI, 1984).

Defining accurate empirical mathematical relationships between these variables is complicated both by the problems in obtaining reliable primary data and by the effect of other variables that introduce uncertainties into the relationships themselves (KOHN and WELKER, 2005). These uncertainties originate from many parameters, comprising biological (including species effects, population variability, variability in use of different water sources), environmental (such as latitudinal effects, rain variability, isotopic variation between potential water sources) and analytical (preparation techniques and measurement uncertainty) effects.

Published equations between temperature and the oxygen isotopic values of bioapatite and precipitation (henceforth referred to as $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–T) are developed using regression analyses to obtain lines of best fit in the form $y(x) = ax + b$ (Table 1). These may be used to calibrate data if the correlation is strong enough (LUCY et al., 2008). Recent examples from the literature make clear, however, that different mathematical practices are currently employed for undertaking the regression, and we will argue that not all methods are equally appropriate.

The spread of the data about a line of best fit represents the combined effect of all the sources of uncertainty. We show that when a best-fit correlation is used to convert new isotopic measurements, this spread makes an important contribution to the resultant uncertainty, and it must be taken into account, even if the line of best fit appears well constrained. If all the uncertainties are acknowedged, then the calibrations can be a useful method for generating first-order estimates of variables of interest in palaeoclimatic research. We will demonstrate that the uncertainties in the empirically-derived isotopic relationships, and the natural variability of new samples about those relationships, lead unavoidably to significant uncertainty in estimates of $\delta^{18}O_{precipitation}$ and temperature. Moreover, the calibrations require several steps of data conversion, and the uncertainties need to be combined appropriately. Whilst some researchers give some information about uncertainties in individual correlations (Bernard et al., 2009; Grimes et al., 2003; Pollard et al., 2011; Pryor et al., 2013; Stevens et al., 2011; van Dam and Reichart, 2009;), others do not explicitly quantify the statistical uncertainties inherent in their calculations (Ukkonen et al., 2007; Iacumin et al., 2010).

Here, we explore the application of standard statistical analysis to the issue of data calibration in the context of generating estimates of past temperature across a wide span of geological time (Arppe and Karhu, 2010; Delgado Huertas et al., 1995; Fabre et al., 2011; Kovács et al., 2012; Krzemińska et al., 2010; Matson and Fox, 2010; Skrzypek et al., 2011; Tütken et al., 2007; Ukkonen et al., 2007; van Dam and Reichart, 2009). Our methods are similar to those used in Pollard et al. (2011) who outline the errors associated with inferring geographical origin from individual human bioapatite measurements We first review some of the methods commonly used for regression analyses that facilitate the conversion of $\delta^{18}O_{bioapatite}$–

$\delta^{18}O_{precipitation}$–T. A regression technique is then established that is statistically valid and appropriate for the datasets being employed, and the reasons for choosing this method are explained in detail. A method for calculating the uncertainties involved in the data calibrations is then presented, introducing the underlying mathematical model and the formulae which comprise the basis of the calculation. A digital spreadsheet that researchers may download and use to process their own data is also presented (Supplementary Data). We then use our model to demonstrate some trends that arise from error calculations and conclude with a series of recommendations concerning the handling of errors when making $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–T conversions. The primary calibration equations discussed in this paper focus on the conversion relationships developed for horse (DELGADO HUERTAS et al., 1995) and elephants (AYLIFFE et al., 1992): although based on small datasets, both are widely applied (ARPPE and KARHU, 2010; BOS et al., 2001; DELGADO HUERTAS et al., 1995; FABRE et al., 2011; KOVÁCS et al., 2012; KRZEMIŃSKA et al., 2010; MATSON and FOX, 2010; SKRZYPEK et al., 2011; TÜTKEN et al., 2007; UKKONEN et al., 2007). We use them as an example to show that correct mathematical handling of the data facilitates a more rigorous data-conversion process, and gives a clearer statement of the inherent uncertainties in the predictions being made from the existing data.

2. Data conversion on enamel carbonates

By convention, the calibration equations of interest (e.g. for Z1) are typically expressed in terms of $\delta^{18}O_{bioapatite}$ values measured on the phosphate moiety in the bioapatite structure, quoted relative to the SMOW/VSMOW isotopic standards. Enamel carbonates offer an alternative source for measuring $\delta^{18}O_{bioapatite}$, almost always measured relative to the PDB/VPDB isotopic standards. Using isotopic data measured on the carbonate moiety of tooth enamel therefore requires up to two preliminary conversions (see Table 1): firstly if the $\delta^{18}O_{bioapatite}$ values were measured relative to the PDB/VPDB isotopic standards, and/or secondly the estimation of a phosphate $\delta^{18}O$ value from an enamel carbonate $\delta^{18}O$ measurement. While these two conversions (described as A1 and A2 in Table 1) each have statistical errors associated with defining the line of best fit through the data points (see below), their correlation coefficient r^2 is very close to 1, meaning the associated errors are minimal. Similarly, measurement errors on oxygen isotopic values are typically negligible compared to the calibration errors. This paper therefore focuses on the implications of much greater uncertainties in conversions from $\delta^{18}O_{bioapatite}$ to $\delta^{18}O_{precipitation}$ and thence to temperature T (Z1 and Z2 in Table 1). Unless specifically stated, all $\delta^{18}O$ values in this paper are given relative to SMOW/VSMOW.

