BayClump: Bayesian Calibration and Temperature Reconstructions for Clumped Isotope Thermometry

C. Román-Palacios[1,2], H. M. Carroll[1], A. J. Arnold[1], R. J. Flores[1], Q. Gan[1], S.V. Petersen[3], K. A. McKinnon[4], A. Tripati[1]

[1]Department of Atmospheric and Oceanic Sciences, Department of Earth, Planetary, and Space Sciences, Institute of the Environment and Sustainability, Center for Diverse Leadership in Science, University of California – Los Angeles, Los Angeles, CA 90095 USA

[2]School of Information, University of Arizona, Tucson, AZ, USA

[3]Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA

[4]Department of Statistics, Department of Atmospheric and Oceanic Sciences, and the Institute of the Environment and Sustainability, University of California, Los Angeles, CA, USA

Corresponding authors: Cristian Román-Palacios (cromanpa94@arizona.edu) and Aradhna Tripati (atripati@g.ucla.edu)

Key Points:

- We implement Bayesian methods for calibrating the carbonate 'clumped' isotope thermometer and reconstructing temperatures.
- Bayesian and ordinary least squares linear models recover regression parameters and reconstruct temperatures with the highest accuracy, and Bayesian regression with the highest precision.
- BayClump is a Shiny dashboard with web-interface and standalone versions, that facilitates the use of Bayesian models which are data intensive and non-Bayesian models for calibration and reconstruction by the broader community.

Abstract

Carbonate clumped isotope thermometry (Δ_{47}) is a temperature proxy that is becoming more widely used in the geosciences. Most calibration studies have used ordinary least squares linear regressions or York models to describe the relationship between Δ_{47} and temperature. However, Bayesian models have not yet been explored for clumped isotopes. There also has not yet been a comprehensive study assessing the performance of commonly used regression models in the field. Here, we use simulated datasets to compare the performance of seven regression models, three of which are new and fit using a Bayesian framework. While Bayesian and non-Bayesian ordinary least squares linear regression models show the best overall accuracy for calibrations, Bayesian models outperform other models in terms of precision, especially if datasets are sufficiently large (>50 data points). For temperature reconstructions where a given regression model is applied to predict temperature from Δ_{47}), Bayesian and non-Bayesian models show variable performance advantages depending on the the structure of errors in the calibration dataset. Overall, our analyses suggest that the advantages of using Bayesian models for calibrating and reconstructing temperatures using clumped isotope paleothermometry are realized through the use of large calibration datasets (>50 data points). When used with large datasets, Bayesian regressions are expected to substantially improve the accuracy and precision of (i) calibration parameter estimates and (ii) temperature reconstructions (e.g., typically improving precision by at least a factor of two). We implement our comparative framework into a new web-based interface, BayClump. This data tool should increase reproducibility by enabling access to the different Bayesian and non-Bayesian regression models.

Plain Language Summary

Inferring past temperatures is central to research in many areas of geoscience, evolutionary biology, and ecology. The carbonate clumped isotope geothermometer is becoming more widely used as a tool for reconstructing temperatures since it allows for direct constraints on carbonate mineral formation temperature. However, to date, no study has critically examined the relative performance of statistical models used to define the relationship between clumped isotopes and formation temperature, that in turn are used for temperature reconstructions. In this study, we develop new Bayesian models that, in contrast to classical linear regression models, are able to account for parameter uncertainty and use information from prior studies to infer regression parameters and reconstruct temperatures. These models have the potential to improve regression parameter estimation for clumped isotope calibrations and reduce uncertainties in paleotemperature predictions.

1 Introduction

A temperature proxy that has emerged as a potentially transformative tool in multiple disciplines is carbonate clumped isotope thermometry, which is based on the analysis of ^{13}C-^{18}O bond abundance in carbonate minerals (e.g., Schauble et al., 2006; Ghosh et al., 2006; Eiler,

2007; Eagle et al., 2010; Passey et al., 2010; Tripati et al., 2010, 2015; Henkes et al., 2013; Meinicke et al., 2020), that is referred to using the notation Δ_{47} (Eiler and Schauble, 2004). A major advantage of clumped isotope thermometry is that it is based solely on thermodynamics, and therefore allows the simultaneous determination of both carbonate formation temperature and the oxygen isotopic composition of source water ($\delta^{18}O_{water}$) from a single measurement of a carbonate sample (Schauble et al., 2006; Hill et al., 2014, 2020). Furthermore, unlike more traditional carbonate-based proxies, the clumped isotope paleothermometer does not rely on assumptions about the phase or the $\delta^{18}O_{water}$ (Ghosh et al., 2006). The temperature dependence of carbonate clumped isotope thermometry has led to applications as broad-ranging as evolutionary biology (e.g., Eagle et al., 2011, 2015; Garzione et al. 2014; Pérez-Escobar et al. 2017), paleoclimate (Eiler 2011; Tripati et al., 2014; Leutert et al., 2019; Kelson et al. 2020), and paleoaltimetry (Garzione et al., 2014).

Δ_{47} has been found to scale linearly with $1/T^2$ across temperature ranges of 0–100 ºC, leading to the use of linear regression models to calibrate the temperature dependence of this proxy, and for the estimation of clumped isotope-derived temperatures (Ghosh et al., 2006; Eiler 2007). Most prior clumped isotope calibration studies have relied on either ordinary least squares linear (Ghosh et al., 2006) or error-in-variables regression models (e.g., Deming; Tripati et al., 2010; e.g., York; Kelson et al., 2017) for inferring model parameters that relate Δ_{47} and $10^6/T^2$ values (i.e., regression slope and intercept). The ordinary least squares linear and York regression models mostly differ in how they treat uncertainty in measured Δ_{47} and $10^6/T^2$. Each has their own advantages and limitations. For instance, ordinary least squares linear are already implemented in many statistical packages and a commonplace in the field. However, a clear limitation of ordinary least squares linear models is the inability to account for errors in $10^6/T^2$ from the modeling framework, even though error is intrinsic to both clumped isotope and temperature measurements used for deriving calibrations. Furthermore, the magnitude of uncertainty in Δ_{47} and $10^6/T^2$ varies for different calibration datasets (e.g., depending on material, instrumentation used, standardization, knowledge of temperature for environment samples are from) and ordinary least squares linear regressions treat all of these equally when different datasets are combined. In contrast, the York and Deming regression models account for error in both variables (e.g., Tripati et al., 2010; Peral et al., 2018; Meinicke et al. 2020; Anderson et al., 2021). Nevertheless, the performance of these two later models is still to be tested under simulated conditions that are relevant to the field.

To date and to our knowledge, no study has critically and comparatively evaluated the performance of error-in-variable regression models on the accuracy and precision of clumped isotope temperature calibrations. Similarly, although Bayesian frameworks have been used for other temperature proxies including TEX_{86} and Mg/Ca and have provided a more robust method for estimating uncertainties in tracer-based estimates of temperature (Tingley and Huybers, 2010; Tierney and Tingley, 2014, 2015; Khider et al., 2015; Tierney et al., 2019; Crampton-Flood et al., 2020; Martinez-Sosa et al., 2021), no study has utilized these frameworks for the calibration of clumped isotopes, or for reconstructing temperatures using Δ_{47}. Thus, it remains unclear whether accounting for uncertainties in both variables actually improves the reliability of inferred regression parameters and reconstructed temperatures using Δ_{47}, and how error-in-variable models compare to Bayesian methods.

In this study, we extend the classic regression approach for calibrating the clumped isotopes paleothermometer into a Bayesian framework, and compare Bayesian and non-Bayesian regression models utilizing synthetic datasets. We focus on answering whether Bayesian models and error-in-variable models outperform models that ignore uncertainty in Δ_{47}. Our main goal is to discuss relative model performance between newly developed and existing models commonly used for calibrating the clumped isotope paleothermometer. In this study, we do not intend to provide a general equation for analyzing clumped isotopes datasets. Instead, we provide a critical overview on the methods that are used in the field. Given the increasing availability of clumped isotope data, the Bayesian models developed in this study should help to pave the way for a unified calibration equation for the clumped isotopes paleothermometer.

Inspired by BAYSPAR, a web-interface for Bayesian models for the TEX_{86} temperature proxy (Tierney and Tingley, 2014), we developed BayClump, a shiny dashboard created in R that provides community-wide access to the Bayesian and non-Bayesian models for the clumped isotope proxy from this study. We derive calibration regression parameters using a published synthesis of calibration data (Petersen et al., 2019; Anderson et al. 2021; Sun et al. 2021). Overall, this work allows us to demonstrate the conceptual and practical advantages of using Bayesian models for inferring model parameters and deriving reliable reconstructions for clumped isotopes, as it has been outlined before in other temperature proxies (Tierney and Tingley 2015).

2 Materials and Methods

2.1 General modeling framework

We examine the performance of Bayesian and non-Bayesian linear models primarily using synthetic datasets (but see section 3.4 for analyses using real-world data). Tables 1, 2, S1, S2 show the range of uncertainties in Δ_{47}, T, and $10^6/T^2$ from existing calibration datasets, respectively. We use these distributions to define "low", "intermediate", and "high" uncertainties in each of the variables. Thus, the analyzed synthetic datasets, assuming the linear relationship between Δ_{47} and $10^6/T^2$, follow different levels of error in Δ_{47} and $10^6/T^2$.

Note that although the general practice in the field is to predict $10^6/T^2$ from Δ_{47} values to reconstruct temperature using a regression model defined from a temperature calibration dataset, our approach relied on a "forward" modeling where regression model parameters are estimated by using Δ_{47} as the response variable. This forward approach is consistent with Δ_{47} being a response to temperature, as opposed to the cause. Using synthetic datasets (with low, intermediate, or high uncertainties in Δ_{47} and $10^6/T^2$), we utilize different models to estimate regression parameters. Specifically, we compare the parameters inferred from each statistical model with the true parameters used to simulate the synthetic datasets. This approach allows us to assess whether different models (ordinary and weighted least squares linear, York, Deming, and Bayesian models) yield accurate and precise values for the slope and intercept. Finally, we utilize the inferred regression parameters and their uncertainties from each model to reconstruct temperatures for specific target Δ_{47} values of 0.600‰, 0.700‰, and 0.800‰ that correspond to

temperatures that are low (~10 ºC), moderate (~19 ºC), and high (~60 ºC). We account for different values of uncertainty in the analyzed target Δ_{47} (low, intermediate, and high).

2.2 Regression models

We fit seven types of regression models to the synthetic clumped isotope-Δ_{47}calibration datasets (Fig. 1). Four models are non-Bayesian regressions and three are Bayesian models. Note that we use Bayesian linear models as a way to propagate uncertainty in regression parameters. Additionally, depending on the Bayesian model, we also propagate uncertainty in the measurements of both $10^6/T^2$ and Δ_{47}. Model performance for proxy calibration is in this section assessed with $10^6/T^2$ as the independent variable and Δ_{47} as the response variable.

2.2.1 Non-Bayesian linear regression models

Ordinary least squares: We first fit an ordinary least squares linear regression model. This regression model is the simplest model used in this study and assumes no errors in $10^6/T^2$ (the independent variable in the regression). We fit the ordinary least squares linear regression model using the lm function in the stats R package version 4.1.0 (R Core Team, 2021) under default parameters. The approach implemented in the lm function in R minimizes the sum of squared error (i.e., sum over the squared of residuals in Δ_{47}) in the relationship between $10^6/T^2$ and Δ_{47}.

Weighted least squares: Second, we fit an ordinary least squares linear model with observations being weighted based on the inverse of their squared uncertainty. In this model, observations with larger residual values (estimated using ordinary least squares linear models) have less importance in estimating the error of alternative proposed lines during the least square optimization of the model. Although this approach accounts for variable uncertainty in Δ_{47}, the weighted least squares model still does not account for uncertainties in $10^6/T^2$. The weighted least squares regression was fit using the lm function in the stats R package version 4.1.0 (R Core Team, 2021). The weights argument is set to the inverse of the squared residuals for the observations.

Deming: Third, we fit a Deming regression using the deming R package version 1.4 (Therneau, 2018). In this study, the Deming regression model is the simplest model that explicitly accounts for measurement error in both Δ_{47} and $10^6/T^2$. With the Deming regression, the ratio of the variance in Δ_{47} and $10^6/T^2$ (calculated in the deming R package using jackknifing-based uncertainties on $10^6/T^2$ and Δ_{47}) is assigned to be constant over all data points (Martin, 2000). To fit this model, we specify values for Δ_{47} and $10^6/T^2$, along with the corresponding inverse of the squared standard error for each of the observations of temperature and Δ_{47}. The Deming model also aims to minimize the sum of squared residuals, where the residuals are a function of the inferred errors in both variables and the specified variance ratio (Deming, 1943).

York: Fourth, we analyzed a York model using the york function in the IsoplotR R package version 3.4 (Vermeesch, 2018). This approach is based on the same ideas that underlie the Deming regression model, specifically accounting for errors in both Δ_{47} and $10^6/T^2$. However, under the York model, the ratio of the weights in Δ_{47} and $10^6/T^2$varies across data points instead of being constant for the whole dataset as in the Deming regression (Martin 2000). Note that the weights are based on the correlation between errors in variables. We specify observations in Δ_{47} and $10^6/T^2$, along with the corresponding standard error (transformed to the inverse of the squared error within IsoplotR's york function) for each observation, when fitting York models.

2.2.2 Bayesian linear regression models

Bayesian linear: Fifth, we fit a Bayesian linear regression, the simplest Bayesian model fit in the study, and is equivalent to the ordinary least squares linear regression model presented above. For this regression, instead of parameter estimates being derived based on ordinary least squares optimization, regression parameters are estimated under a Bayesian framework (see below). Under a Bayesian approach, we use information from prior studies and newly generated clumped isotope data (synthetic datasets) to update the relevant regression parameters (e.g., slope and intercept) that are used in the calibration and reconstruction steps. Below, we present the mathematical definition of this model:

$$\Delta_{47_i} \sim Normal(\mu_i, \tau)$$
$$\mu_i = \alpha + \beta \frac{10^6}{T^2}_i$$
$$\alpha \sim Normal(0.231, 0.065)$$
$$\beta \sim Normal(0.039, 0.004)$$
$$\tau = \frac{1}{\sigma^2}$$
$$\sigma \sim U(0, 100)$$
$$i = 1, \ldots, N$$

Priors for α and β follow previous publications (Table S3; references therein). We use diffuse priors (priors that make weak assumptions about the model) on the precision parameter σ^2. We also present results that utilize diffuse priors for α and β by selecting the same mean parameter value outlined above but with three times a wider standard deviation.

Bayesian linear with errors: Sixth, we fit a linear regression model that accounted for uncertainties in both $10^6/T^2$ and Δ_{47}. The Bayesian linear model with error in variables is defined as the Bayesian linear model outlined above except for the following terms that account for measurement error in both variables:

$$\Delta_{47_i} \sim Normal(\Delta_{47_i}^{true}, \sigma_{\Delta_{47_i}}^{-2})$$
$$\Delta_{47_i}^{true} \sim Normal(\mu_i, \tau)$$
$$\frac{10^6}{T^2}_i \sim Normal(\frac{10^6}{T^2}_i^{true}, \sigma_{\frac{10^6}{T^2},i}^{-2})$$

$$\frac{10^6}{T^2}_i^{true} \sim Normal(11, 0.394)$$

The true values (i.e., those uncontaminated by measurement and sampling error) are indicated as Δ_{47}^{true} and $\frac{10^6}{T^2}^{true}$. We consider Δ_{47}^{true} values taken from realizations of a random variable with an underlying normal distribution, with unknown variance, and whose mean is linearly related to $\frac{10^6}{T^2}$. Therefore, the observed response is Δ_{47_i}, with errors $\sigma_{\Delta_{47}}$ (see also Daëron, 2021). The T explanatory variable is the temperature in the form $\frac{10^6}{T^2}$, with errors $\sigma_{\frac{10^6}{T^2}}$. Note that the prior used for $\frac{10^6}{T^2}_i^{true}$ corresponds to the mean value of temperature within the environmental range and a standard deviation reflecting high temperature uncertainty (Table 1).

