BayClump: Bayesian Calibration and Temperature Reconstructions for Clumped Isotope Thermometry

C. Román-Palacios[1], H. M. Carroll[1], A. J. Arnold[1], R. J. Flores[1], S.V. Petersen[2], K. A. McKinnon[3], 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]Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA

[3]Department of Statistics and the Institute of the Environment and Sustainability, Center for Diverse Leadership in Science, University of California, Los Angeles, CA, USA

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

Key Points:

- We implement Bayesian methods for calibrating the carbonate 'clumped' isotope thermometer and reconstructing temperatures
- Ordinary least squares linear and Bayesian regressions recover true regression parameters with the highest accuracy and precision
- Bayesian temperature reconstructions are generally more precise and accurate than commonly used models
- BayClump is a Shiny dashboard that facilitates the use of Bayesian and non-Bayesian models for calibration and reconstruction

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. All models recover regression parameters within error of true values. Ordinary least squares linear and Bayesian models have the highest precision and accuracy. Congruently, for temperature reconstructions where the fitted model is used to predict temperature from Δ_{47}, Bayesian models generally outperform other regression models in both precision and accuracy. Our analyses suggest that depending on the structure of the examined dataset and relative to classical models, Bayesian regressions could improve the accuracy and precision of (i) calibration parameter estimates and (ii) temperature reconstructions by at least a factor of two. We implement our comparative framework into a new web-based interface, BayClump. This tool should increase reproducibility by enabling access to the different Bayesian and non-Bayesian regression models examined here. We utilize this tool with a published data synthesis to assess regression parameters and show that while both yield similarly accurate results, uncertainty in estimates of the slope and regression are reduced.

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 and comparatively examined the 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 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 et al., 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 modelling 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 isotopes 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 usingΔ_{47}. Thus, it remains unclear whether accounting for uncertainties in both variables actually improves the reliability of inferred regression parameters and reconstructed temperatures usingΔ_{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 a synthetic dataset. We focus on answering whether Bayesian models and error-in-variable models outperform models that ignore uncertainty in $10^6/T^2$. Building on BAYSPAR, a web-interface for Bayesian models for the TEX_{86} temperature proxy (Tierney and

Tingley, 2014), we develop BayClump, an RShiny Dashboard application 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). 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 (Tingley and Huybers 2010; Tierney and Tingley 2015).

2 Materials and Methods

2.1 General modeling framework

Here, we examine the performance of Bayesian and non-Bayesian linear models using synthetic datasets (but see section 4.5.1 for 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. We use the distribution of data to define "low", "intermediate", and "high" uncertainties in each of the variables. Therefore, the analyzed synthetic datasets follow different levels of error in Δ_{47} and $10^6/T^2$. In addition, synthetic datasets assume a linear relationship between 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/or precise values for the slope and intercept. Finally, we utilize the inferred regression parameters and their uncertainties from each model to reconstruct temperatures for assumed Δ_{47} values of 0.600‰, 0.700‰, and 0.800‰ that correspond to temperatures that are low (~10°C), moderate (~19°C), and high (~60°C). We account for different values of uncertainty in the analyzed Δ_{47} (low, intermediate, and high). We highlight that the calibration and reconstruction framework used in this study is useful to evaluate model performance (accuracy/precision) during both the calibration and reconstruction steps.

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. Three are Bayesian models. 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 linear model: 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 model: Second, we fit an ordinary least squares linear model with observations being weighted based on the inverse of their squared uncertainty in the measured Δ_{47}. In this model, observations with higher uncertainty (standard error) 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 standard error of each observation.

Deming regression: 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 D47 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 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 model: 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 for each observation, when fitting York models.

2.2.2 Bayesian linear regression models

Bayesian linear: Fifth, we fit a Bayesian linear regression model, which is 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:

$$\Delta_{47\ i} \sim Normal(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta \frac{10^6}{T^2{}_i}$$

$$\alpha \sim Normal(0.231, 0.065)$$

$$\beta \sim Normal(0.039, 0.004)$$

$$\sigma^2 \sim Gamma(10^{-3}, 10^{-3})$$

$$i = 1, \ldots, N$$

Priors for α and β follow previous publications (Table S3; references therein). We used uninformative priors (priors that make weak assumptions about the model) on the precision parameter σ^2 (see also prior distribution plots in the SI).