3. Calculating conversion relationships using least squares regression

From the perspective of palaeoclimatic investigations, the equations used for conversions Z1 and Z2 are often published in a form that is in the opposite direction to that required when investigating palaeontological and archaeological material: i.e. $y = ax + b$ where *x* is the unknown variable being reconstructed from observations of y (e.g. Table 1). This is because the conversion equations follow the presumed direction of causality, from input to output – thus, $\delta^{18}O_{ingested\ water}$ as the independent variable on the *x*-axis controls resultant $\delta^{18}O_{bioapatite}$ on the *y*-axis and, similarly, air temperature *T* controls resultant $\delta^{18}O_{precipitation}$. Palaeoclimatologists, however, need to work backwards from the known output, which is found and measured, to estimate the input. Researchers have approached this problem in two different ways: some choose to find the least-squares fit $y(x) = ax + b$ and then invert it to obtain $x = (y - b)/a$

(henceforth known as inverted forward regression)(ARPPE and KARHU, 2010; AYLIFFE et al., 1994; TÜTKEN et al., 2007; UKKONEN et al., 2007); others instead swap the *x* and *y* axes of the original data, transposing and re-plotting it, to find a new least-squares fit of the form $x = cy + d$ (henceforth referred to as transposed, or reversed, regression)(BERNARD et al., 2009; FABRE et al., 2011; KOVÁCS et al., 2012; SKRZYPEK et al., 2011; VAN DAM and REICHART, 2009;).

It is important to note that, unless the data are perfectly correlated (with r^2 = 1), the equations $x = (y - b)/a$ and $x = cy + d$ obtained in this way *from the same dataset* will differ in a predictable manner and thus generate predictably different values for '*x*'. Both equations pass through the mean $(\bar{x}, \bar{y})$ of the data, but the slopes $1/a$ and *c* are related by

$$c = r^2 / a \qquad \text{Equation 1}$$

so that the worse the data are correlated (the further r^2 is from 1), the larger the difference between the slope of the inverted forward and the transposed equations. From this relationship it follows that values of '*x*' calculated using a transposed regression fit of *x*(*y*) will be consistently higher than those produced from the inverted forward regression fit of *y*(*x*) for the range of values below the mean $(\bar{x}, \bar{y})$, and consistently lower for those above $(\bar{x}, \bar{y})$ (e.g. Figure 1A).

This discrepancy is a serious problem when attempting quantitative palaeoclimatic reconstruction from isotopic data. For example, across the range of $\delta^{18}O_{bioapatite}$ values typically measured from palaeontological and archaeological samples (c.5–25‰ relative to VSMOW), differences in predicted $\delta^{18}O_{ingested\ water}$ from the forward and transposed fits, *y*(*x*) and *x*(*y*), vary by several permil, owing to the difference in fitted slopes for typical r^2 = 0.75–0.85 (see Table 1). Similarly for temperature, where the values of r^2 are 0.6 or smaller (Table 1) and thus the difference in slopes is much larger, temperatures calculated from $\delta^{18}O_{precipitation}$ using a forward fit *y*(*x*) will always be significantly warmer than those calculated using a transposed fit *x*(*y*) for values below the mean, and the converse is true when above the mean (Figure 1A).

One recent example of the impact this difference in method can have on interpretations of isotopic data is a re-analysis of horse tooth enamel phosphate data from last interglacial-glacial cycle contexts at the Hallera Avenue site, Wrocław (Poland) (3 measurements ranging between 13.4‰ and 14.1‰; SKRZYPEK et al., 2011, Supplementary Data). The isotopic data were interpreted as indicating temperatures 2–4ºC higher than previous estimates for the site based on pollen analyses (SKRZYPEK et al., 2011). In this analysis, the $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–T calibrations were made using transposed fits of a calibration derived from a dataset from SÁNCHEZ CHILLÓN ET AL. (1994). We recalculated these figures using forward and transposed fits of a more commonly-used equation for calibrating horse $\delta^{18}O$ (DELGADO HUERTAS ET AL. 1995; Table 2, Figure 2). When an inverted forward regression fit is used to calibrate the $\delta^{18}O_{bioapatite}$ data, the resulting $\delta^{18}O_{precipitation}$ estimates are 1–2‰ lower, and the estimated temperatures are 5–7ºC lower, than when a transposed regression is used. The point here is not to challenge the specific interpretations given by SKRZYPEK et al. (2011), but to provide a clear illustration of the significant effects that transposing the calibration equations can have on the resulting predicted $\delta^{18}O_{precipitation}$–T values.

Some studies have attempted to avoid the problem of asymmetry between inverting the forward least-squares regression $y(x)$ and the transposed regression

$x(y)$ by instead calculating $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–T conversion relationships using Reduced Major Axis (RMA) regression (van Dam and Reichart, 2009; Matson and Fox, 2010). RMA yields an equation with a slope that can also be related to the correlation coefficient; the RMA slope is $r/a = c/r$, which is equal to the geometric mean of the two slopes given by forward and transposed least-squares regressions, and thus predicts values that fall between these solutions (Figure 1A). The two least-squares regressions and the RMA regression based on the same data all intersect at the mean $(\bar{x}, \bar{y})$. Yet they will systematically diverge from each other, both as the correlation coefficient r^2 becomes smaller, and with increasing distance from the mean. Given these facts, it is pertinent to ask whether one method is more appropriate than another for the interpretation of palaeoclimatic $\delta^{18}O_{bioapatite}$ data? Two main factors are relevant for discussing this question: the partitioning of error between *x* and *y*, and the direction of causality between the variables.