Bayesian linear mixed: Seventh, we fit a Bayesian linear mixed model that accounts for error in both variables (Hilbe et al. 2007). This model is different from the above "linear with errors" in that it assumes that different calibration materials can potentially have distinguishable differences in the relationship between Δ_{47} and $10^6/T^2$. Note that for a single material, this regression model should behave similarly to the previous Bayesian regression model. We use the Bayesian linear mixed model to examine whether a relatively more complex model potentially assuming multiple materials under a single material dataset can still perform similarly to models that intrinsically assume equivalent material behavior. The utility of this model will be used in upcoming papers for assessing if there is evidence for material-specificity in real-world datasets. Below, we present the mathematical definition of this model. Except in the following aspects that allow for material-specific regression parameters, this model is equivalent to the *Bayesian linear with errors* presented above:

$$\mu_i = \alpha_{j(i)} + \beta_{j(i)} \frac{10^6}{T^2}_i$$
$$\alpha_j \sim Normal(0.231, 0.065)$$
$$\beta_j \sim Normal(0.039, 0.004)$$

$j \in \{1,2,\dots\}$, where j is an indicator of the type of material

Therefore, this model allows for material-specific regression parameters. Material identities are indicated under alternative j.

2.2.3 Implementation of Bayesian regression models

All three Bayesian regression models are fit using the jags function in R2jags version 0.6-1 in R version 4.02 under JAGS version 4.3.0 (Plummer 2003). Posterior distributions on parameter estimates are based on 20,000 iterations (three chains), with 50% of samples discarded

as burn-in. We used informative priors for the slope and intercept in alternative analyses that are presented in the supplement to this article. We use the same seed in R for all the analyses conducted in this study (set.seed() set to 3 in R). All the code and datasets used in this study are available on GitHub (https://github.com/Tripati-Lab/BayesPaper; https://github.com/Tripati-Lab/BayClump). BayClump can be accessed at the following link: https://bayclump.tripatilab.epss.ucla.edu/.

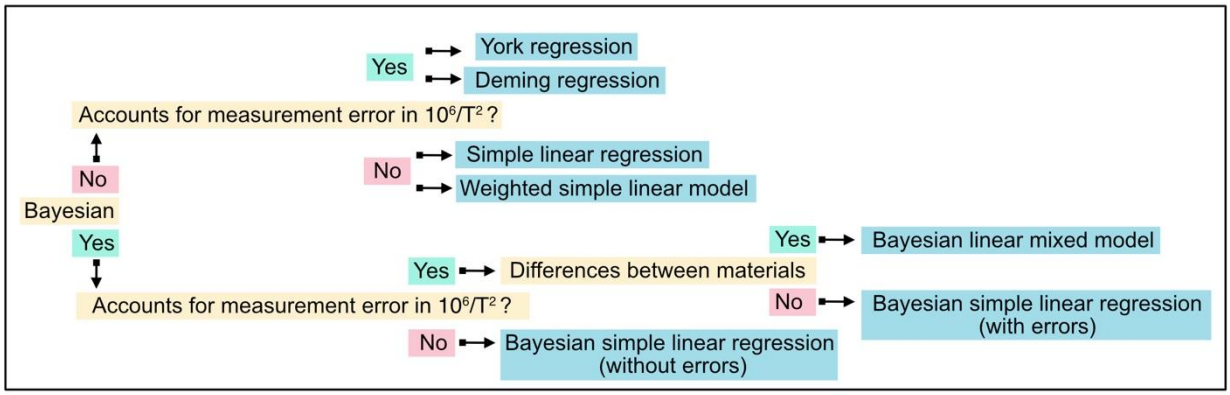

Fig. 1. Conceptual representation of the seven regression models used in this study for the derivation of Δ_{47}-temperature calibrations. We compared the performance of a total of seven regression models in deriving calibration relationships for use in carbonate clumped isotope thermometry. Parameter estimates, or slopes and intercepts of Δ_{47}-temperature calibrations derived for each of these models, were optimized through minimizing squared residuals (non-Bayesian models) or maximizing a likelihood function (Bayesian models). A subset of the classical least squares and Bayesian models account for error-in-variables (i.e., the regression parameters calculated factor in uncertainties in both Δ_{47} and $10^6/T^2$). We also developed a Bayesian model that can potentially account for differences in parameter estimates between materials or other types of sample groups. This model is equivalent in complexity to a Bayesian simple linear regression with errors when the number of materials is one.

2.3 Clumped isotope temperature proxy calibration: model performance based on parameter estimates on the synthetic datasets

Comparisons of model performance for parameter estimates were based on three synthetic datasets of simulated values of $10^6/T^2$ and Δ_{47}. We examine the performance of regression models by simulating errors in $10^6/T^2$ and Δ_{47}. We account for three sources of uncertainty in the $10^6/T^2$ - Δ_{47} relationship: replication error in Δ_{47} across labs, instrument noise in Δ_{47}, and errors in $10^6/T^2$. These three sources of uncertainty closely reflect the main three sources of uncertainty in real clumped isotope datasets, and span a realistic set of values reported across labs and Δ_{47}-$10^6/T^2$ relationships for different materials (Table 1, 2, S1, S2). For each source of error, we model different scenarios with low, intermediate, and high levels of error.

2.3.1 Replication error in Δ_{47} across labs (σ_p parameter)

Reproducibility error across labs can be caused by multiple factors, including the drift in instruments over time, choice of reference frame, and noise in the sample preparation for each replicate, among others. Therefore, this source of error is intrinsic to any clumped isotope dataset. Δ_{47} errors describing reproducibility across labs (see the σ_p parameter below), we follow the distribution of reported Δ_{47} errors in the Petersen et al. (2019) compilation. We estimate typical reproducibility of Δ_{47} by examining the distribution of reproducibility for each lab that participated in the Intercarb interlaboratory exercise (Supplemental Table 1 in Bernasconi et al., 2021; see Section 2.3.2 below). In our lab, long-term reproducibility of better than 0.02‰ is typical, and that for recent instrumentation, better than 0.01‰ is also routinely feasible with sufficiently large numbers of standards being run. We assign reproducibility errors of 0.0125‰, 0.0225‰, and 0.0275‰ corresponding to low, intermediate, and high error scenarios, respectively.

2.3.2 Instrument noise in Δ_{47} across labs (σ_b parameter)

Instrument noise can be related to the stability of beams over a 1 to 2-hour period, or even the length of time that is used to integrate on a single gas. For error in Δ_{47} caused by instrument noise, we use 0.0025‰, 0.0075‰, and 0.0125‰ for low, intermediate, and high error scenarios, respectively (Bernasconi et al. 2021). This error describes stability of the instrument across the ~1 to 2-hour analysis time for each replicate.

2.3.3 Errors in $10^6/T^2$

For measurement error in temperatures ($10^6/T^2$) used for proxy calibration, we define our levels of error by examining the typical uncertainties reported for different types of carbonates used in published studies that have compiled different calibration data (Petersen et al., 2019; Table 1). Low error in formation temperature is defined using reported values for synthetic carbonates (e.g., Ghosh et al., 2006; Tripati et al., 2015; Bonifacie et al., 2017), which have the most well constrained temperatures due to precipitating in controlled environments. Our estimates for high uncertainty for Δ_{47} error across labs reflects either naturally occurring terrestrial carbonates with larger variability in precipitation temperature, such as lacustrine samples (Huntington et al., 2010; Li et al., 2021; Wang et al., 2021) or naturally-occurring dolomites (Winkelstern et al., 2016; Came et al., 2017). Estimates for intermediate error fall in between those for low and high (e.g., foraminifera; Tripati et al., 2010; Meinicke et al., 2020; some naturally forming carbonates with less seasonal variability such as marine mollusks and brachiopods (Eagle et al., 2013; Henkes et al., 2013). We use 0.25°C, 2°C, 5°C as the low, intermediate, and high error scenarios for *T* when prescribing $10^6/T^2$, respectively.

2.3.4 Integrating sources of error into the simulated Δ_{47} and $10^6/T^2$ datasets

We analyze model performance for an "all-low error scenario", with low values of error in measurement error in Δ_{47} errors and measurement error in $10^6/T^2$. Next, we examine model performance in an "all-intermediate error scenario", with intermediate values of error in Δ_{47} and measurement error in $10^6/T^2$. Finally, we use an "all-high error scenario" with high error in Δ_{47} and measurement error in $10^6/T^2$. For simplicity, we refer to each of these as low-, intermediate-, and high-error scenarios for proxy calibration. Analyses in the main text primarily focus on three simplified end-member error scenarios (Data Set S1–S3).

For each error scenario, we simulate a total of 1,000 Δ_{47} and $10^6/T^2$ observations assuming a true value for the slope of 0.0369 and intercept of 0.268. These values were chosen because they represent the mean in the range of values from previous calibrations across different materials (see Table S3 and references therein). We first generated a total of 1,000 observations of $10^6/T^2$ with a normal distribution under the following parameters (informed using Table S2):

$$\frac{10^6}{T^2(true)} \sim Normal(12.03, 2.5)$$

These observations are treated as the true $10^6/T^2$ values (range of true $10^6/T^2$ in our dataset is between 5–19). Next, we simulate random error in $10^6/T^2$ using on a normal distribution with mean 0 and standard error following a given $10^6/T^2$ error scenario (σ; values of 0.019, 0.155, 0.070 for low-, intermediate-, and high-errors based on Table 1):

$$\frac{10^6}{T^2(error)} \sim Normal(0, \sigma)$$

The observed values of $10^6/T^2$ result from the addition of $\frac{10^6}{T^2(true)}$ and $\frac{10^6}{T^2(error)}$. Next, we simulate Δ_{47} values based on a given true slope, intercept, true $10^6/T^2$, and random error in Δ_{47} under a given error scenario (σ_b; values=0.0125‰, 0.0225‰, and 0.0275‰) based on Bernasconi et al. (2021; related to σ^2 in model 6):

$$\Delta_{47_i} \sim Normal(\mu_i, \sigma_b)$$
$$\mu_i = \alpha + \beta \frac{10^6}{T^2(true)}$$

Finally, we account for measurement error in Δ_{47} values representing replication error based on the error σ_p as estimated in Petersen et al. (2019; related to $\sigma^2_{\Delta_{47_i}}$ in model 6; values used in the simulation: 0.0025‰, 0.0075‰, 0.0125‰):

$$\Delta_{47_i}^{error} \sim Normal(0, \sigma_p)$$

Thus, $\Delta_{47_i}^{observed}$ values are calculated as the addition of each initial Δ_{47_i}value to a corresponding $\Delta_{47_i}^{error}$.

2.4 Fitting regression models on the simulated 'clumped isotope' datasets

We examine whether each of the models correctly recover the true slope and intercept for the calibration. For non-Bayesian models, we examine the distribution of slopes and intercepts using 1,000 replicates per model per error scenario. For Bayesian models, we run a total of 20,000 iterations (see details in Section 2.2.3). We analyze datasets with 10, 50, and 500 observations randomly sampled from the original 1,000 data point calibration dataset. Datasets with n=50 generally reflect the size of calibration datasets for individual materials in published clumped isotope calibration studies (e.g., Tripati et al., 2010; Petersen et al., 2019; Anderson et al., 2021).

2.5 Inverting the forward model to predict $10^6/T^2$ from Δ_{47}: Temperature reconstructions for unknowns

In addition to examining whether models accurately and precisely recover regression parameters (Section 2.4), we examine model performance during the temperature reconstruction phase. Paleotemperatures can be inferred by applying a regression model to sample Δ_{47} values. To evaluate model performance for temperature reconstructions, we apply the estimated regression parameters to three Δ_{47} values (0.600‰, 0.700‰, and 0.800‰) with several scenarios for replicate measurement errors in Δ_{47} (i.e., 0.005‰, 0.010‰, and 0.020‰). Our goal is to show whether reconstructions based on these carbonates under- or over-estimate true temperature under each of the examined regression models and scenarios of error. We reconstruct temperatures and their uncertainties following the usual practice in the field:

$$T\ (°\mathrm{C}) = \sqrt{\frac{\beta\ * 10^6}{\Delta_{47_i} - \alpha}} - 273.15$$

$$SE, T\ (°\mathrm{C}) = \left(\sqrt{\frac{\beta\ * 10^6}{\Delta_{47_i} - \alpha}} - 273.15\right) - \left(\sqrt{\frac{\beta\ * 10^6}{\left(\Delta_{47_i} + se\left(\Delta_{47_i}\right)\right) - \alpha}} - 273.15\right)$$

where, $T\ (°\mathrm{C})$ is the reconstructed temperature in degrees Celsius, $SE, T\ (°\mathrm{C})$ is the error in reconstructed temperature (also in degrees Celsius), Δ_{47_i} is the analyzed target Δ_{47_i}, $se\left(\Delta_{47_i}\right)$ the uncertainty in the analyzed target Δ_{47_i}, and both α and β the intercept and slope, both estimated during the calibration step. For non-Bayesian models, inversions were conducted using the invest function in the investr R package (Greenwell and Kabban, 2014). For Bayesian models, temperature reconstructions were conducted using Jags in two major steps (with uncertainty propagated across steps). First, we estimated the point estimate of temperature as follows:

$$\Delta_{47_k} \sim (\mu_k, \tau)$$

$$\mu_k\ = \alpha\ + \beta \frac{10^6}{T^2}_k$$

$$T_k \sim \sqrt{\frac{\beta\ * 10^6}{\Delta_{47_k} - \alpha}} - 273.15$$

$$\frac{10^6}{T^2}_k \sim Normal(11, 0.394)$$

$k \in \{1,2,\dots\}$, where k is an indicator of the target Δ_{47_k}

where, Δ_{47_k} is the analyzed target Δ_{47} , α and β, the intercept and slope (both corresponding to samples of the post-burnin distribution in the calibration step of the analyses), τ is the precision (also based on the post-burnin distribution in the calibration phase of the analyses), $\frac{10^6}{T^2}_k$ is the reconstructed temperature (in °K), with priors set to reflect an environmental temperature range (mean) and high uncertainty (standard deviation; Table 1). Finally, the reconstructed point estimate of temperature in degrees Celsius is summarized in T_k.

Note that this Bayesian reconstruction model follows the same structure as the three Bayesian calibration models presented before in Section 2.2.2. However, in the reconstruction section (Section 2.2.2), we focus on k target Δ_{47}, instead of i Δ_{47} from the calibration set. Similarly, values for α, β, and τ are derived from the posterior distribution estimates in the calibration step (Section 2.2.2) and not calculated here during the reconstruction phase of the analyses. We randomly sample 500 observations from the post-burnin posterior distribution for each of these parameters based on the calibration step. The α, β for Bayesian Mixed models are specific to the material that is utilized as an archive for reconstructing temperature and are matched to calibration-derived values.