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 follows:

$$\Delta_{47\ i} \sim Normal(\Delta^{true}_{47\ i}, \sigma^2_{\Delta 47;\, i})$$

$$\Delta^{true}_{47\ i} \sim Normal(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta \frac{10^6}{T^2{}_i}$$

$$\frac{10^6}{T^2{}_i} \sim Normal(\frac{10^6}{T^2{}_i}^{true}, \sigma^2_{\frac{10^6}{T^2};i})$$

$$\frac{10^6}{T^2{}_i}^{true} \sim Normal(0, 10^{-3})$$

$$\alpha \sim Normal(0.231, 0.065)$$

$$\beta \sim Normal(0.039, 0.004)$$

$$\sigma^2 \sim Gamma(10^{-3}, 10^{-3})$$

$i = 1, \ldots, N$, where i is an index of each sample

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 $\Delta_{47\ i}$, with errors $\sigma_{\Delta 47}$, and the explanatory variable is the temperature in the form $\frac{10^6}{T^2}$, with errors $\sigma_{\frac{10^6}{T^2}}$. Finally, $\mu = \alpha + \beta\frac{10^6}{T^2}$ is the linear predictor, α is the intercept, β the slope, and σ^2 is the model error.

Previous research has found support for a normal prior on the relationship between Δ_{47}^{true} and observed Δ_{47} (Daëron, 2021). Note that the definition of this model presented above utilizes a prior on $\frac{10^6}{T^2}^{true}$ that can yield negative prior values of true $\frac{10^6}{T^2}$. A more appropriate prior to the general patterns of 'clumped-isotope' data is used in BayClump: $\frac{10^6}{T^2}_i^{true} \sim Normal(7, 10^{-2})$. We use uninformative priors (i.e. priors containing little relevant information) for defining α, β, and σ^2.

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.

In this paper, 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, this model is equivalent to the *Bayesian linear with errors* presented above:

$$\mu_i = \alpha_{j(i)} + \beta_{j(i)} \frac{10^6}{T^2}_i$$

$$\alpha_j \sim Normal(0.231, 0.065)$$

$$\beta_j \sim Normal(0.039, 0.004)$$

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

Therefore, this model allows for material-specific regression parameters. 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. The MCMC parameters are selected based on preliminary analyses that recovered convergence in datasets with variable size and different levels of error in variables. 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 used to run each of the models analyzed in this study is implemented in BayClump (see section 4.5 below).