3.1 Error partitioning

In a least squares regression analysis, the effects of any (measurement) uncertainties in the independent controlling variable *x* are assumed to be negligible in comparison to the statistical variability in the dependent variable *y* for a given value of *x*. The underlying statistical model is $y = ax + b + e$, where the coefficients *α* and *β* give the true correlation line for the whole population from which the data sample is drawn (whereas *a* and *b* are estimates of *α* and *β* from the data), and where e is a random variable with a zero mean that reflects natural variability about any less-than-perfect correlation, perhaps due to unknown variables other than *x* that also affect *y*. The forward least-squares fit *y*(*x*) is calculated by minimising the sum of the squared *y*-distances between each datapoint and the best fit line (Figure 1B). This assumes that

100% of the residual misfit is associated with the variability or uncertainty in *y*, including when the formula is used in its inverted form $x = (y - b)/a$. Conversely, the transposed fit *x*(*y*) minimizes the sum of the squared *x*-distances between the datapoint and the line, assuming that 100% of the residual misfit is associated with uncertainty in *x* (Figure 1C).

It is obvious in practice that the datasets used to generate equations for palaeoclimatic reconstruction have measurement errors in both *x* and *y*, which should be considered additional to the errors associated with natural variability in the dependent variable *y*. For example, in conversion Z1, $\delta^{18}O_{drinking\ water}$ is typically poorly known, being estimated using $\delta^{18}O_{precipitation}$ data from local or regional International Atomic Energy Agency monitoring stations that may not include (or be restricted to) data from the years when the analysed fauna were alive, rather than being estimated from water sources actually consumed by fauna (AYLIFFE et al., 1992; HOPPE, 2006; SÁNCHEZ CHILLÓN et al., 1994); $\delta^{18}O_{bioapatite}$ can generally be measured more precisely, yet sources of sampling variability may include such factors as the time period represented by the analysed sample. If the sizes of the errors were known – typically they are not – then a generalised least-squares method could be used to assign a specified proportion of the misfit to each variable, and the resultant slope would fall between those of the inverted forward fit and the transposed fit. RMA constitutes a specific example of this, making the overly simplistic assumption that the errors in *x* and *y* are proportional to the magnitude of the overall range in each variable (SMITH, 2009), which is equivalent to minimising the sum of the triangular areas formed between each datapoint and the line of best fit in both the *x* and *y* directions (Figure 1D). The best argument for this assumption is that *x* and *y* are treated symmetrically in the minimisation, and thus calibrations produced using RMA do not depend on whether the data is transposed or not. It is not an appropriate assumption, however, when most of the misfit is probably due to natural variability in *y*.

3.2 Direction of causality

The symmetry of RMA analysis between *x* and *y*, and the acknowledgement of error in both axes, suggests that it may be appropriate in situations where the two variables are co-dependent on other causes, and it seems arbitrary which variable is placed on which axis. For example, in conversion between $\delta^{18}O_{phosphate}$ and $\delta^{18}O_{carbonate}$ (A2), the two variables are directly related but one is not dependent on the other; rather, they co-vary according to the composition of a third variable – the $\delta^{18}O$ of body water. Accordingly, we suggest that RMA be considered for conversions A1 and A2 (although both datasets show such high r^2 coefficients that the difference between the least squares and RMA solutions would be small).

In contrast, we argue here that RMA is not the appropriate method for conversions Z1 and Z2 due to the causal relationship between the two variables in each conversion, which are related because one is dependent on the other, i.e. there is a causal stimulus and resulting effect. For example, the value of $y=\delta^{18}O_{bioapatite}$ is a dependent variable, controlled by the independent variable $x=\delta^{18}O_{drinking\ water}$ (with some natural variability due to other factors such as physiology and food) and no possibility for $\delta^{18}O_{bioapatite}$ to impact back directly on $\delta^{18}O_{drinking\ water}$. The critical point here is the asymmetry of the relationship being investigated. In situations where *x* "causes" *y*, it is statistical good practice and appropriately representative of the physical relationship between the variables to place the independent variable on the *x*-axis and calculate a fit of *y*(*x*), thus preserving the direction of cause and effect (see also POLLARD et al., 2011 and SMITH, 2009). For $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–T conversions, the most appropriate method is thus a forward least squares analysis, following the direction of causality and then inverting the relationship to $x=(y-b)/a$; this is indeed consistent with the way in which the vast majority of conversion relationships have been published. We discourage the use of transposed regression and RMA for these conversions, as statistically inappropriate for the causal relationships used in the Z1 and Z2 calibrations, and we note again that they are possibly misleading since they have lower slopes, r^2/a and r/a respectively, than the slope $1/a$ of inverted forward regression (see earlier discussion of slopes).

3.3 Theory of error and error estimation

Palaeoclimatic researchers have an understandable desire to draw firm conclusions about past temperatures from the isotopic measurements of palaeontological and archaeological samples. It is important, nevertheless, to keep track of the statistical uncertainties that are inevitably associated with reconstructions based on least-squares regressions, and these are not always quoted. In this section we discuss the nature of the statistical uncertainties, explain how they can be calculated and conclude with two key equations 5 and 6 that may be used for error estimation in the conversions Z1 and Z2. In the next section we then illustrate the use of these equations by way of case studies.