Second, we estimated the uncertainty in reconstructed temperatures within a Bayesian framework by using a uniform prior for the true error in Δ_{47_k}. This prior on the temperature error was based on the uncertainty of the target Δ_{47_k} ($se(\Delta_{47_k})$, which is here assumed to be 0.005‰, 0.1‰, 0.2‰:

$$\Delta_{47_k}^{error} \sim Uniform(0.0001, se(\Delta_{47_k}))$$
$$\Delta_{47_k}^{*} = \Delta_{47_k} + \Delta_{47_k}^{error}$$
$$\mu_k^{*} = \alpha + \beta \frac{10^6}{T^2}_k^{*}$$
$$\frac{10^6}{T^2}_k^{*} \sim Normal(11, 0.394)$$
$$T_k{}^{*} \sim \sqrt{\frac{\beta * 10^6}{\Delta_{47_k}^{*} - \alpha}} - 273.15$$
$$T_{error;\,k} = T_k - T_k{}^{*}$$
$$Temp_k \sim Normal(T_k, T_{error;k}{}^{-2})$$

$k \in \{1,2,\dots\}$, where k is an indicator of the target Δ_{47_k}

Here, all the parameters follow the same structure outlined above for the point estimates of temperature reconstructions under Bayesian models. However, we note that instead of focusing on Δ_{47_k}, we used $\Delta_{47_k}^{*}$ for reconstructions. $\Delta_{47_k}^{*}$ is here used to estimate a second point estimate of temperature ($T_k{}^{*}$), which is subtracted from the first point estimate of temperature

(T_k) to calculate the error in reconstructed temperature ($T_{error;\,k}$). Finally, $Temp_k$ summarize the estimated temperature (in °C) with propagated uncertainty.

3 Results and Discussion

3.1 Model performance in calibration datasets with 50 data points

Most of the existing calibration datasets have 20–50 data points (e.g., Tripati et al., 2010; Petersen et al., 2019; Anderson et al., 2021). Therefore, we focus on results based on examining model performance on datasets with 50 observations. We found differences in the performance of classical and Bayesian models with our synthetic datasets for each error scenario considered. Our results suggest that while all the examined models are able to correctly recover true regression parameters, regressions differ in their accuracy and precision during the calibration stage (Figs. 2, 3). In general, Bayesian and non-Bayesian simple linear models generally outperform other methods for inferring regression parameters with the highest accuracy and precision. Precision and accuracy in parameter estimate under Bayesian and non-Bayesian linear models are not strongly affected by the analyzed error scenario. Deming and York regression models are consistently the least accurate and precise of the models. These models show the best performance when uncertainty in the analyzed calibration scenario is intermediate, again in datasets with n ~ 50 datapoints. Our results suggest that York regressions generally underestimate the intercept and overestimate the slope more strongly than any other model examined in this study. Conversely, Deming regressions generally overestimate the intercept and underestimate the slope.

3.1.1 Patterns of accuracy and precision between models within scenarios of error

Under a low-error scenario, accuracy and precision is relatively high for all the analyzed models (Figs. 2, 3). All the models slightly underestimate the true intercept. The slope is slightly underestimated by Bayesian models and slightly overestimated by non-Bayesian models. Overall, weighted and unweighted linear regression models show the best accuracy among the examined models and Bayesian models show the best precision. Specifically, York regressions underestimate the intercept by 2.7% and overestimate the slope by 1.5%. Deming regressions underestimate the intercept by 1.2% and overestimate the slope by 0.25%. Weighted and unweighted linear regression models underestimate the intercept by 1.3–1.6% and overestimate the slope by 0.7–0.87%. Finally, Bayesian models recover intercept values that are 1.3–1.4% smaller than true intercept and estimate slope values 0.15–0.2% smaller than the true slope. In terms of precision, Bayesian regressions recover at least 1.7 times less parameter uncertainty than any of the non-Bayesian models analyzed in this study.

Under an intermediate-error scenario, Bayesian models overestimate the intercept and underestimate the slope. Conversely, non-Bayesian models underestimate the intercept and overestimate the slope. Weighted and unweighted linear models show the best accuracy and Bayesian regressions the best precision. York regressions underestimate the true intercept by 7%

and overestimate the slope by 4%. Deming regressions underestimate the intercept by ~4% and overestimate the slope by 1.4%. Weighted and unweighted linear models underestimate the intercept by 2.6% and overestimate the slope by 1.4%. Finally, Bayesian models overestimate the intercept by ~4% and underestimate the slope by 3%. We also note that Bayesian models recover parameter estimates with at least 1.6 times less uncertainty than any of the alternatives.

Under a high-error scenario, patterns of model performance are similar to an intermediate-error scenario. York regressions underestimate the intercept by 8% and overestimate the slope by 5%. Deming regressions underestimate the intercept by 4% and overestimate the slope by 1%. Weighted and unweighted linear models underestimate the intercept by ~2.6% and overestimate the slope by ~1.5%. Bayesian linear models overestimate the intercept by ~3–4% and underestimate the slope by 3–4%. Parameter uncertainty is >1.4 times smaller in Bayesian models relative to any of the other models.

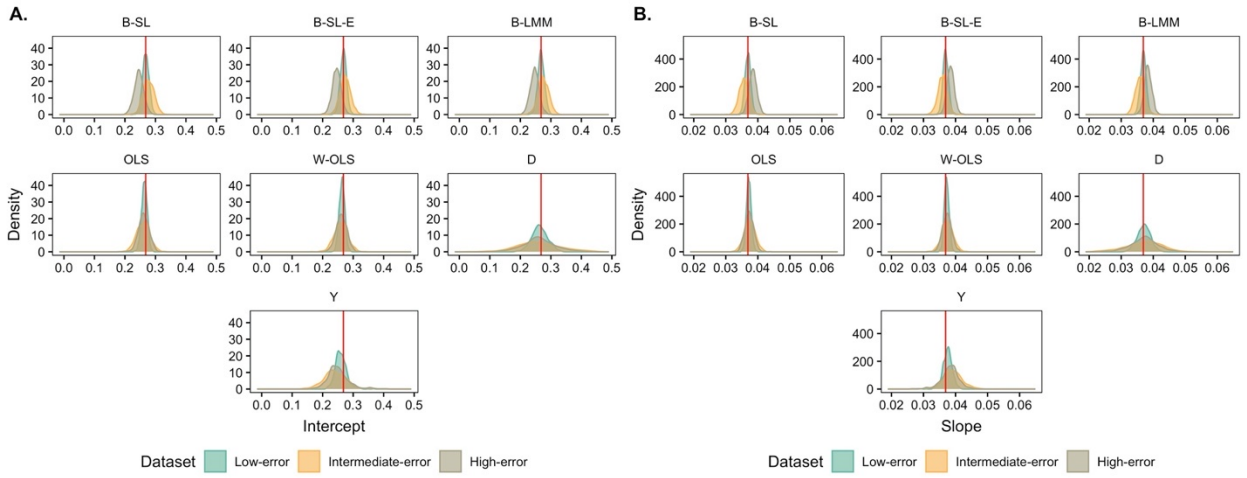

Fig. 2. Performance of different statistical models for deriving regression parameters for clumped isotope temperature calibrations, evaluated using a synthetic dataset with different levels of uncertainty in both Δ_{47} and T. We show the distribution of regression parameters (A-intercept; B-slope) based on re-sampling of the calibration dataset. Results are for seven models (panels), and different datasets (colors within panels) indicating alternative error scenarios, reflecting a low-error scenario (measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25 °C), an intermediate-error scenario (measurement error in Δ_{47}= 0.0075‰, instrument error Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2 °C), and a high-error scenario (measurement error in Δ_{47}= 0.0125‰, instrument error Δ_{47}= 0.0275‰, measurement error in $10^6/T^2$= 5 °C). Also shown are the true slope = 0.0369 and intercept = 0.268 (red vertical lines). In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

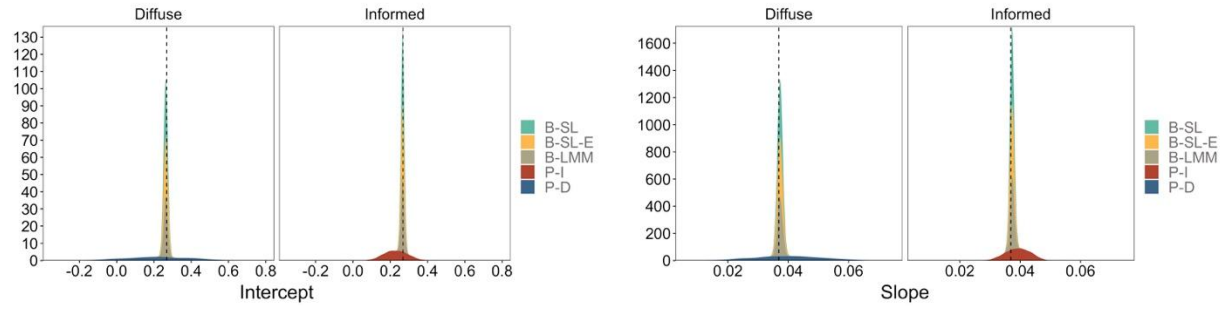

Fig. 3. Prior and posterior distributions for Bayesian analyses on the slope and intercept based on analyses run using informed and diffuse priors on regression parameters. Bayesian analyses yield posterior distributions that are robust. Results are for a low-error scenario (measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25 °C). The true slope = 0.0369 and intercept = 0.268, and the vertical dashed black line in each panel indicates the true parameter value. Results shown for datasets with 50 calibration samples. We show results for analyses run using the Bayesian simple linear model with errors, including both informed and diffuse priors on regression parameters. In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "P-I" for Prior – Informative, and "P-D" Prior–Diffuse.

3.1.2 Conclusions on model performance with synthetic datasets with 50 data points

In general, our analyses suggest that for calibration purposes (assuming 50-datapoint datasets) and across all error scenarios examined in this study, Bayesian and non-Bayesian linear models perform the best in terms of accuracy and precision. Non-Bayesian linear models generally outperform other models in terms of accuracy and Bayesian in terms of accuracy. We note that York and Deming regressions perform similarly across error scenarios, with these models over- or under-estimating the slope and intercept more strongly than any of the other models examined in this study. Therefore, our results suggest that Bayesian and non-Bayesian simple linear regression models should be used for calibrating the 'clumped isotopes' paleothermometer instead of York and Deming regression models.

We acknowledge that our results on the performance of the York model in the calibration part of our study might be unexpected for some readers. To our knowledge, only two studies have examined the performance of York regressions relative to any other model, within any context (clumped isotopes or otherwise). First, Wu and Yu (2018) compared the performance of multiple regression models, including York and ordinary least squares linear regressions. Using synthetic data with errors in both variables, these authors concluded that parameter estimates under York were less biased than those estimated under simple linear models. Their approach also involved examining the distribution parameter estimates under a given set of independent "runs" of each model. However, critical details on how the characteristics of each of these "runs" are missing from the study. For instance, it is unclear whether each of these "runs" was conducted on the complete dataset or a smaller set of observations. If analyses were run on subsampled datasets, the size of each smaller datasets is not provided. Similarly, we note that Wu

and Yu (2018) concluded that ordinary least squares models tend to fail to recover true parameters when the mean Y to X ratio is larger than 1. Therefore, our results suggesting that OLS models outperform York regression potentially reflect the fact that the mean Y to X ratio in clumped isotopes datasets is consistently well below 1. Second, results presented in Höhener and Imfeld (2021) show similar patterns to the ones reported in our study. For instance, Höhener and Imfeld (2021) indicate that while ordinary least squares linear models produce narrower error estimates, York regressions recover true regression parameters with a larger error. Nevertheless, we highlight that results in Höhener and Imfeld (2021) are also not generalizable to our study. For instance, the simulated datasets analyzed in Höhener and Imfeld (2021) assumed no error in variables. Finally, we note that York et al. (2004) does not provide a direct comparison of the performance of the York regression relative to other models.

3.2 Small and large calibration datasets: Δ_{47}-T calibration: Model performance with synthetic datasets

In addition to examining model performance on calibration datasets with intermediate sample size (n=50; Figs. 2, 3), we compare precision and accuracy between models in calibration datasets with a smaller (i.e., n=10 data points) and larger sample size (i.e., n=500).

Within our 10-datapoint datasets, Bayesian models show the worst performance across all the examined models regardless of the analyzed error scenario. For instance, under an intermediate error scenario, while the intercept was underestimated by ~20% in all the Bayesian models, non-Bayesian models underestimated the same parameter by only 2–7%. Among non-Bayesian models, Deming, simple, and weighted linear regressions show a similar performance, greater than the one for the York regression. Deming regressions outperform the simple and weighted linear models when error in the calibration dataset is high. Therefore, we recommend using (1) simple or weighted linear models when uncertainty in the calibration dataset is low, (2) Deming, simple, or weighted models when uncertainty is intermediate, and (3) a Deming model when uncertainty is high. In general, we recommend avoiding the use of Bayesian regression models when datasets are small (~10 observations).

Under 500-datapoint datasets, Bayesian regressions outperform any of the other models examined in this study in terms of both accuracy and precision. These results are independent of the analyzed scenario of error. For instance, under an intermediate scenario of errors, Bayesian models underestimate the true intercept by 0.5–0.7% and overestimate the slope by 0.3–0.5%. Non-Bayesian models underestimate the intercept by 2–5% and overestimate the slope by 1–3%. Therefore, Bayesian regressions are consistently a better choice for improving parameter estimation when the calibration dataset is large (n~500 datapoints). Therefore, as the size of datasets increase, the implementation of our Bayesian regressions will certainly help to improve the overall performance of reconstructions based on clumped isotope data.