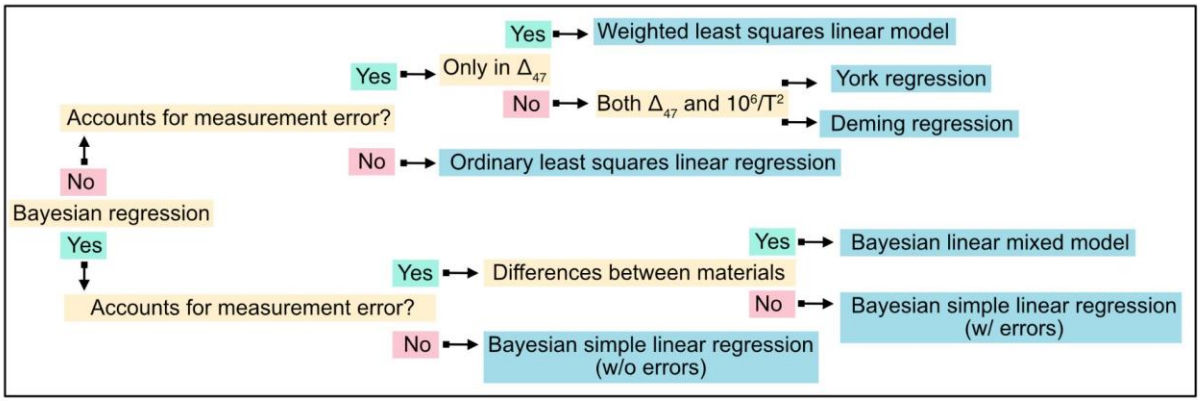

Fig. 1. Conceptual representation of the seven regression models used in this study for the derivation of Δ_{47}-temperature calibrations. We compared the performance of a total of seven regression models in deriving calibration relationships for use in carbonate clumped isotope thermometry. Parameter 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}. Specifically, 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 assume different scenarios with low, intermediate, and high levels of error. 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). For the Δ_{47} error across labs (see the σ_p parameter below), we follow the distribution of reported Δ_{47} errors in the Petersen et al. (2019) compilation. Low measurement error is defined using reported measurements of 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). Thus, our estimates for intermediate error fall in between our estimates 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). Finally, we estimate additional the levels of Δ_{47} error by examining the distribution of reproducibility for each lab that participated in the Intercarb interlaboratory exercise (Supplemental Table 1 in Bernasconi et al., 2021). Additionally, we have shown that 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. For Δ_{47} errors (see the σ_b parameter below), data is binned in 0.005‰ increments, and grouped into low, intermediate, and high error, with the intermediate error case being the most abundant (Table 1). For error in Δ_{47} across labs, we use 0.0125‰, 0.0225‰, as 0.0275‰ low, intermediate, and high error scenarios, respectively. For error the reproducibility of Δ_{47} caused by instrument noise, we use 0.0025‰, 0.0075‰, and 0.0125‰ for low, intermediate, and high error scenarios, respectively. 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.

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 analyze a total of 1,00010$^6/T^2$ observations with a normal distribution following parameters (informed based on Table S2):

$$\frac{10^6}{T^2} \sim Normal(12.02585352, 0.8759)$$

These observations are treated as the true $10^6/T^2$values. Next, we simulate random error in $10^6/T^2$ using on a normal distribution centered in 0 and with standard error following a given $10^6/T^2$ error scenario (σ):

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

The observed $10^6/T^2$ values resulted of 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) based on Bernasconi et al. (2021; see 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 additional 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):

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

Using model 6 (Bayesian linear model with errors) as our generative model, $\Delta_{47\ i}^{observed}$values are created by the addition of each initial $\Delta_{47\ i}$ value to a corresponding $\Delta_{47\ i}^{error}$. Analyses in the main text primarily focus on three simplified end-member error scenarios (Data Set S1–S3). First, we present results 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 examined 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 used a "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.

We use a bootstrapping (1,000 replicates per model per error scenario) approach for examining whether each of the models correctly recovered the true slope and intercept for the calibration. We present parameter estimates per model based on a total of 1,000 simulated bootstrap replicates. In these analyses and within each bootstrap replicate, we sample 50 observations from the original dataset. This number of samples (n=50) generally reflects the size of calibration datasets for individual materials in published clumped isotope calibration studies (e.g. Tripati et al., 2010; Petersen et al., 2019; Andersen et al., 2021). In addition, we examine results for calibration datasets of different sizes, assuming calibration datasets with n =10 and n = 500. Finally, we use the n bootstrapped replicates to estimate the median and error in the slope and intercept for each model. The 95% confidence intervals on slope and intercept range between the 2.5 and 97.5 percentiles, stating the range in which an estimate for the regression

coefficients is 95% likely to occur. Finally, we examine model performance based on the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) estimated for a given regression model with each error scenario.

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

Paleotemperatures can be inferred by applying a regression model to sample Δ_{47} values. To evaluate model performance for temperatures reconstructions, we apply the estimated regression parameters to three target Δ_{47}values (0.6‰, 0.7‰, and 0.8‰) with several scenarios for replicate measurement errors in Δ_{47} (i.e. 0.005‰, 0.01‰, 0.02‰, 0.04‰, 0.06‰). Note that each of the three target Δ_{47}values represent hypothetical carbonates that might be of particular interest to researchers. Our goal is to show whether reconstructions based on these particular carbonates under- or overestimate true temperature under each of the examined regression models and scenarios of error. Several different reconstruction approaches are used here. Our approach is simplified in Fig. 2 and described in the next two paragraphs.