The uncertainties in conversions may be divided into two main categories: (1) those concerning the initial calibration by estimation of the line of best fit for the population from a finite dataset and (2) those concerning the natural variation of new samples around the line. Both are ultimately due to the fact that there is a natural spread of data around any correlation that cannot therefore be described as providing a direct prediction of *y* from *x*. This is often due to the impact of other external factors, for example, the impact of humidity, evapotranspiration effects or intra-population variability on the $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$ conversion (see also the discussion of natural variation in SMITH, 2009). As the variables $\delta^{18}O_{bioapatite}$ and $\delta^{18}O_{precipitation}$ are not 100% dependent upon each other, deviations from a line of best fit are inevitable even if the measurement errors are negligible. This variation cannot be controlled or reduced by the investigator, but is a natural property of the system being investigated, and it should be estimated when using the conversion formula to calibrate isotopic data.

Recall that the underlying statistical model is $y = ax + b + \epsilon$, where α and β give the true correlation line for the whole population, and ε is a random variable that represents the effects of all the unknown variables that impact on the calibration relationship. (The parameters α and β are unknown because we can only ever have a sample from the whole population.) When α and β are estimated by a least-squares fit ($y = ax + b$) to a dataset containing a random sample of *n* values (x_i, y_i) from this population, the inherent uncertainty, if reported, is often given in the form $y = (a \pm da)x + (b \pm db)$. It is, however, statistically more appropriate to write $y=ax+b \pm \delta y$, where the formula

$$\delta y = \sqrt{\delta \bar{b}^2 + \delta a^2 (x - \bar{x})^2}$$

Equation 2

gives a one-standard-deviation estimate of the uncertainty in the least-squares fit at position *x*, and

$$\delta a = \frac{s_{y/x}}{\sqrt{\sum(x_i - \bar{x})^2}} \qquad \delta \bar{b} = \frac{s_{y/x}}{\sqrt{n}}$$

and

$$s_{y/x} = \sqrt{\frac{\sum(y_i - a x_i - b)^2}{n-2}}$$

Equation 3

Here, δa is an estimate of the uncertainty in the slope, $\delta \bar{b}$ is an estimate of the uncertainty in the fit at $x = \bar{x}$, and $s_{y/x}$ is an estimate of the standard deviation of the natural variability in ε. Three critical points to note are: (i) the uncertainty in the fit is proportional to the natural variation $s_{y/x}$ about the fit; (ii) the uncertainty decreases as the size n of the dataset increases; (iii) the uncertainty increases with distance $x - \bar{x}$ from the mean of the dataset, which is a warning against extrapolation. We note also that regression software typically returns the value $\delta b = \delta \bar{b} + |\delta a \, \bar{x}|$ of the uncertainty in the fit at $x = 0$ rather than $\delta \bar{b}$, and thus δb may substantially overestimate the uncertainties of calibrated $\delta^{18}O$ or temperature values if, as is usual, these are not centred around $x = 0$ (which is sometimes known as the lever effect).

We now apply this model to assess the magnitude of the errors in categories (1) and (2) when evaluating data using an inverted calibration equation $x = (y - b)/a$. First, we note that the least-squares fit is itself uncertain. Following MILLER and MILLER (1984), we can approximate the uncertainty in the inverted correlation line by writing $x = (y - b)/a + \delta x$, where:

$$\delta x = \frac{s_{y/x}}{a}\sqrt{\frac{1}{n} + \frac{(y-\bar{y})^2}{a^2\sum(x_i-\bar{x})^2}}$$

Equation 4

(Equation 4 can be derived from Equations 2 and 3 and the relationship $(y-\bar{y}) = a(x-\bar{x})$ which follows from $b = \bar{y} - a\bar{x}$.)

Second, we note that when using sample data for palaeoclimatic reconstruction, each of these samples is subject to the natural variability ε. Therefore the mean y_0 of the samples is not equivalent to the population mean y at a given location, just as a particular mammoth tooth is unlikely to be typical of the population as a whole. If we have *m* independent samples (where *m* may only be 1) and the mean of those samples y_0 then the value of $x_0 = (y_0 - b)/a$ inferred from the calibration relationship is subject to an uncertainty (MILLER and MILLER, 1984; POLLARD et al., 2011):

$$\delta x_0 = \frac{s_{y/x}}{a}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_0-\bar{y})^2}{a^2\sum(x_i-\bar{x})^2}}$$

Equation 5

In many practical examples, the number *n* of datapoints used to generate the correlation is much greater than the number *m* of independent samples, and thus the natural variability of these samples will then dominate any uncertainty from the correlation.

Finally, there are many situations where researchers may wish to take estimates x_0 of $\delta^{18}O_{precipitation}$ generated by conversion Z1, and use a further calibration $T = (x - b_T) / a_T$ to generate an estimate of temperature from the value of x_0 (conversion Z2). The uncertainty in this temperature can be obtained using a similar formula to Equation 5, but this time using the uncertainty δx_0 previously calculated for the $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$ calibration in place of a sample variability $s_{x/T} / \sqrt{m}$. This gives:

$$\delta T_0 = \frac{1}{a_T} \sqrt{\delta x_0^2 + \frac{s_{x/T}^2}{n_T} + \frac{s_{x/T}^2 (x_0 - \bar{x}_T)^2}{a_T^2 \sum (T_i - \bar{T})^2}}$$

Equation 6

where n_T and $\bar{x}_T$ are values from the temperature calibration dataset. It is important to note that Equation 6 is used to estimate errors at the Z2 conversion stage only when using values of x_0 inferred from conversion Z1 with uncertainty dx_0 inferred from Equation 5. (If a Z2 conversion were applied to m_T direct observations of x_0 ($\delta^{18}O_{precipitation}$) then an equation analagous to Equation 5 would be used instead.)