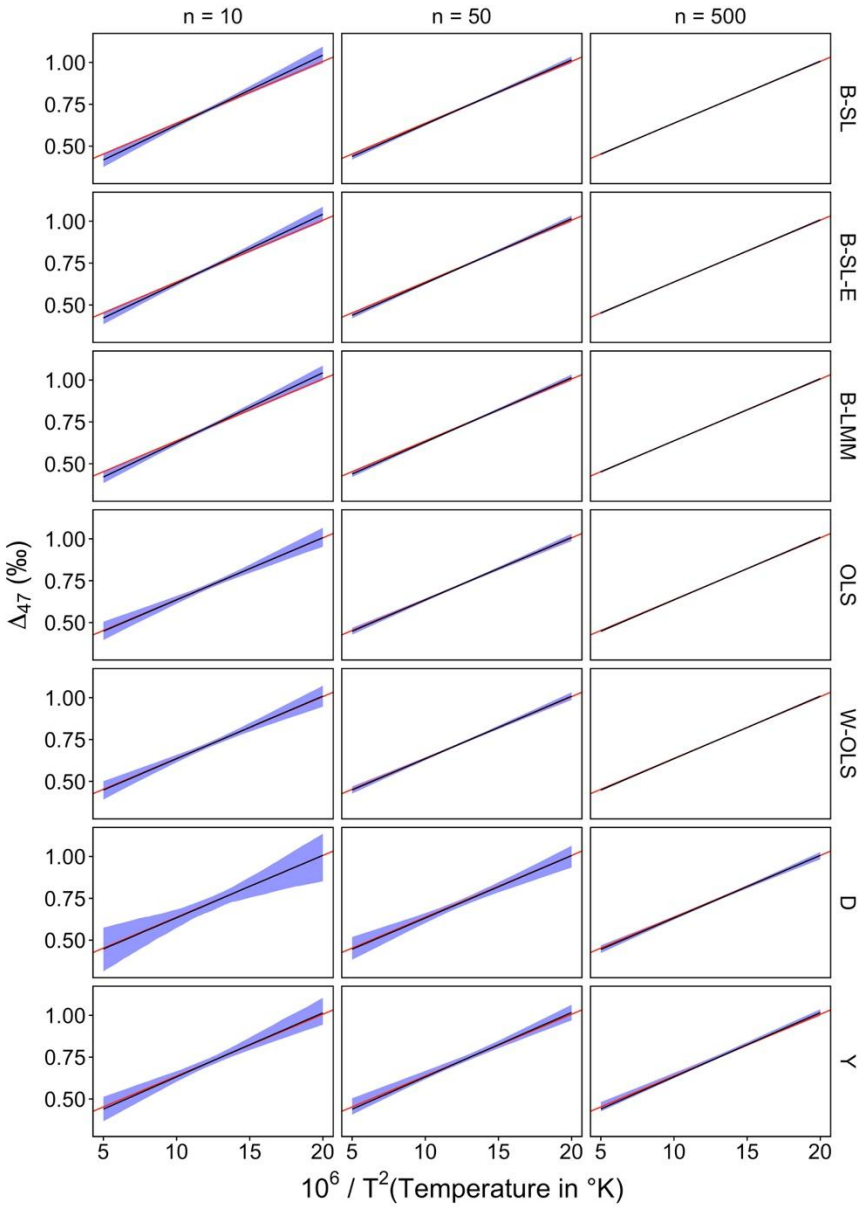

Fig. 4. Patterns of parameter uncertainty across regression models based on calibration datasets of variable size. Results are shown for an intermediate-error scenario (measurement error in Δ_{47}= 0.0075‰, instrument error Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2 °C). Patterns are largely similar to those for low- and high-error scenarios. We present mean parameter estimates for regression models (black line), the associated 95% CI (shaded blue region), and true model (red line). We fit regression models based on datasets with 10, 50, and 500 data points (columns) and using seven regression models (indicated by the column headings). As described in the text, for large calibration datasets (n > 50), Bayesian models typically yield the most accurate and precise regression parameters, while for n < 50, OLS is the best performing model. In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian

linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

3.3 General recommendations for selecting models to use for calibration in clumped isotope studies

We use the results presented in sections 3.1 and 3.2 to generalize major patterns of model performance between calibration datasets of different sizes and provide recommendations. Our analyses suggest that when the calibration dataset is small (n=10), Deming, weighted and unweighted OLS show the best accuracy. In datasets with intermediate sample size (n=50), weighted and unweighted OLS models recover the best accuracy. In large datasets (n=500), Bayesian regressions show the best performance in terms of precision and accuracy. Our analyses highlight the relative downsides of using York regressions in datasets of different sample size and with variable level of error. Instead, analyses conducted under OLS models (small and intermediate sample size) and Bayesian models (large sample size) should be prioritized for providing more reliable parameter estimates than other approaches. Due to the increasing trend in the number of observations in calibration datasets, our Bayesian framework provides a natural pathway to appropriately analyze large syntheses of clumped isotope data.

Finally, we also suggest that significant improvements on the performance of Bayesian models could be achieved in small datasets (e.g., 10 datapoints) by using maximum a posteriori estimation of regression parameters based on models fit in a larger number of replicates for the calibration dataset. In this study, parameter estimates under Bayesian models were based on fitting each model only once in each of the analyzed datasets. A more computationally intensive approach, fitting each model multiple times (e.g., 100–1,000) in each dataset and summarizing parameter estimates across the same number of posterior distributions, could have led to a better performance of Bayesian models in small datasets.

3.4 Inverting the forward model to predict $10^6/T^2$ from Δ_{47}: Temperature reconstructions for unknowns

We evaluate model performance for temperature reconstructions. We use regression parameters derived from calibration datasets with a total of 50 observations. The corresponding results, summarized in the sections presented below, are also included in the supplement. We present results in the main text that reflects patterns associated with an intermediate level of error for Δ_{47} (0.01‰).

Note that patterns of accuracy across models outlined in this section are expected to generally reflect the overall patterns of model performance during the calibration step (see Section 3.1). Given that Bayesian models were fit a single time in each of the datasets, performance for these models in this section likely reflect only an example of parameter estimates that are possible for 50-datapoint datasets. Due to the sample size dependency in accuracy that was outlined in Section 3.2, slightly different results can be obtained if

reconstructions are conducted using an alternative version of the 50-datapoint dataset (see distributions in Fig. 2).

3.4.1 Reconstructions for low-temperature carbonates (Δ_{47}= 0.8‰; T ~ 10 ºC)

Fig. 5 shows that for low-temperature carbonates, non-Bayesian models generally recover true temperature with better accuracy than Bayesian models. Under a low-error scenario, Deming, weighted and unweighted OLS overestimates true temperature by 0.5%, York by ~2%, and Bayesian models between 0.4% (no errors) and 1% (Bayesian linear mixed model). Under an intermediate-error scenario, Deming regression overestimates true temperature by 0.34%, weighted and unweighted OLS by 0.67%, York models by 4.5%, Bayesian simple linear models by 0.7%, and the Bayesian linear mixed model underestimates true temperature by 0.38%. Finally, under a high-error scenario, while non-Bayesian models overestimate temperature between 0.19% (weighted, unweighted OLS, and Deming regressions) and 8% (York model), Bayesian models underestimate temperature by 20%.

3.4.2 Reconstructions for intermediate-temperature carbonates (Δ_{47} = 0.7‰; T ~ 19 ºC)

Fig. 6 shows that for intermediate-temperature carbonates, Bayesian reconstructions only outperform any other approach when the error associated with the calibration dataset is low. However, non-Bayesian models are more appropriate when the error in the calibration dataset is intermediate or high. Under a low-error scenario, Bayesian models overestimate true temperature by 0.4–0.6% and non-Bayesian models underestimate temperature by ~1%. Under an intermediate error scenario, Bayesian models underestimate true temperature by ~6% and non-Bayesian models by ~2%. Finally, under a high-error scenario, Bayesian models underestimate true temperature by 7–8% and non-Bayesian models by 2–3%. In terms of precision, Bayesian reconstructions exhibit ~1.7 times less uncertainty in temperature relative to any of the alternative non-Bayesian models examined in this study.

3.4.3 Reconstructions for high-temperature carbonates (Δ_{47}= 0.6‰; T ~ 60 ºC)

Fig. 7 shows that for high-temperature carbonates, Bayesian reconstructions are more accurate when the uncertainty in the calibration dataset is low or high. However, non-Bayesian models show a better performance when the uncertainty in the calibration set is intermediate in value. Under a low-error scenario, Bayesian models overestimate true temperature by 0.2% and non-Bayesian models underestimate temperature between 1–2%. Under an intermediate error scenario, non-Bayesian models underestimate temperature between 1.8% (Deming regression) and 3.6% (York regression). Bayesian models underestimate temperature by ~6% under the same error scenario. Under a high-error scenario, Bayesian models underestimate temperature by 0.3–1.5% and non-Bayesian models between 2–5%. In terms of precision, Bayesian reconstructions exhibit 1.5–1.8 less uncertainty in temperature relative to any of the alternative non-Bayesian models examined in this study.

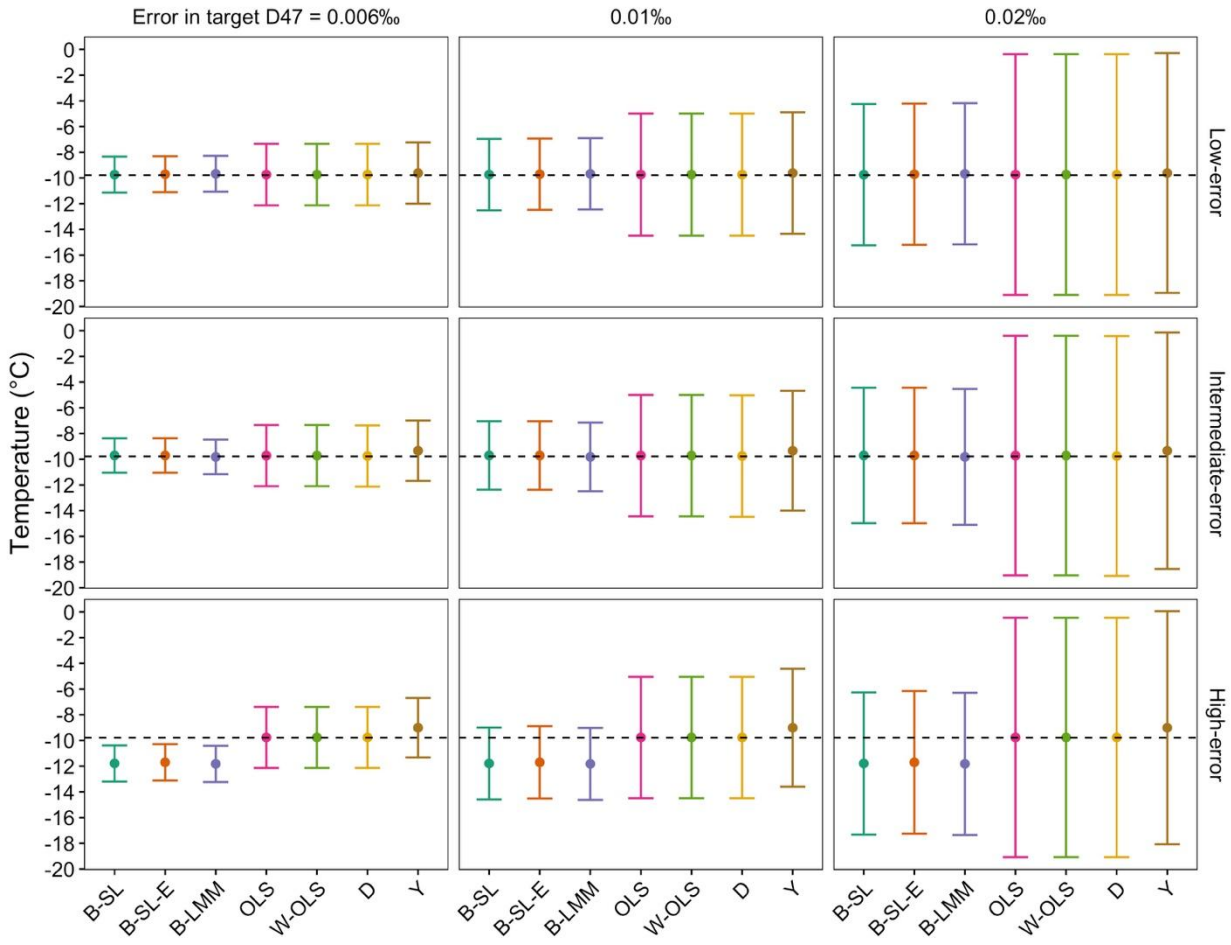

Fig. 5. Comparison of model performance in temperature reconstructions for low-temperature carbonates (Δ_{47} = 0.8‰). Mean reconstructed temperatures are shown in circles indicating accuracy. The precision of the reconstructions is also shown (standard error – error bars). We show results for Δ_{47} = 0.8‰, and multiple errors in Δ_{47} (corresponding to 0.005‰, 0.01‰, and 0.02‰). We also present results for a low-error scenario (measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C), intermediate-error scenario (measurement error in Δ_{47}= 0.0075‰, instrument error Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2°C), and high-error scenario (measurement error in Δ_{47}= 0.0125‰, instrument error Δ_{47}= 0.0275‰, measurement error in $10^6/T^2$= 5°C). In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

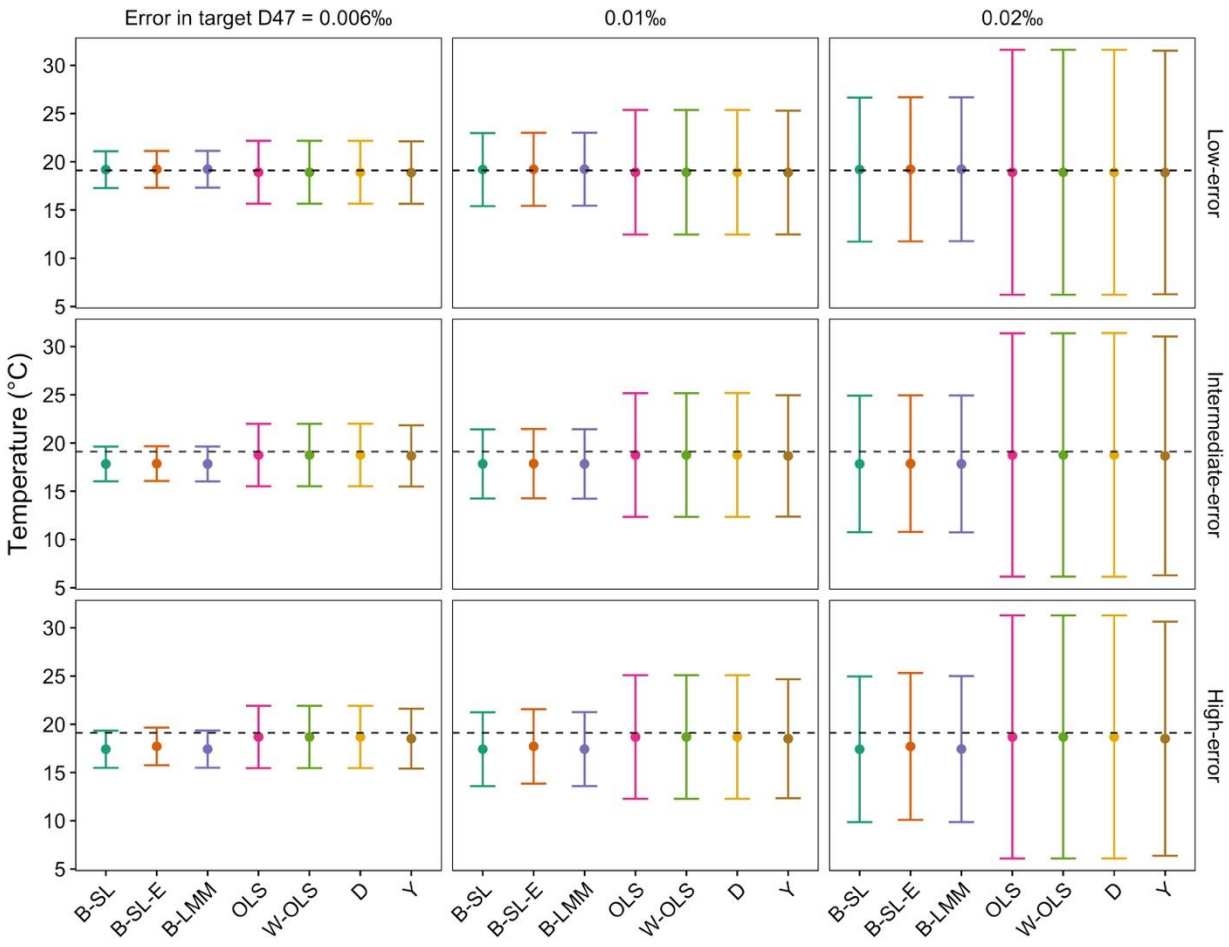

Fig. 6. Comparison of model performance in temperature reconstructions for intermediate-temperature carbonates (Δ_{47} = 0.7‰). Mean reconstructed temperatures are shown in circles indicating accuracy. The precision of the reconstructions is also shown (standard error – error bars). We show results for Δ_{47} = 0.7‰, and multiple errors in Δ_{47} (corresponding to 0.005‰, 0.01‰, and 0.02‰). We also present results for a low-error scenario (measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C), intermediate-error scenario (measurement error in Δ_{47}= 0.0075‰, instrument error Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2°C), and high-error scenario (measurement error in Δ_{47}= 0.0125‰, instrument error Δ_{47}= 0.0275‰, measurement error in $10^6/T^2$= 5°C). In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