2.4.1. Inverting non-Bayesian models

We conduct an assessment of model performance using two scenarios that either ignore or account for uncertainty in regression parameters. When we ignore error in regression parameters, we simply account for error in Δ_{47} by predicting temperature based on the limits of the associated 95% CI in the target Δ_{47}. Alternatively, we present the 95% CI in predictions when we sample across our bootstrapped estimates of slope and intercept, as well as target Δ_{47}. We account for correlation between slope and intercept by paired regression parameter estimates. We note that previous research has suggested that in particular proxies, accounting for error in regression parameters is less important than accounting for uncertainty in the regression model (McClelland et al. 2021). This later aspect will be discussed below in section 4.4.

2.4.2. Inverting Bayesian models

We perform temperature reconstructions either within or outside of a Bayesian framework. For temperature reconstructions outside of a Bayesian framework, we use regression parameters (slope and intercept) extracted from a Bayesian structure. We follow the same approach as the one described in the paragraph above for non-Bayesian models (i.e. ignore or account for parameter error). For temperature reconstructions within a Bayesian framework, we conducted "backwards" predictions within Jags for the median, lower, upper limits in the 95% CI of Δ_{47}. For Bayesian and non-Bayesian reconstructions, temperature estimates were based on the median and percentile-based 95% confidence intervals of regression parameters across 1,000 replicates.

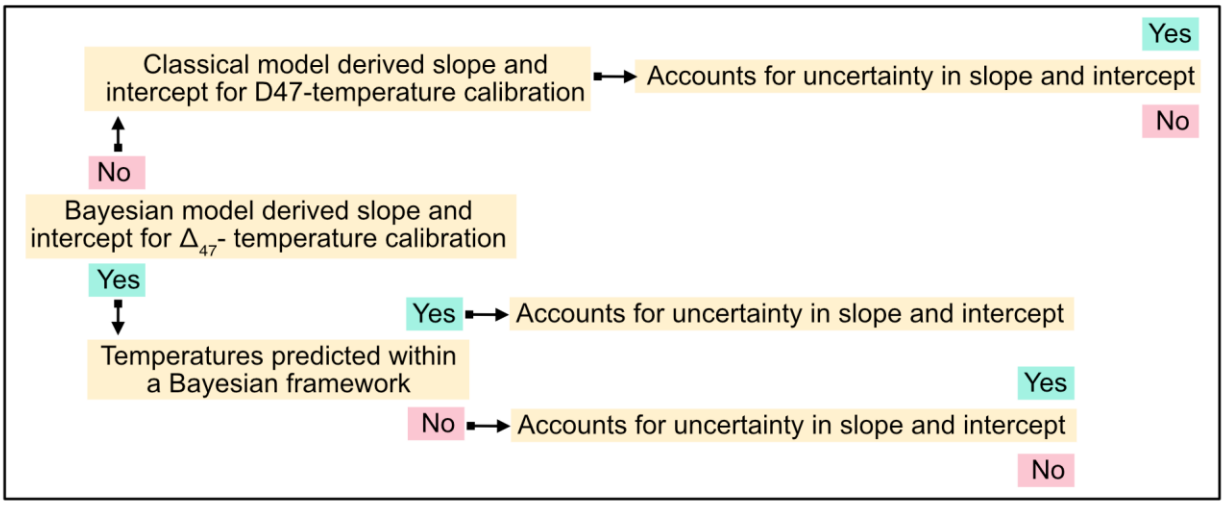

Fig. 2. Conceptual representation of the different methods that can be used to calculate temperatures from Δ_{47}using calibrations derived from classical and Bayesian regression models. Fig. 1 showed different classical and Bayesian regression models that we explored. We then compared three different approaches for reconstructing temperature using model-derived slopes and intercepts, and associated uncertainties. Temperature reconstructions can be derived using slopes and intercepts by applying classical regression models to calibration datasets, and can either factor in uncertainties in model-derived slopes and intercepts, or not. Calibration slopes and intercepts derived using Bayesian models can be applied to data outside of a Bayesian framework, and can either factor in uncertainties in model-derived slopes and intercepts, or not. Alternatively, calibration slopes and intercepts derived using Bayesian models can be applied to data within a Bayesian framework, and therefore will factor in uncertainties in model slopes and intercepts.