Equations 4–6 are all simple estimates of one-standard-deviation uncertainty for the relevant variable. This is certainly sufficient to get a feel for the magnitude of the uncertainties, though rigorous hypothesis testing should be based on confidence intervals in a Student's *t*-test (POLLARD et al., 2011). For ease of use, these equations have been programmed into a spreadsheet that is available with this article, downloadable from the journal website (Supplementary Data).

4. Application and propagation of errors

Having outlined the theory of error and error estimation, we now assess some of the implications for the way that palaeoclimatic inferences are drawn from isotopic data, and provide examples of the conversion $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–T using published data. A key point is that this is a two-stage process, and that errors produced in the first stage must be propagated through to the second stage. Our approach has been developed for a particular context, that of vertebrate isotopic data, but may be used in other geochemical contexts.

4.1 Errors in the conversion from $\delta^{18}O_{bioapatite}$ to $\delta^{18}O_{precipitation}$ (Z1)

To illustrate the errors associated with this conversion, we have re-analysed two datasets from previous studies (horse and mammoth $\delta^{18}O_{bioapatite}$)(AYLIFFE et al., 1992; DELGADO HUERTAS et al., 1995) using Equations 4 and 5 to obtain the error estimates for an inverted forward regression (Figure 2). The error lines show how uncertainty in the lines of best fit is least around the dataset mean $(\bar{x}, \bar{y})$ and increases with distance from the mean, for both the uncertainty in the fit, calculated using Equation 4 (dark grey region in Figure 2) and the total uncertainty δx_0 incorporating the natural variability of the population, calculated using Equation 5 (light grey region in Figure 2). The total error associated with converting a single $\delta^{18}O_{bioapatite}$ measurement (i.e. $m = 1$) to $\delta^{18}O_{precipitation}$ using $x = (y - b)/a$ remains relatively constant for different values of y, since it is dominated by the estimate of the natural variability in the sample data (the first term in the square root of Equation 5).

Considering Equation 5, it is clear that the errors associated with calibration will be smaller if a larger number of samples are averaged together, thus reducing the size of the term $1/m$. The effects of sample size may be illustrated by calculating the errors associated with converting $\delta^{18}O_{bioapatite}$ values in the range 10‰–20‰ to estimates of $\delta^{18}O_{precipitation}$. Comparing conversions from increasing sample sizes of 1, 5 and 20 individuals with a mean $\delta^{18}O_{bioapatite}$ value of 10‰, we see that the errors are reduced from 1.7‰ to 1.1‰ in mammoth and 2.8‰ to 1.6‰ in horses; larger reductions are seen for mean $\delta^{18}O_{bioapatite}$ values of 20‰ since these are closer to the regression mean (Table 3). Whilst increasing sample sizes does reduce the error, a larger reduction is always seen between sample sizes of 1 and 5 than between 5 and 20 (indeed, the largest drop is from m = 1 to m = 2). That the greatest reduction in error is seen when analysing two samples rather than just one emphasises that it is worth making a significant effort to get more than one sample from each layer; however, after a few samples, the extra effort of continuing to reduce 1/m has little extra impact, as the error tends towards that of the regression line. These calculations clearly indicate the benefit of sampling multiple individuals to obtain a better estimate of the population-level mean $\delta^{18}O_{bioapatite}$, which can more than halve the error compared to single measurements in some cases.

The effects of sample size can be further illustrated with an example of recently published data. In their investigation of early-mid Pleniglacial climate in Poland, SKRZYPEK et al. (2011) calibrate their oxygen isotopic data from bioapatite to temperature using transposed fits of $x(y)$ but do not report the associated errors. When their data for mammoth and horse samples are reprocessed using the methods outlined in this paper (using the equations of AYLIFFE et al.1992 and DELGADO HUERTAS et al. 1995), the errors in T are calculated to be ±4.3–4.6ºC and ±8.0ºC respectively. Treating each sample individually, these errors are too large to offer a detailed interpretation of palaeoclimate. However, by using the mean of two mammoth samples and the two horse samples from the same layer, the errors fall to ±3.3ºC and

±5.9ºC respectively. If ten individuals had been sampled for each layer these errors could have been reduced to <2ºC.

A previous assessment of calibration errors investigated the conversion of human $\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$, and calculated errors of at least 1–3.5‰ (POLLARD et al., 2011). This study concluded that these errors were too large for the calculated $\delta^{18}O_{precipitation}$ values to be used for pin-pointing the geographic origin of individuals within the UK due to the limited natural variability in UK groundwaters. This is an interpretive problem in which it is desired to interpret each sample individually, and thus averaging between individuals cannot be used to reduce the uncertainty. In situations where multiple individuals can be sampled, however, such as the investigation of palaeotemperature through faunal remains as discussed in this article, it is possible to reduce the uncertainty by increasing *m* and obtain a more accurate estimate of the mean value of *y* (i.e. of y_0 in equations 4, 5 and 6). This substantially reduces the conversion errors overall. The sensitivity of the calibration equations to the number of measured samples has critical importance for determining whether the research questions of interest can legitimately be answered when calibrating the data, or whether the associated errors will be too large. Calibration may not be sufficient to answer the question, particularly for individual samples or smaller assemblages where a cohesive group of samples cannot be obtained.