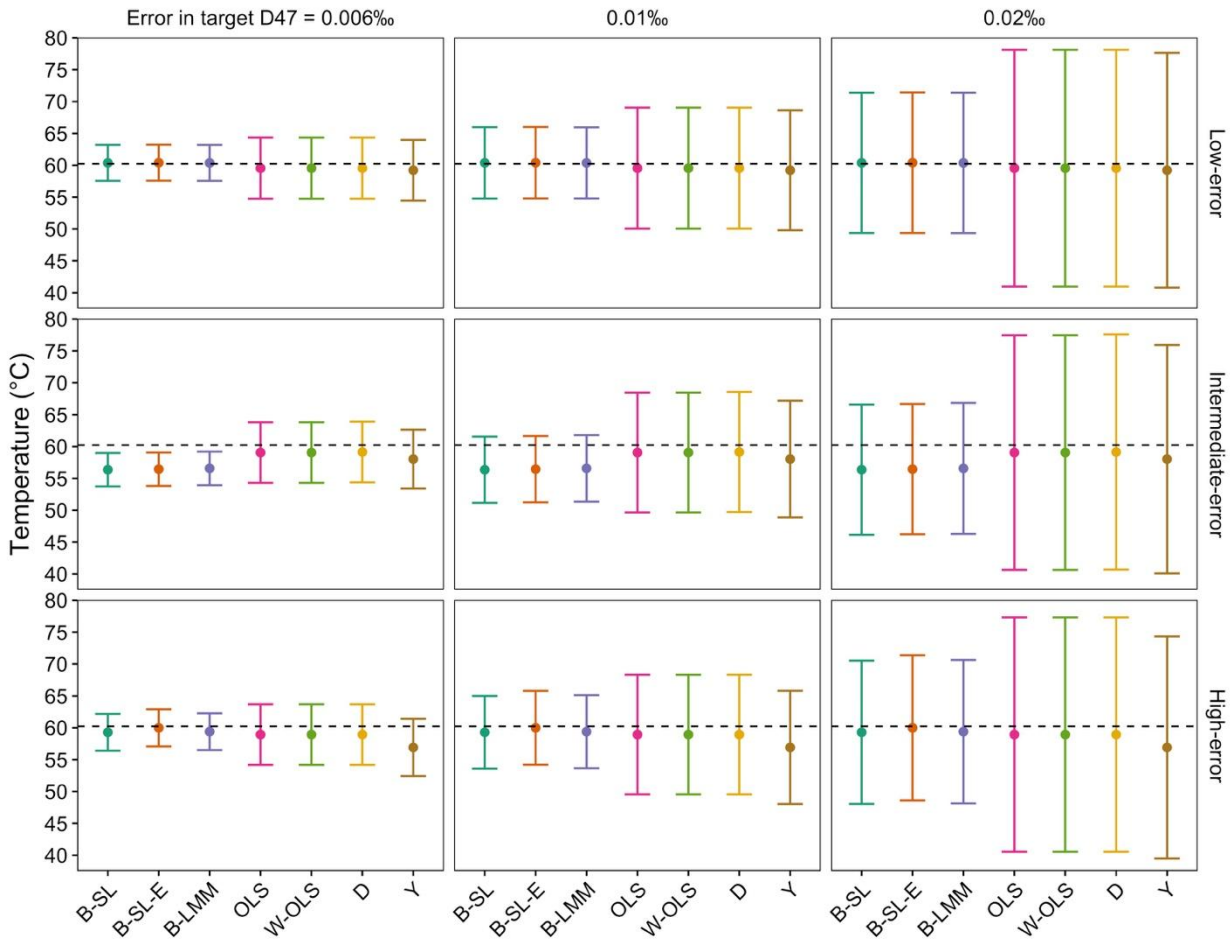

Fig. 7. Comparison of model performance in temperature reconstructions for high-temperature carbonates (Δ_{47} = 0.6‰). Mean reconstructed temperatures are shown in circles indicating accuracy. The precision of the reconstructions is also shown (standard error – error bars). We show results for Δ_{47} = 0.6‰, and multiple errors in Δ_{47} (corresponding to 0.005‰, 0.01‰, and 0.02‰). We also present results for a low-error scenario (measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C), intermediate-error scenario (measurement error in Δ_{47}= 0.0075‰, instrument error Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2°C), and high-error scenario (measurement error in Δ_{47}= 0.0125‰, instrument error Δ_{47}= 0.0275‰, measurement error in $10^6/T^2$= 5°C). In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

3.4.4 Recommendations

We make a series of recommendations for calibration datasets with n=50 or more observations. For low-temperature carbonates (Δ_{47} ~0.8‰), non-Bayesian linear models reconstruct temperatures with the highest precision (e.g., Deming and weighted/unweighted OLS) and Bayesian models with the highest accuracy. For intermediate-temperature carbonates (Δ_{47} ~ 0.7‰), Bayesian models outperform any other models when uncertainty in the calibration dataset is low. However, non-Bayesian models (e.g., York, Deming, weighted/unweighted) may be preferred when uncertainty in the calibration set is intermediate or larger. For high-temperature carbonates (Δ_{47} ~ 0.6‰), Bayesian models reconstruct temperatures more accurately when the uncertainty in the calibration dataset is at either extreme (small or high). We recommend using non-Bayesian regression models (with the exception of York regressions) regression models when uncertainty in the calibration set is intermediate.

3.5 BayClump: A Shiny app for paleothermometry using 'clumped isotopes'

To support the use of Bayesian models and the analytical framework developed in this study for clumped isotope calibration and for temperature reconstructions, and to facilitate comparisons of Bayesian and classical models, we present a self-contained R Shiny Dashboard application, BayClump (Fig. 8). BayClump fits both classical and Bayesian linear regressions to calibration datasets and performs temperature reconstructions. It uses most of the models described in this study in a graphical user interface (GUI) environment, without the need for expertise in R or programming of any kind. BayClump is open source and analyses are highly reproducible. BayClump is available as a web application at https://bayclump.tripatilab.epss.ucla.edu, as a local application for users familiar with R and RStudio (https://github.com/Tripati-Lab/BayClump), or as a standalone Electron desktop application which requires no additional software (the Electron application will be made available through Zenodo upon acceptance for publication), as freeware with the only requirements being citation of this study, and including an appropriate statement if the software or calibrations are modified.

BayClump currently includes two preloaded calibration datasets, compiled from Petersen et al. (2019) (Model 1) and Anderson et al. (2021) (Model 2), which will be updated as new calibration studies are published. Regression models can be developed using existing datasets or users can upload their own calibration data to work with by using a template available within the app. Users can also combine new calibration data with posted datasets to create a larger calibration set if desired. Any data that a user works with is not made available to others.

Based on current best practices (Bernasconi et al., 2021; Upadhyay et al., 2021), both Model 1 and Model 2 calibration sets are provided in the Intercarb Carbon Dioxide Equilibrium Scale (I-CDES). For this, we used an AFF to adjust values (i.e., 0.088 from Petersen et al. (2019)) so they are projected to the same acid digestion temperature (I-CDES is anchored to values that assume a reaction T of 90°C, not 25°C). Users may project their data into the I-

$CDES_{90}$ reference frame prior to adding into the template and uploading, for compatibility with default datasets, or they can exclusively use their own calibration data in any reference frame. Calibration models may be selected independently of one another, and options are available to scale data if needed. BayClump provides the ability for the user to download full calibration regression model output and any of the associated calibration regression model plots.

We also provide a GUI in BayClump for reconstructing temperature using both Bayesian and non-Bayesian models. A separate template in comma separated value (.csv) format is provided where users can add a table of sample Δ_{47} values and the combined error from measurement and standardization, and then download calculated temperatures in an Excel file. Currently, users can implement the Bayesian linear regression model with errors, utilize a Bayesian framework for estimating temperature that intrinsically accounts for both uncertainty in parameter estimates and error in target Δ_{47}.

Alternatively, BayClump users can transfer over a distribution of Bayesian or non-Bayesian regression parameters derived from their own datasets from the Calibration tab to use for temperature reconstruction in the Reconstruction tab, either within or outside of a Bayesian framework. For non-Bayesian temperature estimates, reconstructed values of temperature will be shown for each of the selected models (in the calibration tab) when (1) parameter error is ignored in temperature reconstructions, or (2) when parameter error is accounted for in the reconstruction step.

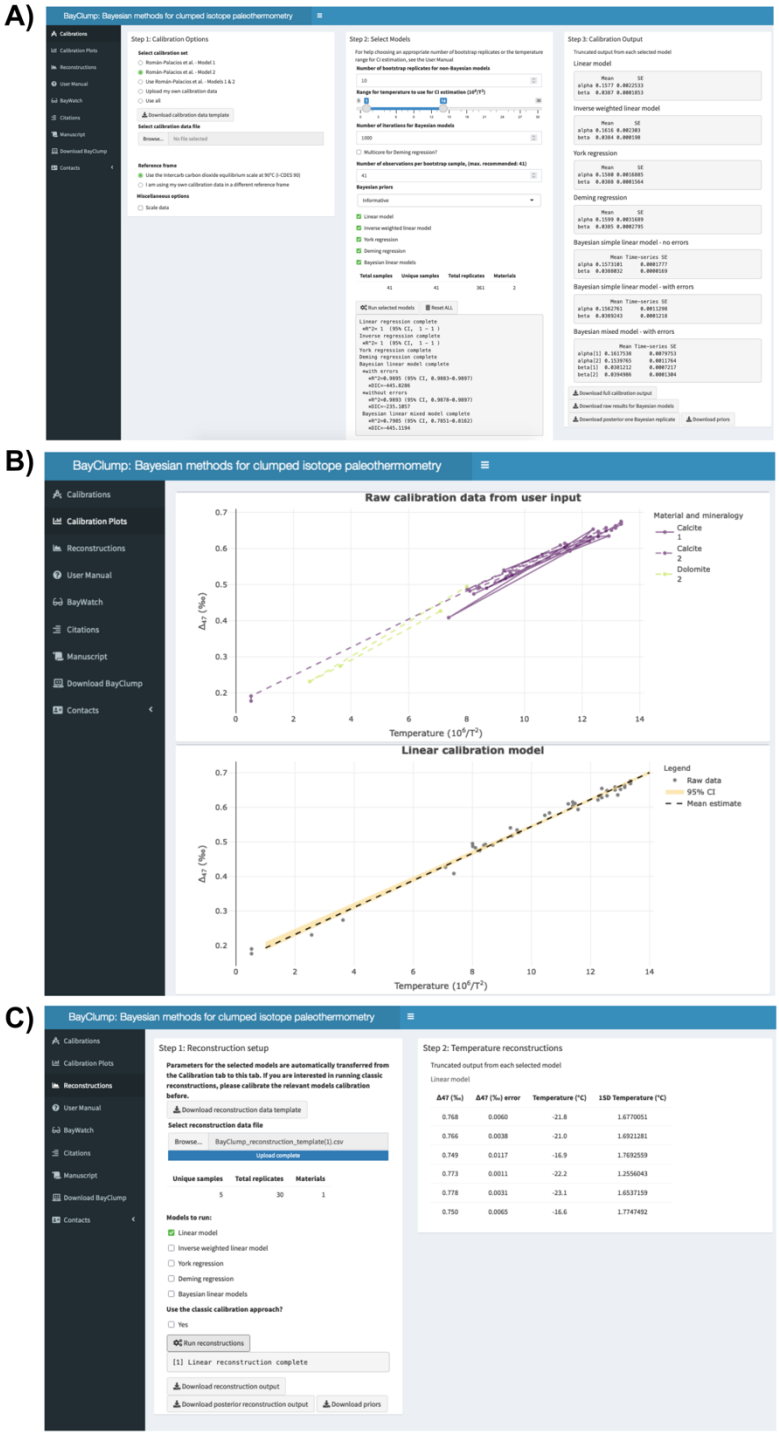

Fig. 8. BayClump application screenshots for proxy calibration and for temperature reconstruction. A) Calibration options are chosen in the first tab of the application. Multiple preloaded datasets are included, or the user may opt to upload their own calibration data.

Summary information is provided for each selected regression model upon run completion. **B)**: Plots of calibration data and each selected regression model are provided in the 'Calibration Plots' tab. Plots are fully interactive and downloadable. **C)** The application automatically transfers calibration models and data to the 'Reconstructions' tab. Users upload their own Δ_{47} data for temperature reconstructions using the provided template. Summary information is provided at run completion (output shown above is truncated). Both 'Calibration' and 'Reconstruction' tabs provide buttons to download full model output in tabbed and labeled Excel spreadsheets.

3.5.1 Reanalysis of the Petersen et al. (2019) dataset using BayClump

In addition to providing a general summary of model performance using synthetic datasets, we utilize BayClump to estimate regression parameters for a published synthesis of calibration data that contained results for 451 samples measured in several different laboratories on the Carbon Dioxide Equilibrium Scale reference frame (Petersen et al. 2019). In that study, the authors provided estimates of the slope and intercept using a Monte Carlo least squares regression based on 10,000 replicates (Table 3). Here, we analyze the same dataset using all seven regression models analyzed in this study (Fig. 9). We perform a total of 1,000 replicates of each non-Bayesian model implemented in BayClump. Bayesian models were analyzed using 50,000 iterations (50% burnin). Note that given our main focus was on re-examining the main equation in Petersen et al. (2019), we do not fit material-specific regression models (e.g., using B-LMM). The authors of the same study extensively discuss material specificity in the same article. Similarly, we focus on deriving a calibration for the full temperature range of the dataset (also largely the main goal of their study).

Fig. 9 compares regression coefficients and their associated uncertainties. Except for York models, Bayesian and non-Bayesian linear models differed from the slope estimated in Petersen et al. (2019) by less than 1.5%. York recovered a slope estimate that was ~17% larger than the one in Petersen et al. (2019; slope in York=0.0397 vs Petersen 0.0383). In terms of the intercept, Deming and York models recovered parameter estimates that were between 5% (Deming) and 26% smaller than the ones in Petersen et al. (2019; intercept Deming=0.245 vs York=0.191 vs Petersen = 0.258). Based on our analysis of synthetic datasets, the divergence of York and Deming models from the rest of the regression models is expected. These two models generally show a poor performance and under- or over-estimate calibration regression parameters relative to Bayesian and OLS models.