3 Results and Discussion

3.1 Model performance in calibration datasets with 50 data points

We found major differences in the performance of classical and Bayesian models with our synthetic datasets for each error scenario considered. However, the 95% CI of all regression models overlap with the true regression slope and intercept (Table S4). Therefore, 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.

Results on model performance are congruent across all the scenarios of error examined in this study. In general, we found that Bayesian and ordinary least squares linear models recover true parameters with the highest precision and accuracy. Parameter estimates from York models are the least accurate and precise among all the examined models. In particular, York regressions strongly underestimate the intercept and overestimate the slope. The remaining models (i.e. weighted ordinary least squares linear and deming) underestimate the slope but overestimate the intercept at an extent that is intermediate between York, Bayesian, and ordinary least squares

regression models. Below, we provide additional details on model performance within each scenario of simulated error.

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

Under a low-error scenario, the York regression is the most inaccurate and imprecise model (Fig. 3A–F; Data Set S4). This regression model underestimates the true intercept by 12% and overestimates the true slope by 7% (Fig. 3E). The weighted least squares and Deming regressions overestimate the true intercept by ~4–5% and underestimate the true slope by ~3% (Figs. 3D and 3F, respectively). Bayesian and ordinary least squares linear models are the most accurate, overestimating the true intercept by 2–3% and underestimating the true slope by ~1% (Figs. 3A–C). Relative to ordinary least squares linear and Bayesian models, the uncertainty in the intercept and slope was at least 36% higher in the weighted least squares regression model, 85% higher in the Deming model, and 283% higher in the York model.

We found similar results under an intermediate-error scenario. Our analyses suggest that a York regression is the most inaccurate model (Data Set S5; Fig. S1). This regression model underestimates the true intercept by 23% and overestimates the true slope by 13%. The weighted least squares and Deming regressions overestimate the true intercept by 7–10% and underestimate the true slope by 3–7%. Again, the Bayesian and ordinary least squares linear regressions are the most accurate models, overestimating the true intercept by ~8% and underestimating the true slope by 4–5%. Relative to Bayesian and ordinary least squares linear models, uncertainty in the estimated intercept and slope is at least 29%, 90%, and 145% higher in the weighted least squares regression model, Deming model, and York model, respectively.

Finally, our analyses suggest that a York model is also the most inaccurate model under a high-error scenario. Specifically, the true intercept is consistently and strongly underestimated using a York model (Fig. S2). This parameter is also overestimated using both weighted least squares and Deming regression models (Data Set S6). The York regression underestimates the true intercept by 42% and overestimates the true slope by 25%. The weighted least squares and Deming regressions overestimate the true intercept by 35–40% and underestimate the true slope by 22–25%. Bayesian and ordinary least squares linear models overestimate the true intercept by ~21–28% and underestimate the true slope by 14–17%. Relative to Bayesian and ordinary least squares linear models, uncertainty in the intercept and slope is at least 43% higher in the weighted least squares regression model, 94% higher in the Deming regression model, and 126% higher in the York model.