4.2 Propagation of errors into the conversion from $\delta^{18}O_{precipitation}$ to temperature (Z2)

Moving to the second stage of the conversion process, we now consider what are the implications of the quantified errors in the Z1 conversion when propagated through into the Z2 conversion of $\delta^{18}O_{precipitation}$ to temperature. Unlike for conversion Z1, there are no standard equations for this stage, but rather there are many equations that have been used, which follow from a particular choice of dataset to construct each equation. Researchers typically generate a $\delta^{18}O_{precipitation}$–T conversion dataset relevant to their study by compiling the readily available data from one or a number of monitoring stations in the GNIP network over a global, continental, or regional geographic area (Kovács et al., 2012; Skrzypek et al., 2011); other potential calibration equations have also been calculated (Duliński et al., 2001; Gourcy et al., 2005; Rozanski et al., 1993; Tütken et al., 2007; Ukkonen et al., 2007; von Grafenstein et al., 1996). Each of these datasets will generate a slightly different estimated temperature for a given value of $\delta^{18}O_{precipitation}$. For example, Table 4 shows the temperatures and errors estimated from horse $\delta^{18}O_{bioapatite}$ using five different datasets taken from the GNIP network for the Z2 conversion (see also Table 1). We illustrate the effect of varying numbers of enamel analyses (1, 5, 10, 20), but all with a mean $\delta^{18}O_{bioapatite}$ of 15‰, equating to $\delta^{18}O_{precipitation}$ of –10.7‰.

Three significant points are highlighted. Firstly, the crucial effect of palaeo-sample size *m* is again evident: the dominant influence on the errors at the Z2 conversion stage is the number of horse samples analysed (*m*) and the consequent magnitude of the error in the Z1 conversion (δx_0). The term δx_0^2 dominates the other terms in the square root in Equation 6 so that, to a good approximation, $\delta T_0 \approx \delta x_0 / a_T$, and the statistical uncertainty in the regression line for a particular dataset has little effect (see Figure 3). But as we discuss below, it does not follow that the choice of dataset has little effect.

Secondly, the choice of dataset and thus regression equation can make a big difference to the estimated magnitude of error for a given number of samples. In the example we show, conversions based on annual temperature/precipitation data give markedly smaller errors than the equations based on monthly data (compare the conversions based on data from Kraków and Vienna: Table 4). This is because the spread of the annual and monthly data are different, influencing the slope a_T of the $\delta^{18}O_{precipitation}$–T regression line: for the annually averaged data, the slope is approximately twice as large as that for the monthly data and, as noted above, $dT_0 \approx dx_0 / a_T$. The choice between monthly and annual data should, however, be made on grounds of biological suitability, such as the nature of the temporal averaging in the faunal sample, rather than simply to minimise error estimates.

Thirdly, though the statistical uncertainty in the regression line for a given dataset is typically less than 0.2ºC (Table 1), the temperatures inferred from the different datasets vary from 5.8ºC (General Europe) to 8.7ºC (Vienna, annual). However, if the number of faunal samples is small then, allowing for the uncertainty in the Z1 conversions, the temperature ranges predicted by the various equations largely overlap with each other (Figure 4). Only if 10 or 20 samples are available do the temperature ranges inferred from annual data at different locations start to separate.

The above discussion suggests that whilst the errors are mainly generated by the Z1 conversion ($\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$) and depend on sample size, the way that these errors are mapped through to temperature ranges depends on the choice of regression line for the Z2 conversion ($\delta^{18}O_{precipitation}$–T).

5. Concluding comments and recommendations

The correlations between temperature and the oxygen isotopic values of bioapatite and precipitation motivate the use of calibration for generating first-order estimates of palaeoclimatic variables indicated by faunal isotopic compositions. Calibration also permits direct comparisons between measurements based on $\delta^{18}O_{bioapatite}$ data and estimates of $\delta^{18}O_{groundwater}$ or temperature measured in other proxies such as palaeoaquifer waters, chironomids or pollen. Such multi-proxy comparative approaches represent a valuable interpretive tool in palaeoclimatic studies provided the limits and uncertainties of each method are acknowledged, which is not universally done. We offer the equations in this paper as a suitable means of quantifying the uncertainties associated with calibrating isotopic data.

In summary, we advocate the use of multiple samples where possible, but that a balance must be struck between reduced uncertainty and feasibility, both in terms of number of analyses and comparative data. The use of multiple samples (m>1) for each investigated assemblage reduces the population-level uncertainty through the factor 1/m in Equation 5. But after a certain point, when 1/m becomes smaller than other terms inside the square root of Equation 5, adding more samples will not significantly reduce the Z1 conversion error ($\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$) any further. For conversions of $\delta^{18}O_{bioapatite}$ data to temperature through both the Z1 and Z2 conversions ($\delta^{18}O_{bioapatite}$–$\delta^{18}O_{precipitation}$–Temperature), the use of larger numbers of samples results in smaller errors at both conversion stages. But the limiting factor on temperature estimates may often be the availability of appropriate comparative datasets. In such circumstances, one should be aware of the accuracy needed to make meaningful interpretations in a given case study.

We conclude by listing three recommendations for the statistical treatment of errors in the conversion of bioapatite oxygen isotope data to precipitation oxygen isotope values and temperature:

1. Use appropriate regression for the datasets being employed – we recommend inverted forward regression for conversions Z1 and Z2, and not transposed or RMA regressions.
2. To report errors in a regression line, use Equations 2 and 3 rather than the form $y = (a \pm da)x + (b \pm db)$, as is commonly produced by spreadsheet software.
3. To report errors in data conversion, use Equations 5 and 6 which appropriately estimate this uncertainty.