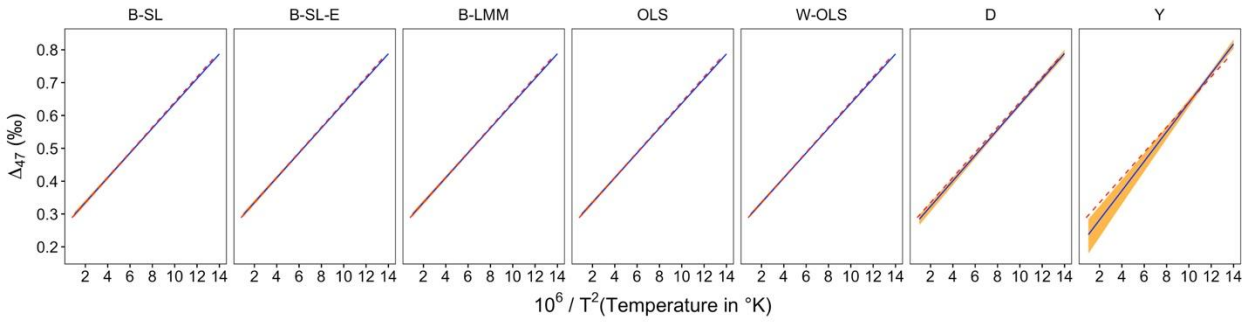

Fig. 9. Reanalysis of the Petersen et al. (2019) dataset using BayClump. We present mean parameter estimates for each of the seven regression models implemented in BayClump (blue line), the associated 95% CI (shaded orange region), and the regression reported by Petersen et al. (2009) (red dashed line; 95% CI in grey). In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

3.5.2 Reanalysis of the Anderson et al. (2021) dataset using BayClump

Using BayClump, we also reanalyze the Anderson et al. (2021) dataset (Fig. 10) using the same methodology as described in this study in Section 3.5.1. Anderson et al. (2021) provided parameter estimates based on York regression models. Instead of using a single version of the full dataset as in Petersen et al. (2019), we use two versions of the Anderson et al. (2021) calibration dataset. First, we fit regressions to the whole dataset, over the full range of temperature values. Second, we perform analyses that include only samples in the environmental temperature range (0–35 ºC; Table 3). Specifically, given that the simulated datasets closely reflect an environmental temperature range, we subsampled the Anderson et al. (2021) dataset for being able to use our results in the synthetic component of this paper as a reference. Note that patterns of association between $10^6/T^2$ from Δ_{47} in a limited temperature range (~0–35 ºC) we examined in Anderson et al. (2019) but not in Petersen et al. (2019). Similarly, we examined material-specific regression parameters (e.g., B-LMM) given that this aspect was not extensively discussed in Anderson et al. (2021).

Note that the uncertainty in the full and subsampled Anderson et al. (2021) dataset with a mean temperature error ~0.5 °C and mean error in Δ_{47}=0.012‰ (0.040‰ in the subsampled), roughly corresponds to the error structure of our low-error scenario. When using the entire Anderson et al. (2021) dataset, our parameter estimates for the based on the York model fit in BayClump differ by -0.5% (slope) and 1.1% (intercept) from parameter estimates in Anderson et al. (2021; Table 3). Our Bayesian simple linear models recover estimates of the slope that are -0.5–1% different than the one in Anderson et al. (2021) and intercept values that are 0.04–1.27% larger than the published ones. The rest of the models differed from parameter estimates in Anderson et al. (2021) by less than 3%. The most important difference is not actually between models but related to the analyzed partition of the dataset. Relative to the original Anderson et al. (2021) calibration for their full dataset, all of our models yield different regression parameters

(15–29% smaller slope estimates and 27–55% larger intercept values relative to the published estimates) when the reduced Anderson et al. (2021) dataset, that includes only samples in the environmental temperature range, is used. Finally, we note that our Bayesian linear mixed models did detect differences in parameter estimates (full Anderson dataset: mostly related to the slope; ATM Anderson dataset: related to both the slope and intercept; Table 3) between calcite and dolomite samples in Anderson et al. (2021).

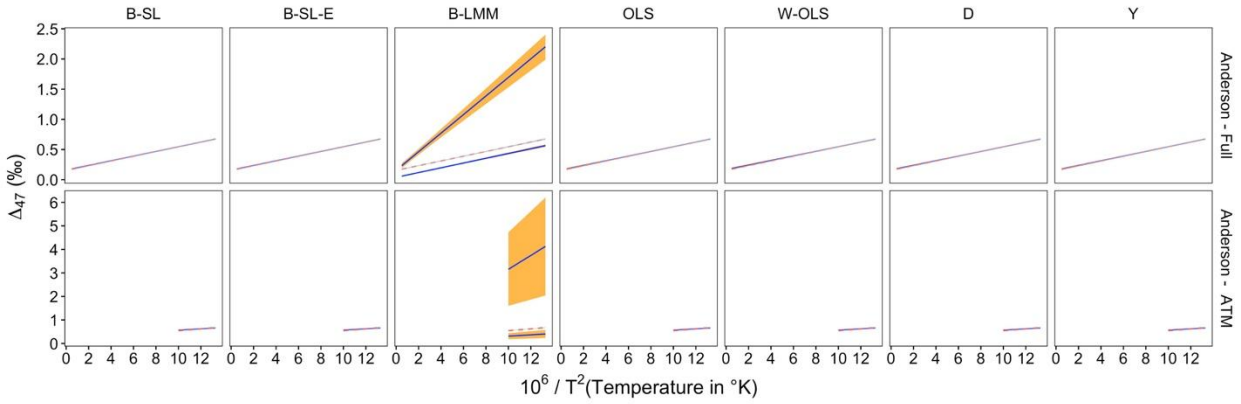

Fig. 10. Reanalysis of the Anderson et al. (2021) dataset synthesis using BayClump. We present mean parameter estimates for each of the seven regression models implemented in BayClump (blue line), the associated 95% CI (shaded orange region), and the published syntheses (red dashed line; 95% CI in grey). Results are shown for Anderson et al. (2021) synthesis for all samples (top row), and Anderson et al. (2021) for samples in an environmental temperature range (0–35 ºC; bottom row). Note that Bayesian linear mixed models estimate material-specific regression models (two lines and associated uncertainties in plot). However, parameter estimates for each material under might have suffered from further partitioning the already small dataset in Anderson et al. (2021). In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

3.5.3 Reanalysis of the Sun et al. (2021) dataset using BayClump

We recalculate temperatures using data from Sun et al. (2021) for Late Miocene and recent carbonates from the Shuitangba hominoids site in Yunnan, China with BayClump (Fig. 11). This study represents a relevant study case where the main focus is directly related to addressing whether temperatures between two sites differed at different times (~0 and 6.2 Ma). For this reanalysis, we take regression parameters for the different models from Table 3 that are calculated using the three datasets described above (dataset 1 - the Petersen et al. (2019) calibration dataset from section 3.5.1; dataset 2 – the full sample set from the Anderson et al. (2021) calibration from section 3.5.2; dataset 3 – a restricted subset of the Anderson et al. (2021)

calibration that uses their samples from environmental temperatures (0–35 °C) from section 3.5.2). Next, we perform temperature reconstructions that account for both regression parameter uncertainty (Table 3) and error in Δ_{47} (Sun et al., 2021; Table 4). Note that the uncertainty in Petersen et al. (2019) (mean temperature error=1.49 °C, and mean error in Δ_{47}=0.010‰), roughly corresponds to the error structure of our intermediate-error scenario (measurement error in Δ_{47}= 0.0075‰, error in true Δ_{47}= 0.0225‰, measurement error in temperature=2 °C). The mean Δ_{47} in Sun et al. (2021) is similar to our intermediate-temperature carbonates scenario (~0.7‰), and most of their samples have an intermediate level of error (~0.01‰). Given our results based on synthetic datasets (section 3.4.2; Table 3), we expect that Bayesian models should be the most accurate and precise among the examined models. We also compare reconstructed temperatures to the results reported in Sun et al. (2021) that utilized published calibrations (Henkes et al., 2013; Tripati et al., 2015).

Overall, our mean estimates of temperature based on Bayesian regression models suggest that there are not major differences in reconstructed temperatures between sites (Table 3). Temperature estimates are, however, very similar to those reported by Sun et al. (2021) based on calibrations from Tripati et al. (2015) and Henkes et al. (2013) (see also Table 4). Our Bayesian framework generally yields temperature values with much smaller uncertainties, depending on the analyzed dataset.

In general, our results based on the utilization of Bayesian and non-Bayesian regression models and a Bayesian framework for reconstructing temperatures support one of the main conclusions in Sun et al. (2021). Specifically, the authors and our study suggest similar temperatures in Shuitangba during the Late Miocene and the present-day temperatures in the Fuxian Lake area (both sites, ~17 ºC). Results in Sun et al. (2021) noted that uncertainties in reconstructed temperatures were larger than the mean temperature difference between areas. Our Bayesian reconstructions were able to shrink uncertainty around median temperature values while also concluding that temperatures are similar between sites.

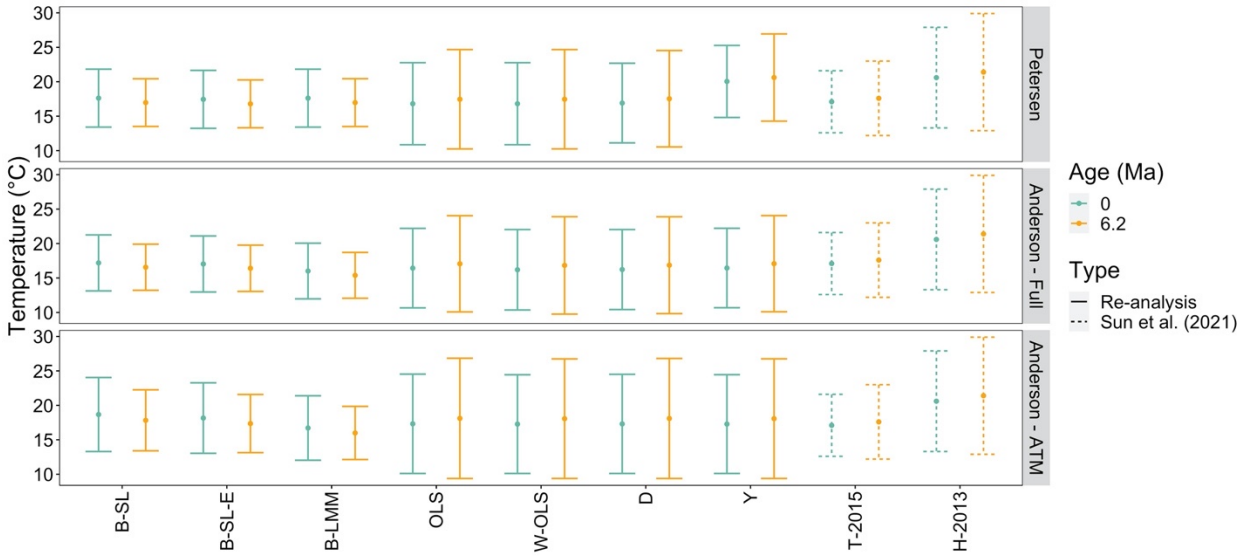

Fig. 11. Temperature reconstructions presented in Sun et al. (2021) for a hominoid locality in Yunnan, China compared to those derived using BayClump. We present mean and associated error for reconstructed temperatures from Sun et al. (2021) based on calibrations by Henkes et al. (2013; "H-2013") and Tripati et al. (2015; "T-2015"). Note that "H-2013" and "T-2015" are the same across rows in the figure (dotted lines). These reconstructions are used in each panel as a reference to the new calibrations performed in BayClump (solid lines). Our analyses in BayClump (i.e., solid lines) are shown for calibrations performed on the Petersen et al. (2019; top row), Anderson et al. (2021)'s full calibration dataset (second row), and Anderson et al. (2021)'s dataset reduced to environmental samples (bottom row). In each of these datasets, we fit seven regression models and reconstructed temperatures for the two sited in Sun et al. (2021). In the figure, "B-SL" stands for Bayesian simple linear model without errors, "B-SL-E" for Bayesian simple linear model with errors, "B-LMM" for Bayesian linear mixed model, "OLS" for ordinary least squares regression, "W-OLS" for weighted ordinary least squares regression, "D" for Deming regression, and "Y" for York model.

4 Conclusions

We examine the performance of seven regression models for calibrating the clumped isotope thermometer, including the first Bayesian implementations of regression models. We implement a Bayesian linear mixed model that can accommodate differences in regression parameters between groups, so that a range of materials can have different slopes and/or intercepts. Using simulated calibration datasets with variable number of observations (from 10 to 500 samples) and degrees of error in clumped isotope measurements and temperature, we find that Bayesian and non-Bayesian ordinary least squares linear models consistently outperform other regression models in terms of accuracy and precision under most synthetic scenarios when reconstructing true regression parameters. The performance of Bayesian linear models strongly improves when the number of observations in the calibration dataset exceeds 50 data points.

We also utilized different frameworks for reconstructing temperatures and found differences in temperature reconstruction performance between regression models. In general, Bayesian reconstructions were more precise and accurate than non-Bayesian reconstructions when error in the examined Δ_{47} was small (<0.01‰). Non-Bayesian reconstructions using Bayesian model-derived regression parameters were generally more robust than other approaches, and accurately recovered temperatures in a range of scenarios. Based on our analyses, we summarized the models that showed the best performance during the calibration and reconstruction phase. A Bayesian regression model when applied to published calibration syntheses yields reduced uncertainties. Some of the differences between temperature reconstructions based on published syntheses may originate from material-specific patterns in the calibration datasets, which should be investigated in future work.

The analytical tools developed in this study are available in BayClump, a Shiny dashboard with data templates that facilitates the use of Bayesian methods for both calibration and temperature reconstruction. Application to published clumped isotope data from Late Miocene and recent samples from the Shuitangba hominoids site in Fuxian Lake (Yunnan, China) show the potential of BayClump as a tool for resolving relatively small temperature

changes with confidence in paleoclimatology and support the published conclusions. We expect the tools developed in this study to provide a basis for robustly choosing to use particular regression models for temperature calibration and reconstruction in clumped isotopes, and to reduce the uncertainty of temperature reconstructions.

Author Contribution Statement

A. Tripati initiated the project, developed the project team, mentored the early career researchers, and funded the work. A. Tripati and C. Román-Palacios designed the project. C. Román-Palacios developed the Bayesian regression models and wrote the manuscript, with input from A. Tripati and all co-authors. H. Carroll developed BayClump with input from C. Román-Palacios and A. Tripati, A. J. Arnold and R. J. Flores, and the Tripati Lab Group. A. J. Arnold and R. J. Flores provided input on datasets and parameters for analysis. K. McKinnon provided input on analysis. Q. Gan helped revise the datasets analyzing in this study and write the stan version of our Bayesian regressions. All authors provided edits to the manuscript.

Acknowledgments

This work benefited from discussions with the Tripati Lab Group including R. Ulrich and D. Brown. It also was improved through discussions and edits from R. Eagle and N. Kraft, a review from Jessica Tierney, and editorial comments. We are grateful to Rod O'Connor for application hosting and support. C. Román-Palacios, H. M. Carroll, A. J. Arnold, R. J. Flores, and A. Tripati were supported by grants to A. Tripati from DOE BES DE-SC0010288, NSF EAR-1936715, the Heising-Simons Foundation, and by the Center for Diverse Leadership in Science which is supported by the Packard Foundation and the Silicon Valley Community Foundation. H. M. Carroll was also supported through a postdoctoral fellowship by the Institutional Research and Academic Career Development Awards (IRACDA) program at UCLA (Award # K12 GM106996).