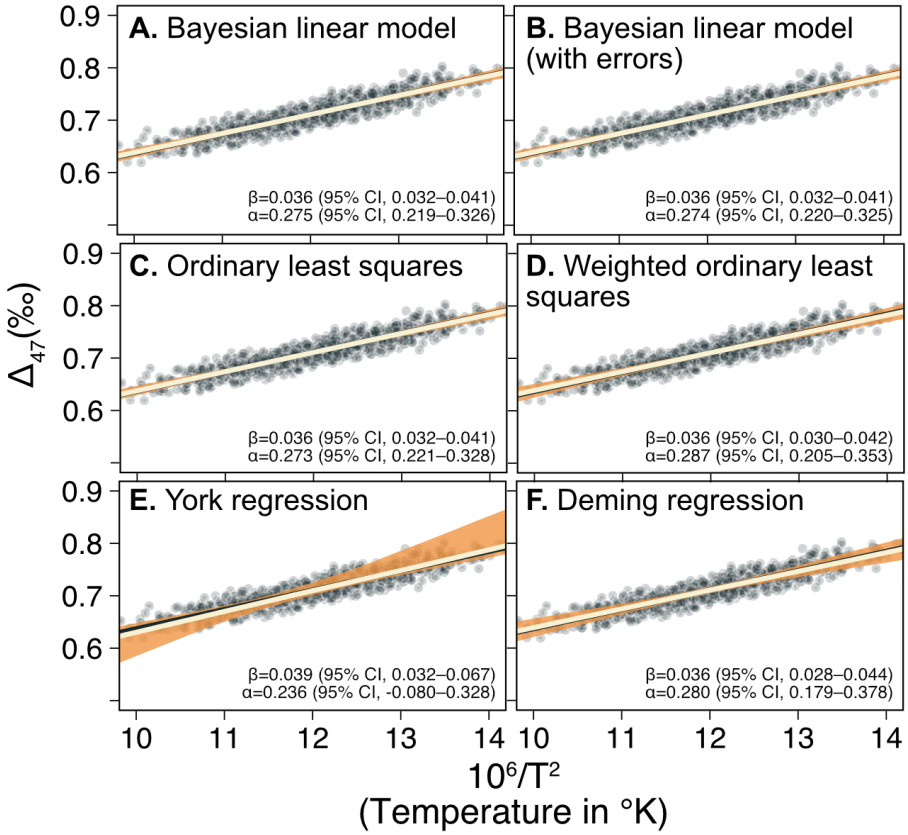

Fig. 3. Performance of different statistical models for deriving clumped isotope temperature calibrations evaluated using a synthetic dataset with low uncertainties in both Δ_{47} and T. Results shown for a low-error scenario with measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C. The true slope = 0.0369 and intercept = 0.268. In black, we show the synthetic dataset, with each symbol corresponding to an individual observation, with 50 "sample values" and associated uncertainties in both variables used to bootstrap 1,000 observations in total. For each model, we show the median regression line (yellow solid line) and associated 95% CI (orange area) based on the 1,000 bootstrapped values. Values for the slope (β) and intercept (α) and their uncertainties are also shown. See Table S4 and Data Set S4 for additional details. Results for intermediate-error scenarios are provided in Table S4 and Data Set S5. Figures for the corresponding error scenarios are also presented in Fig. S1 and S2. Results for high-error scenarios are provided in Data S4 and Data Set S6.

3.1.2 Standardized errors

We also assess model performance using standardized errors RMSE (Tables S4–S6) and MAE (Fig. 4). Results are shown for calibration datasets of intermediate size (n=50. Table S4; see Tables S5 and S6 for additional results). Both indices are measures of the difference between predicted and true Δ_{47} values, and showed similar values for most models. For brevity, we describe trends observed in MAE, but note that a similar pattern is observed in RMSE (Tables

S4–S6). First, the Bayesian and ordinary least squares linear models show the lowest values for MAE. MAE values are also almost indistinguishable between the two models across all three error scenarios. Second, the weighted least squares regression yields a similar but slightly higher MAE than Bayesian and ordinary least squares linear models (1.3–2.4% higher; range across models and error scenarios). Third, the Deming regression model shows 5–8% more error than Bayesian and ordinary least squares linear models. Finally, York regression models recover 14–21% higher errors than those in Bayesian and ordinary least squares linear models.