These recommendations are not a comprehensive list, but offer an important set of guidelines regarding the calculation of error estimates.

Acknowledgements

We thank the members of the Dorothy Garrod Isotopes Laboratory at the McDonald Institute for Archaeological Research for their helpful comments on this text. We are grateful to the University of Cambridge (AJEP) and the Royal Society (RES) for financial support.

References

Arppe, L. and Karhu, J. A., 2010. Oxygen isotope values of precipitation and the thermal climate in Europe during the middle to late Weichselian ice age. *Quaternary Sci Rev* **29**, 1263-1275.

Ayliffe, L. K., Chivas, A. R., and Leakey, M. G., 1994. The retention of primary oxygen-isotope compositions of fossil elephant skeletal phosphate. *Geochimica et Cosmochimica Acta* **58**, 5291-5298.

Ayliffe, L. K., Lister, A. M., and Chivas, A. R., 1992. The preservation of glacial-interglacial climatic signatures in the oxygen isotopes of elephant skeletal phosphate. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **99**, 179-191.

Bernard, A., Daux, V., Lécuyer, C., Brugal, J.-P., Genty, D., Wainer, K., Gardien, V., Fourel, F., and Jaubert, J., 2009. Pleistocene seasonal temperature variations recorded in the $\delta^{18}O$ of *Bison priscus* teeth. *Earth Planet Sci. Lett.* **283**, 133-143.

Bos, J. A. A., Bohncke, S. J. P., Kasse, C., and Vandenberghe, J., 2001. Vegetation and climate during the Weichselian Early Glacial and Pleniglacial in the Niederlausitz, eastern Germany - macrofossil and pollen evidence. *Journal of Quaternary Science* **16**, 269-289.

Dansgaard, W., 1964. Stable isotopes in precipitation. *Tellus* **16**, 436-468.

Delgado Huertas, A., Iacumin, P., Stenni, B., Sánchez Chillón, B., and Longinelli, A., 1995. Oxygen isotope variations of phosphate in mammalian bone and tooth enamel. *Geochimica et Cosmochimica Acta* **59**, 4299-4305.

Duliński, M., Florkowski, T., Grabczak, J., and Różański, K., 2001. Twenty-five years of systematic measurements of isotopic composition of precipitation in Poland. *Przegląd Geologiczny* **49**, 250-256.

Fabre, M., Lécuyer, C., Brugal, J.-P., Amiot, R., Fourel, F., and Martineau, F., 2011. Late Pleistocene climatic change in the French Jura (Gigny) recorded in the $\delta^{18}O$ of phosphate from ungulate tooth enamel. *Quaternary Res.* **75**, 605-613.

Forbes, M. S., Kohn, M. J., Bestland, E. A., and Wells, R. T., 2010. Late Pleistocene environmental change interpreted from $\delta^{13}C$ and $\delta^{18}O$ of tooth enamel from the Black Creek Swamp Megafauna site, Kangaroo Island, South Australia. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **291**, 319-327.

Fricke, H. C., O'Neil, J. R., and Lynnerup, N., 1995. SPECIAL REPORT: Oxygen isotope composition of human tooth enamel from medieval Greenland: Linking climate and society. *Geology* **23**, 869-872.

GNIP/WISER database: Water Isotope System for Data Analysis, Visualization, and Electronic Retrieval. Accessed February 2011.

Gourcy, L. L., Groening, M., and Aggarwal, P. K., 2005. Stable oxygen and hydrogen isotopes in precipitation. In: Aggarwal, P. K., Gat, J. R., and Froehlich, K. F. O. (Eds.), *Isotopes in the Water Cycle*. Springer, Vienna. p. 39-51.

Grimes, S. T., Mattey, D. P., Hooker, J. J., and Collinson, M. E., 2003. Paleogene paleoclimate reconstruction using oxygen isotopes from land and freshwater organisms: the use of multiple paleoproxies. *Geochimica et Cosmochimica Acta* **67**, 4033-4047.

Hallin, K. A., Schoeninger, M. J., and Schwarcz, H. P., 2012. Paleoclimate during Neandertal and anatomically modern human occupation at Amud and Qafzeh, Israel: the stable isotope data. *J. Hum. Evol.* **62**, 59-73.

Hoppe, K., 2006. Correlation between the oxygen isotope ratio of North American bison teeth and local waters: Implication for paleoclimatic reconstructions. *Earth Planet Sci. Lett.* **244**, 408-417.

Iacumin, P., Di Matteo, A., Nikolaev, V., and Kuznetsova, T. V., 2010. Climate information from C, N and O stable isotope analyses of mammoth bones from northern Siberia. *Quaternary Int.* **212**, 206-212.

Kohn, M., 1996. Predicting animal δ^{18}O: Accounting for diet and physiological adaptation. *Geochimica et Cosmochimica Acta* **60**, 4811-4829.

Kohn, M. J. and Welker, J. M., 2005. On the temperature correlation of δ^{18}O in modern precipitation. *Earth Planet Sci. Lett.* **231**, 87-96.

Kovács, J., Moravcová, M., Újvári, G., and Pintér, A. G., 2012. Reconstructing the paleoenvironment of East Central Europe in the Late Pleistocene using the oxygen and carbon isotopic signal of tooth in large mammal remains. *Quaternary Int.* **276-277**, 145-154.