Data Availability Statement

BayClump is available on GitHub (https://github.com/Tripati-Lab/BayClump) and deployed on the following link https://bayclump.tripatilab.epss.ucla.edu/. All data and code used in this paper is available on GitHub (https://github.com/Tripati-Lab/BayesPaper).

References

Anderson, N. T., Kelson, J. R., Kele, S., Daëron, M., Bonifacie, M., Horita, J., ... & Bergmann, K. D. (2021). A unified clumped isotope thermometer calibration (0.5–1,100 C) using carbonate-based standardization. *Geophysical Research Letters*, *48*(7), e2020GL092069.

Bernasconi, S., Daëron, M., Bergmann, K. D., Bonifacie, M. and A. N. Meckler (2021), InterCarb: A community effort to improve inter-laboratory standardization of the carbonate clumped isotope thermometer using carbonate standards, *Geochem. Geophys.*, 22(5), e2020GC009588, doi:10.1002/essoar.10504430.3.

Bonifacie, M., Calmels, D., Eiler, J. M., Horita, J., Chaduteau, C., Vasconcelos, C., ... & Bourrand, J. J. (2017). Calibration of the dolomite clumped isotope thermometer from 25 to 350 C, and implications for a universal calibration for all (Ca, Mg, Fe) CO3 carbonates. *Geochimica et Cosmochimica Acta*, *200*, 255-279.

Brandon M. Greenwell and Christine M. Schubert Kabban (2014). investr: An R Package for Inverse Estimation. The R Journal, 6(1), 90-100. URL http://journal.r-project.org/archive/2014-1/greenwell-kabban.pdf.

Came, R. E., Eiler, J. M., Veizer, J., Azmy, K., Brand, U. and C. R. Weidman (2007), Coupling of surface temperatures and atmospheric CO2 concentrations during the Palaeozoic era, *Nature*, 449(7159), 198–201, doi:10.1038/nature06085.

Crampton-Flood, E. D., Tierney, J. E., Peterse, F., Kirkels, F. M., & Damsté, J. S. S. (2020). BayMBT: A Bayesian calibration model for branched glycerol dialkyl glycerol tetraethers in soils and peats. *Geochimica et Cosmochimica Acta*, *268*, 142-159.

Daëron, M.: Full propagation of analytical uncertainties in Δ_{47} measurements, *Geochem. Geophys.*, *22*(5), e2020GC009592, doi:10.1002/essoar.10505298.1, 2020.

Deming, W. E. (1964), Statistical adjustment of data, Dover publications.

Eagle, R. A., Eiler, J. M., Tripati, A. K., Ries, J. B., Freitas, P. S., Hiebenthal, C., ... & Roy, K. (2013). The influence of temperature and seawater carbonate saturation state on 13 C–18 O bond ordering in bivalve mollusks. *Biogeosciences*, *10*(7), 4591-4606.

Eagle, R. A., Enriquez, M., Grellet-Tinner, G., Pérez-Huerta, A., Hu, D., Tütken, T., ... & Eiler, J. M. (2015). Isotopic ordering in eggshells reflects body temperatures and suggests differing thermophysiology in two Cretaceous dinosaurs. *Nature communications*, *6*(1), 1-11.

Eagle, R. A., Schauble, E. A., Tripati, A. K., Tutken, T., Hulbert, R. C. and J. M. Eiler (2010), Body temperatures of modern and extinct vertebrates from 13C-18O bond abundances in bioapatite, *Proc. Natl. Acad. Sci. U.S.A.*, 107(23), 10377–10382, doi:10.1073/pnas.0911115107.

Eagle, R. A., Tütken, T., Martin, T. S., Tripati, A. K., Fricke, H. C., Connely, M., ... & Eiler, J. M. (2011). Dinosaur body temperatures determined from isotopic (13C-18O) ordering in fossil biominerals. *science*, *333*(6041), 443-445.

Eiler, J. M., & Schauble, E. (2004). 18O13C16O in Earth's atmosphere. *Geochimica et Cosmochimica Acta*, *68*(23), 4767-4777.

Eiler, J. M. (2007), On the Origins of Granites, *Science*, 315(5814), 951–952, doi:10.1126/science.1138065.

Eiler, J. M. (2011), Paleoclimate reconstruction using carbonate clumped isotope thermometry. *Quat. Sci. Rev.*, 30(25-26), 3575-3588.

Garzione, C. N., Auerbach, D. J., Smith, J. J.-S., Rosario, J. J., Passey, B. H., Jordan, T. E., Eiler, J. M. (2014), Clumped isotope evidence for diachronous surface cooling of the Altiplano and pulsed surface uplift of the Central Andes. *Earth Planet. Sci. Lett.,* 393, 173–181.

Ghosh, P., Adkins, J., Affek, H., Balta, B., Guo, W., Schauble, E.A., Schrag, D. and Eiler, J.M. (2006), 13C–18O bonds in carbonate minerals: a new kind of paleothermometer. *Geochimica et Cosmochimica Acta*, *70*(6), pp.1439-1456.

Henkes, G. A., Passey, B. H., Wanamaker, A. D., Grossman, E. L., Ambrose, W. G. and M. L. Carroll (2013), Carbonate clumped isotope compositions of modern marine mollusk and brachiopod shells, *Geochim. Cosmochim. Acta*, 106, 307–325, doi:10.1016/j.gca.2012.12.020.

Hilbe, J. M., De Souza, R. S., and E. E. Ishida (2017), *Bayesian models for astrophysical data: using R, JAGS, Python, and Stan*. Cambridge University Press.

Hill, P. S., Schauble, E. A., & Tripati, A. (2020). Theoretical constraints on the effects of added cations on clumped, oxygen, and carbon isotope signatures of dissolved inorganic carbon species and minerals. *Geochimica et Cosmochimica Acta*, *269*, 496-539.

Hill, P. S., Tripati, A. K., & Schauble, E. A. (2014). Theoretical constraints on the effects of pH, salinity, and temperature on clumped isotope signatures of dissolved inorganic carbon species and precipitating carbonate minerals. *Geochimica et cosmochimica acta*, *125*, 610-652.

Höhener, P., and G. Imfeld (2021), Quantification of Lambda (Λ) in multi-elemental compound-specific isotope analysis. *Chemosphere*, 267, 129232.

Huntington, K. W., Wernicke, B. P., & Eiler, J. M. (2010). Influence of climate change and uplift on Colorado Plateau paleotemperatures from carbonate clumped isotope thermometry. *Tectonics*, *29*(3).

Kelson, J. R., Huntington, K. W., Breecker, D. O., Burgener, L. K., Gallagher, T. M., Hoke, G. D., and S. V. Petersen (2020), A proxy for all seasons? A synthesis of clumped isotope data from Holocene soil carbonates. Quat. Sci. Rev., 234, 106259.

Kelson, J. R., Huntington, K. W., Schauer, A. J., Saenger, C., & Lechler, A. R. (2017). Toward a universal carbonate clumped isotope calibration: Diverse synthesis and preparatory methods suggest a single temperature relationship. *Geochimica et Cosmochimica Acta*, *197*, 104-131.

Khider, D., Huerta, G., Jackson, C., Stott, L. D. and J. Emile-Geay (2015), A Bayesian, multivariate calibration for *Globigerinoides ruber* Mg/Ca, *Geochem. Geophys.*, 16(9), 2916–2932, doi:10.1002/2015gc005844.

Leutert, T. J., Sexton, P. F., Tripati, A., Piasecki, A., Ho, S. L., & Meckler, A. N. (2019). Sensitivity of clumped isotope temperatures in fossil benthic and planktic foraminifera to diagenetic alteration. *Geochimica et Cosmochimica Acta*, *257*, 354-372.

Li, H., Liu, X., Arnold, A., Elliott, B., Flores, R., Kelley, A. M., & Tripati, A. (2021). Mass 47 clumped isotope signatures in modern lacustrine authigenic carbonates in Western China and other regions and implications for paleotemperature and paleoelevation reconstructions. *Earth and Planetary Science Letters*, *562*, 116840.

Martin, R. F. (2000) General Deming Regression for Estimating Systematic Bias and Its Confidence Interval in Method-Comparison Studies, *Clin. Chem.*, 46(1), 100–104, doi:10.1093/clinchem/46.1.100.

Martínez-Sosa, P., Tierney, J. E., Stefanescu, I. C., Crampton-Flood, E. D., Shuman, B. N., & Routson, C. (2021). A global Bayesian temperature calibration for lacustrine brGDGTs. *Geochimica et Cosmochimica Acta*, *305*, 87-105.

Meinicke, N., Ho, S., Hannisdal, B., Nürnberg, D., Tripati, A., Schiebel, R. and A. Meckler (2020), A robust calibration of the clumped isotopes to temperature relationship for foraminifers, *Geochim. Cosmochim. Acta*, 270, 160–183, doi:10.1016/j.gca.2019.11.022.

Passey, Q. R., Bohacs, K. M., Esch, W. L., Klimentidis, R. and S. Sinha (2010), From Oil-Prone Source Rock to Gas-Producing Shale Reservoir – Geologic and Petrophysical Characterization of Unconventional Shale-Gas Reservoirs, doi:10.2118/131350-ms.

Peral, M., Daëron, M., Blamart, D., Bassinot, F., Dewilde, F., Smialkowski, N., Isguder, G., Bonnin, J., Jorissen, F., Kissel, C., Michel, E., Riveiros, N. V. and C. Waelbroeck (2018), Updated calibration of the clumped isotope thermometer in planktonic and benthic foraminifera, *Geochim. Cosmochim. Acta*, 239, 1–16, doi:10.1016/j.gca.2018.07.016.

Pérez-Escobar, O. A., Gottschling, M., Chomicki, G., Condamine, F. L., Klitgård, B. B., Pansarin, E., and G. Gerlach (2017), Andean mountain building did not preclude dispersal of lowland epiphytic orchids in the Neotropics, *Sci. Rep.*, 7, 1–10.

Petersen, S. V., Defliese, W. F., Saenger, C., Daëron, M., Huntington, K. W., John, et al. (2019), Effects of Improved 17O Correction on Interlaboratory Agreement in Clumped Isotope Calibrations, Estimates of Mineral-Specific Offsets, and Temperature Dependence of Acid Digestion Fractionation, *Geochem. Geophys.*, 20(7), 3495–3519, doi:10.1029/2018gc008127.

Plummer, M. (2003), JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing, 124(125), 1–10.

R Core Team (2021), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Schauble, E. A., Ghosh, P. and J. M. Eiler (2006), Preferential formation of 13C–18O bonds in carbonate minerals, estimated using first-principles lattice dynamics, *Geochim. Cosmochim. Acta*, 70(10), 2510–2529, doi:10.1016/j.gca.2006.02.011.

Sun, F., Wang, Y., Jablonski, N. G., Hou, S., Ji, X., Wolff, B., ... & Yang, X. (2021). Paleoenvironment of the late Miocene Shuitangba hominoids from Yunnan, Southwest China: Insights from stable isotopes. *Chemical Geology*, *569*, 120123.

Therneau, T. (2018), deming: Deming, Theil-Sen, Passing-Bablock and Total Least Squares Regression. R package version 1.4. https://CRAN.R-project.org/package=deming

Tierney, J. E., Malevich, S. B., Gray, W., Vetter, L., & Thirumalai, K. (2019). Bayesian calibration of the Mg/Ca paleothermometer in planktic foraminifera. *Paleoceanography and Paleoclimatology*, *34*(12), 2005-2030.

Tierney, J. E. and M. P. Tingley (2014), A Bayesian, spatially-varying calibration model for the TEX86 proxy, *Geochim. Cosmochim. Acta*, 127, 83–106, doi:10.1016/j.gca.2013.11.026.

Tierney, J. E. and M. P. Tingley (2015), A TEX86 surface sediment database and extended Bayesian calibration, *Sci. Data*, 2, doi:10.1038/sdata.2015.29.

Tingley, M. P. and P. Huybers (2010), A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison with the Regularized Expectation–Maximization Algorithm, J. Clim., 23(10), 2782–2800, doi:10.1175/2009jcli3016.1.

Tripati, A. K., Eagle, R. A., Thiagarajan, N., Gagnon, A. C., Bauch, H., Halloran, P. R. and J. M. Eiler (2010), 13C–18O isotope signatures and 'clumped isotope' thermometry in foraminifera and coccoliths, *Geochim. Cosmochim. Acta*, 74(20), 5697–5717, doi:10.1016/j.gca.2010.07.006.

Tripati, A. K., Hill, P. S., Eagle, R. A., Mosenfelder, J. L., Tang, J., Schauble, E. A., ... & Henry, D. (2015). Beyond temperature: Clumped isotope signatures in dissolved inorganic carbon species and the influence of solution chemistry on carbonate mineral composition. *Geochimica et Cosmochimica Acta*, *166*, 344-371.

Tripati, A. K., Sahany, S., Pittman, D., Eagle, R. A., Neelin, J. D., Mitchell, J. L., & Beaufort, L. (2014). Modern and glacial tropical snowlines controlled by sea surface temperature and atmospheric mixing. *Nature Geoscience*, *7*(3), 205-209.

Upadhyay, D., Lucarelli, J., Arnold, A., Flores, R., Bricker, H., Ulrich, R. N., ... & Tripati, A. (2021). Carbonate clumped isotope analysis (Δ47) of 21 carbonate standards determined via gas-source isotope-ratio mass spectrometry on four instrumental configurations using carbonate-based standardization and multiyear data sets. *Rapid Communications in Mass Spectrometry*, *35*(17), e9143.

Vermeesch, P. (2018). IsoplotR: A free and open toolbox for geochronology. *Geoscience Frontiers*, *9*(5), 1479-1493.

Wang, H., Liu, W., He, Y., Zhou, A., Zhao, H., Liu, H., ... & Liu, Z. (2021). Salinity-controlled isomerization of lacustrine brGDGTs impacts the associated MBT5ME'terrestrial temperature index. *Geochimica et Cosmochimica Acta*, *305*, 33-48.

Winkelstern, I. Z., and Lohmann, K. C. (2016), Shallow burial alteration of dolomite and limestone clumped isotope geochemistry. *Geology*, 44(6), 467-470.

Wu, C., and J. Z. Yu (2018), Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting. *Atmos. Meas. Tech.*, 11(2), 1233–1250.

References from the Supporting Information

Bernasconi, S. M., Müller, I. A., Bergmann, K. D., Breitenbach, S. F., Fernandez, A., Hodell, D. A., et al. (2018), Reducing uncertainties in carbonate clumped isotope analysis through consistent carbonate-based standardization. Geochem. Geophys., 19(9), 2895–2914.

Breitenbach, S. F., Mleneck-Vautravers, M. J., Grauel, A. L., Lo, L., Bernasconi, S. M., Müller, I. et al. (2018), Coupled Mg/Ca and clumped isotope analyses of foraminifera provide consistent water temperatures. Geochim. Cosmochim. Acta, 236, 283-296.

Davies, A. J., and C. M. John (2019), The clumped (13C18O) isotope composition of echinoid calcite: Further evidence for "vital effects" in the clumped isotope proxy. Geochim. Cosmochim. Acta, 245, 172-189.

Defliese, W. F., Hren, M. T., and K. C. Lohmann (2015), Compositional and temperature effects of phosphoric acid fractionation on Δ47 analysis and implications for discrepant calibrations. Chem. Geol., 396, 51-60.