Krzemińska, A., Stefaniak, K., Zych, J., Wojtal, P., Skrzypek, G., Mikołajczyk, A., and Wiśniewski, A., 2010. A late Pleistocene woolly mammoth from lower Silesia, SW Poland. *Acta Zool. Cracoviensia* **53A**, 51-64.

Lécolle, P., 1985. The oxygen isotope composition of landsnail shells as a climatic indicator: Applications to hydrogeology and paleoclimatology. *Chemical Geology: Isotope Geoscience section* **58**, 157-181.

Longinelli, A., 1984. Oxygen isotopes in mammal bone phosphate: a new tool for paleohydrological and paleoclimatological research? *Geochimica et Cosmochimica Acta* **48**, 385-390.

Lucy, D., Robertson, I., Aykroyd, R. G., and Pollard, A. M., 2008. Estimates of uncertainty in the prediction of past climatic variables. *Applied Geochemistry* **23**, 2961-2965.

Luz, B. and Kolodny, Y., 1985. Oxygen isotope variations in phosphate of biogenic apatites, IV. Mammal teeth and bones. *Earth Planet Sci. Lett.* **75**, 29-36.

Luz, B., Kolodny, Y., and Horowitz, M., 1984. Fractionation of oxygen isotopes between mammalian bone-phosphate and environmental drinking water. *Geochimica et Cosmochimica Acta* **48**, 1689-1693.

Matson, S. D. and Fox, D. L., 2010. Stable isotopic evidence for terrestrial latitudinal climate gradients in the Late Miocene of the Iberian Peninsula. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **287**, 28-44.

Miller, J. C. and Miller, J. N., 1984. *Statistics for analytical chemistry*. Ellis Horwood, Chichester.

Navarro, N., Lécuyer, C., Montuire, S., Langlois, C., and Martineau, F., 2004. Oxygen isotope compositions of phosphate from arvicoline teeth and Quaternary climatic changes, Gigny, French Jura. *Quaternary Res.* **62**, 172-182.

Pollard, A. M., Pellegrini, M., and Lee-Thorp, J. A., 2011. Technical note: some observations on the conversion of dental enamel $\delta^{18}O_p$ values to $\delta^{18}O_w$ to determine human mobility. *Am. J. Phys. Anthropol.* **145**, 499-504.

Pryor, A. J. E., O'Connell, T. C., Wojtal, P., Krzemińska, A., and Stevens, R. E., 2013. Investigating climate at the Upper Palaeolithic site of Kraków Spadzista Street (B), Poland, using oxygen isotopes. *Quaternary Int.* **294**, 108-119.

Rozanski, K., Araguás-Araguás, L., and Gonfiantini, R., 1992. Relation between long-term trends of oxygen-18 isotope composition of precipitation and climate. *Science* **258**, 981-985.

Rozanski, K., Araguás-Araguás, L., and Gonfiantini, R., 1993. Isotopic patterns in modern global precipitation. In: Swart, P. K., Lohmann, K. C., McKenzie, J., and Savin, S. (Eds.), *Climate change in Continental Isotopic Records*. American Geophysical Union, Washington DC. p. 1-36.

Sánchez Chillón, B., Alberdi, M. T., Leone, G., Bonadonna, F. P., Stenni, B., and Longinelli, A., 1994. Oxygen isotopic composition of fossil equid tooth and bone phosphate: an archive of difficult interpretation. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **107**, 317-328.

Skrzypek, G., Wiśniewski, A., and Grierson, P. F., 2011. How cold was it for Neanderthals moving to Central Europe during warm phases of the last glaciation? *Quaternary Sci Rev* **30**, 481-487.

Smith, R. J., 2009. Use and misuse of the reduced major axis for line-fitting. *Am. J. Phys. Anthropol.* **140**, 476-486.

Stevens, R. E., Balasse, M., and O'Connell, T. C., 2011. Intra-tooth oxygen isotope variation in a known population of red deer: Implications for past climate and seasonality reconstructions. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **301**, 64-74.

Tütken, T., Furrer, H., and Walter Vennemann, T., 2007. Stable isotope compositions of mammoth teeth from Niederweningen, Switzerland: Implications for the Late Pleistocene climate, environment, and diet. *Quaternary Int.* **164-165**, 139-150.

Ukkonen, P., Arppe, L., Houmark-Nielsen, M., Kjaer, K. H., and Karhu, J. A., 2007. MIS 3 mammoth remains from Sweden - implications for faunal history, palaeoclimate and glaciation chronology. *Quaternary Sci Rev* **26**, 3081-3098.

van Dam, J. A. and Reichart, G. J., 2009. Oxygen and carbon isotope signatures in late Neogene horse teeth from Spain and application as temperature and seasonality proxies. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **274**, 64-81.

von Grafenstein, U., Erlenkeuser, H., Müller, J., Trimborn, P., and Alefs, J., 1996. A 200 year mid-European air temperature record preserved in lake sediments: An extension of the $\delta^{18}O_p$-air temperature relation into the past. *Geochimica et Cosmochimica Acta* **60**, 4025-4036.

Zanazzi, A., Kohn, M. J., MacFadden, B. J., and Terry, D. O. J., 2007. Large temperature drop across the Eocene-Oligocene transition in central North America. *Nature* **445**, 639-642.

Zanchetta, G., Leone, G., Fallick, A. E., and Bonadonna, F. P., 2005. Oxygen isotope composition of living land snail shells: Data from Italy. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **223**, 20-33.