Fernandez, A., Tang, J., and B. E. Rosenheim (2014), Siderite 'clumped' isotope thermometry: A new paleoclimate proxy for humid continental environments. Geochim. Cosmochim. Acta, 126, 411-421.

García del Real, P., Maher, K., Kluge, T., Bird, D. K., Brown Jr, G. E., and C. M. John (2016), Clumped-isotope thermometry of magnesium carbonates in ultramafic rocks. Geochim. Cosmochim. Acta, 193, 222-250.

Henkes, G. A., Passey, B. H., Wanamaker Jr, A. D., Grossman, E. L., Ambrose Jr, W. G., and M. L. Carroll (2013), Carbonate clumped isotope compositions of modern marine mollusk and brachiopod shells. Geochim. Cosmochim. Acta, 106, 307-325.

Jautzy, J. J., Savard, M. M., Dhillon, R. S., Bernasconi, S. M., and A. Smirnoff (2020), Clumped isotope temperature calibration for calcite: Bridging theory and experimentation. Geochem. Perspect. Lett, 14, 36-41.

Katz, A., Bonifacie, M., Hermoso, M., Cartigny, P., and D. Calmels (2017), Laboratory-grown coccoliths exhibit no vital effect in clumped isotope (Δ47) composition on a range of geologically relevant temperatures. Geochim. Cosmochim. Acta, 208, 335-353.

Kele, S., Breitenbach, S. F., Capezzuoli, E., Meckler, A. N., Ziegler, M., Millan, I. M., et al. (2015), Temperature dependence of oxygen-and clumped isotope fractionation in carbonates: a study of travertines and tufas in the 6–95 C temperature range. Geochim. Cosmochim. Acta, 168, 172-192.

Kelson, J. R., Huntington, K. W., Schauer, A. J., Saenger, C., and A. R. Lechler (2017), Toward a universal carbonate clumped isotope calibration: Diverse synthesis and preparatory methods suggest a single temperature relationship. Geochim. Cosmochim. Acta, 197, 104-131.

Kluge, T., and C. M. John (2015), Effects of brine chemistry and polymorphism on clumped isotopes revealed by laboratory precipitation of mono-and multiphase calcium carbonates. Geochim. Cosmochim. Acta, 160, 155-168.

Löffler, N., Fiebig, J., Mulch, A., Tütken, T., Schmidt, B. C., Bajnai, D., et al. (2019), Refining the temperature dependence of the oxygen and clumped isotopic compositions of structurally bound carbonate in apatite. Geochim. Cosmochim. Acta, 253, 19-38.

Meinicke, N., Ho, S. L., Hannisdal, B., Nürnberg, D., Tripati, A., Schiebel, R., and A. N. Meckler (2020), A robust calibration of the clumped isotopes to temperature relationship for foraminifers. Geochim. Cosmochim. Acta, 270, 160-183.

Müller, I. A., Rodriguez-Blanco, J. D., Storck, J. C., do Nascimento, G. S., Bontognali, T. R., Vasconcelos, C., et al. (2019), Calibration of the oxygen and clumped isotope thermometers for (proto-) dolomite based on synthetic and natural carbonates. Chem. Geol., 525, 1-17.

Peral, M., Daëron, M., Blamart, D., Bassinot, F., Dewilde, F., Smialkowski, N., et al. (2018). Updated calibration of the clumped isotope thermometer in planktonic and benthic foraminifera. Geochim. Cosmochim. Acta, 239, 1-16.

Petersen, S. V., Defliese, W. F., Saenger, C., Daëron, M., Huntington, K. W., John, C. M., et al. (2019). Effects of improved 17O correction on interlaboratory agreement in clumped isotope calibrations, estimates of mineral-specific offsets, and temperature dependence of acid digestion fractionation. Geochem. Geophys., 20(7), 3495-3519.

Petrizzo, D. A., Young, E. D., and B. N. Runnegar (2014), Implications of high-precision measurements of 13C–18O bond ordering in CO2 for thermometry in modern bivalved mollusc shells. Geochim. Cosmochim. Acta, 142, 400-410.

Piasecki, A., Bernasconi, S. M., Grauel, A. L., Hannisdal, B., Ho, S. L., Leutert, T. J., et al. (2019). Application of clumped isotope thermometry to benthic foraminifera. Geochem. Geophys., 20(4), 2082-2090.

Tang, J., Dietzel, M., Fernandez, A., Tripati, A. K., and B. E. Rosenheim (2014), Evaluation of kinetic effects on clumped isotope fractionation (Δ47) during inorganic calcite precipitation. Geochim. Cosmochim. Acta, 134, 120-136.

van Dijk, J., Fernandez, A., Storck, J. C., White, T. S., Lever, M., Müller, I. A., et al. (2019). Experimental calibration of clumped isotopes in siderite between 8.5 and 62° C and its application as paleo-thermometer in paleosols. Geochim. Cosmochim. Acta, 254, 1-20.

Wacker, U., Fiebig, J., Tödter, J., Schöne, B. R., Bahr, A., Friedrich, O., et al. (2014). Empirical calibration of the clumped isotope paleothermometer using calcites of various origins. Geochim. Cosmochim. Acta, 141, 127-144.

Winkelstern, I. Z., and Lohmann, K. C. (2016), Shallow burial alteration of dolomite and limestone clumped isotope geochemistry. Geology, 44(6), 467-470.

Table captions

Table 1. Distribution of measurement error in temperature that was used to design the synthetic datasets for this study. We provide examples of the materials that correspond to each category. For example, calibration datasets for synthetic carbonates grown at known temperatures, or benthic foraminifera from intermediate and deep-ocean sites, often have very well-constrained temperatures with errors of less than 0.5 °C (e.g., Ghosh et al., 2006; Tripati et al., 2010). Levels of error were defined based on the distribution of typical uncertainties reported for calibration temperatures in a recent synthesis of calibration data (Petersen et al., 2019). Temperatures in degree °C are transformed into $10^6/T^2$, with T in °K, for calibration purposes (after Ghosh et al., 2006).

Table 2. Distribution of measurement error in Δ_{47} used to design our synthetic datasets. Measurement error in Δ_{47} was estimated using the distribution of reported Δ_{47} errors in a synthesis of calibration data (Petersen et al., 2019). We present distributions for natural, synthetic, calcite, and aragonite samples. Distribution across sample types are under the "all materials" heading.

Table 3. Comparison of regression parameters estimated in Petersen et al. (2019) relative to our new estimates based on the same dataset. Results in Petersen et al. (2019) are based on a Monte Carlo sampling (10,000 replicates) for synthetic carbonate samples only (n = 451 replicates). We use a total of 10,000 replicates for estimating each regression parameter under each of the regression models implemented in BayClump. For our newly fit regressions, we report the mean and SE for each parameter.

Table 4. Comparison of Miocene and Recent temperatures reconstructed by Sun et al. (2021) using published calibrations from Tripati et al. (2015) and Henkes et al. (2013) to values obtained using BayClump for different calibration datasets and the York, OLS, and Bayesian regression models. We use results presented in Table 2 from Sun et al. (2021), with data provided in their Table S4. Temperature uncertainties are in SE and 95% CIs.

Temperature (°C)	Temperature ($10^6/T^2$)	Low		Intermediate		High		Very high
		0.25 °C	0.5 °C	1 °C	2 °C	3 °C	5 °C	10 °C
0	13.403	0.025	0.049	0.099	0.198	0.299	0.504	1.038
5	12.925	0.023	0.047	0.093	0.188	0.283	0.478	0.982
10	12.473	0.022	0.044	0.089	0.178	0.269	0.452	0.930
15	12.044	0.021	0.042	0.084	0.169	0.255	0.429	0.882
20	11.636	0.020	0.040	0.080	0.160	0.242	0.407	0.836
25	11.249	0.019	0.038	0.076	0.152	0.230	0.387	0.794
30	10.881	0.018	0.036	0.072	0.145	0.219	0.368	0.755
35	10.531	0.017	0.034	0.069	0.138	0.208	0.350	0.718
40	10.198	0.016	0.033	0.065	0.132	0.198	0.334	0.684
45	9.880	0.016	0.031	0.062	0.125	0.189	0.318	0.652
50	9.576	0.015	0.030	0.060	0.120	0.180	0.303	0.621
	Average:	0.019	0.038	0.077	0.155	0.234	0.394	0.808
Temperature uncertainty typical of calibrations for these material types		Synthetic carbonates, benthic foraminifera		Planktic foraminifera, lake carbonates		Some terrestrial carbonates		Natural dolomites

All data		Natural data		Synthetic data		Calcite		Aragonite		Level of error
Bin (‰)	Count	Bin (‰)	Count	Bin (‰)	Count	Bin (‰)	Count	Bin (‰)	Count	
0.000-0.005	123	0.000-0.005	96	0.000-0.005	20	0.000-0.005	88	0.000-0.005	9	Low
0.005-0.010	244	0.005-0.010	182	0.005-0.010	58	0.005-0.010	191	0.005-0.010	26	Intermediate
0.010-0.015	103	0.010-0.015	60	0.010-0.015	42	0.010-0.015	55	0.010-0.015	25	
0.015-0.020	44	0.015-0.020	18	0.015-0.020	24	0.015-0.020	10	0.015-0.020	4	High
0.020-0.025	18	0.020-0.025	7	0.020-0.025	11	0.020-0.025	8	0.020-0.025	2	
0.025-0.030	7	0.025-0.030	2	0.025-0.030	5	0.025-0.030	1	0.025-0.030	1	
0.030-0.035	3	0.030-0.035	1	0.030-0.035	2	0.030-0.035	2	0.030-0.035	0	
0.035-0.040	2	0.035-0.040	1	0.035-0.040	1	0.035-0.040	1	0.035-0.040	0	
0.040-0.045	2	0.040-0.045	1	0.040-0.045	1	0.040-0.045	0	0.040-0.045	0	
0.045-0.050	2	0.045-0.050	1	0.045-0.050	1	0.045-0.050	1	0.045-0.050	1	
0.050-0.055	1	0.050-0.055	0	0.050-0.055	1	0.050-0.055	0	0.050-0.055	0	
0.055-0.060	1	0.055-0.060	0	0.055-0.060	1	0.055-0.060	1	0.055-0.060	0	
0.060-0.065	1	0.060-0.065	0	0.060-0.065	1	0.060-0.065	1	0.060-0.065	0	Very High
Average:	0.01	Average:	0.0086	Average:	0.0133	Average:	0.0086	Average:	0.0105	

Study	Regression model	Slope	SE	Intercept	SE
Petersen et al. (2021)	Original calibration	0.0383	1.70E-06	0.258	1.70E-05
	Bayesian simple linear model (w/o errors)	0.0377	7.20E-06	0.260	8.11E-05
	Bayesian simple linear model (w/ errors)	0.0378	1.05E-05	0.259	1.16E-04
	Bayesian mixed model	0.0378	1.02E-05	0.260	1.15E-04
	Deming regression	0.0389	2.91E-05	0.245	2.85E-04
	Linear model	0.0377	1.04E-05	0.260	1.14E-04
	Inverse weighted linear model	0.0377	9.50E-06	0.259	1.03E-04
	York regression	0.0447	8.33E-05	0.191	9.79E-04
Anderson et al. (2021; full range)	Original calibration	0.0391	4.00E-04	0.154	4.00E-03
	Linear model	0.0388	2.45E-05	0.158	2.82E-04
	Inverse weighted linear model	0.0382	2.70E-05	0.164	3.10E-04
	York regression	0.0389	2.43E-05	0.157	2.73E-04
	Deming regression	0.0384	2.41E-05	0.160	2.83E-04
	Bayesian simple linear model (w/o errors)	0.0388	1.20E-05	0.157	1.25E-04
	Bayesian simple linear model (w/ errors)	0.0389	2.25E-05	0.157	2.29E-04
	Bayesian mixed model (Calcite)	0.0388	1.14E-04	0.154	1.34E-03
	Bayesian mixed model (Dolomite)	0.0395	3.54E-05	0.154	3.46E-04
Anderson et al. (2021; environmental temperature)	Linear model	0.0312	7.50E-05	0.250	9.43E-04
	Inverse weighted linear model	0.0315	9.20E-05	0.246	1.16E-03
	York regression	0.0313	8.04E-05	0.249	1.00E-03
	Deming regression	0.0313	6.20E-05	0.248	7.89E-04
	Bayesian simple linear model (w/o errors)	0.0311	4.64E-05	0.251	5.78E-04
	Bayesian simple linear model (w/ errors)	0.0305	9.09E-04	0.259	1.15E-02
	Bayesian mixed model (Calcite)	0.0331	1.10E-03	0.224	1.37E-02
	Bayesian mixed model (Dolomite)	0.0276	1.60E-03	0.296	1.97E-02

Sample Type	Age	Location	Lat/Long.	Elevation (m)	Δ47 (Sun et al. 2021; CDES)
Fossil aragonite shells	~6.2 Ma	huitangba, Zhaotong, Yunnai	27°19′44″ N, 103°44′15″ E	1918	0.706 ± 0.023
Modern aragonite shells	0	Fuxian Lake, Yunnan	24°26′55.5″ N, 102°51′18.3″ E	1722	0.708 ± 0.019

Published calibrations	Age	Average T (°C) (Henkes et al., 2013)	Average T (°C) (Tripati et al., 2015)	Average T (°C) (Petersen et al., 2019)	
Sun et al. (2021; Table 2)	~6.2 Ma	21.4 ± 8.5	17.6 ± 5.4	22.7 ± 7.6	
	0	20.6 ± 7.3	17.1 ± 4.5	26.9 ± 4.4	

BayClump - Dataset 1	Age	Average T (°C) (this study - OLS)	Average T (°C) (this study - B - LM)	Average T (°C) (this study - B - LM - E)	Average T (°C) (this study - B - LMM)
Petersen et al. (2019)	~6.2 Ma	17.5 ± 7.2	17.0 ± 3.5	16.8 ± 3.5	17.0 ± 3.5
	0	16.8 ± 5.95	17.6 ± 4.2	17.4 ± 4.2	17.6 ± 4.2

BayClump - Dataset 2	Age	Average T (°C) (this study - OLS)	Average T (°C) (this study - B - LM)	Average T (°C) (this study - B - LM - E)	Average T (°C) (this study - B - LMM)
Anderson et al. (2021) – Full range	~6.2 Ma	17.1 ± 6.9	16.6 ± 3.3	16.4 ± 3.3	15.4 ± 3.2
	0	16.4 ± 5.7	16.2 ± 5.8	17.0 ± 4.1	16.0 ± 4.0

BayClump - Dataset 3 (Anderson, Env)	Age	Average T (°C) (this study - OLS)	Average T (°C) (this study - B - LM)	Average T (°C) (this study - B - LM - E)	Average T (°C) (this study - B - LMM)
Anderson et al. (2021) – Environmental range	~6.2 Ma	18.1 ± 8.7	17.8 ± 4.4	17.4 ± 4.2	16.0 ± 3.8
	0	17.3 ± 7.2	18.7 ± 5.1	18.2 ± 5.1	16.7 ± 4.6

Δ_{47} (Sun et al. 2021; CDES90)
0.618 ± 0.023
0.620 ± 0.019
(this study - York)
20.6 ± 6.3
20.0 ± 5.2
(this study - York)
17.1 ± 6.9
16.5 ± 5.7
(this study - York)
18.1 ± 8.6
17.3 ± 7.2