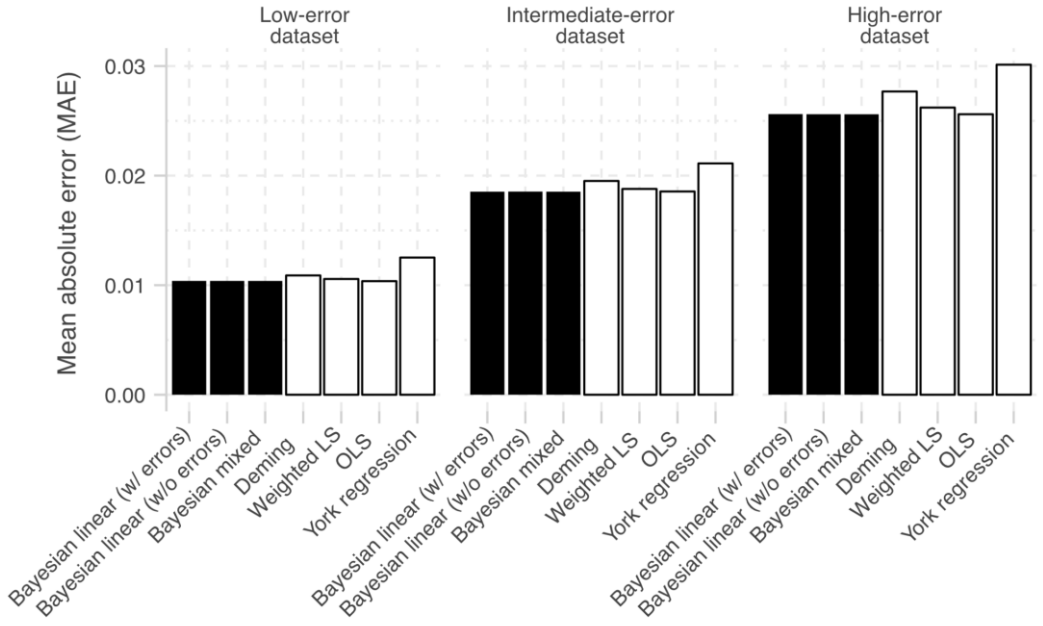

Fig. 4. Estimates of the mean absolute error (MAE) for each regression model used for deriving a Δ_{47}-T calibration. Results are based on a synthetic Δ_{47}-T calibration dataset, and three different error scenarios. For a given model, the higher the estimated Δ_{47}, the larger the difference between predicted and true Δ_{47}. Numerical estimates for MAE and RMSE are presented in Table S4. Bayesian models are shown in black. In the figure, OLS: ordinary least squares regression; Weighted LS: weighted least squares regression. Results for alternative error scenarios and calibration datasets with different sample sizes are provided in Tables S3–S5 and Figures S3–S4.

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

Our recommendations are summarized in Table 3 (see also Tables S4–S6). In general, our analyses suggest that for calibration purposes and across all error scenarios examined in this study, ordinary least squares linear and Bayesian models performed the best in terms of both

accuracy and precision. Weighted least squares models consistently show an intermediate performance across all datasets. Deming regression models perform similarly to weighted least squares models in low and intermediate error scenarios. However, Deming and York regression models perform similarly and are particularly poor in performance in high-error scenarios. York regression models consistently underestimate the true intercept (12–42% across error scenarios), overestimate the true slope (14–51% across error scenarios) and recover the highest uncertainty in regression parameters across all the models (95–287% higher than Bayesian models; Tables S4–S6). Among the Bayesian models, the Bayesian linear model with error in variables performs slightly better, particularly in scenarios with increased error in variables (i.e. low- vs. high-error scenario; Tables S4–S6).

We acknowledge that our results on the performance of the York model 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. We also note that differences between our study and Wu and Yu (2018) could simply be explained by how bias was calculated. Specifically, Wu and Yu (2018) calculated bias in parameter estimates for each model using mean values (we used median estimates). Finally, 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, their results are not generalizable to our study given that the mean Y to X ratio in our 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. The simulated datasets analyzed in this study assumed no error in variables. As an additional note, York et al. (2004) does not provide a direct comparison of the performance of the York regression relative to other models. Therefore, we suggest that evidence on whether models that account for error in both variables ("error-in-variables models") actually outperform other regression types in every dataset merits further study.

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

In addition to examining model performance on datasets with intermediate sample size (n=50; Figs. 3–4; Table S4; Data Sets S4–S6), we compare precision and accuracy between models in datasets with a smaller (i.e. n=10; Figs. S5–S7; Table S5; Data Sets S7–S9) and larger

sample size (n=500 data points; Figs. S8–S10; Table S6; Data Sets S10–S18). Patterns of model performance (precision and accuracy) in 10- and 500-datapoint datasets largely resemble those based on 50-datapoint datasets (Tables S4–S6). However, although precision is consistently affected by the size of the calibration dataset, model accuracy is largely similar across calibration datasets of varying size (Fig. 5). Therefore, larger datasets are generally ideal for any regression model when reducing uncertainty in regression parameters is desirable.

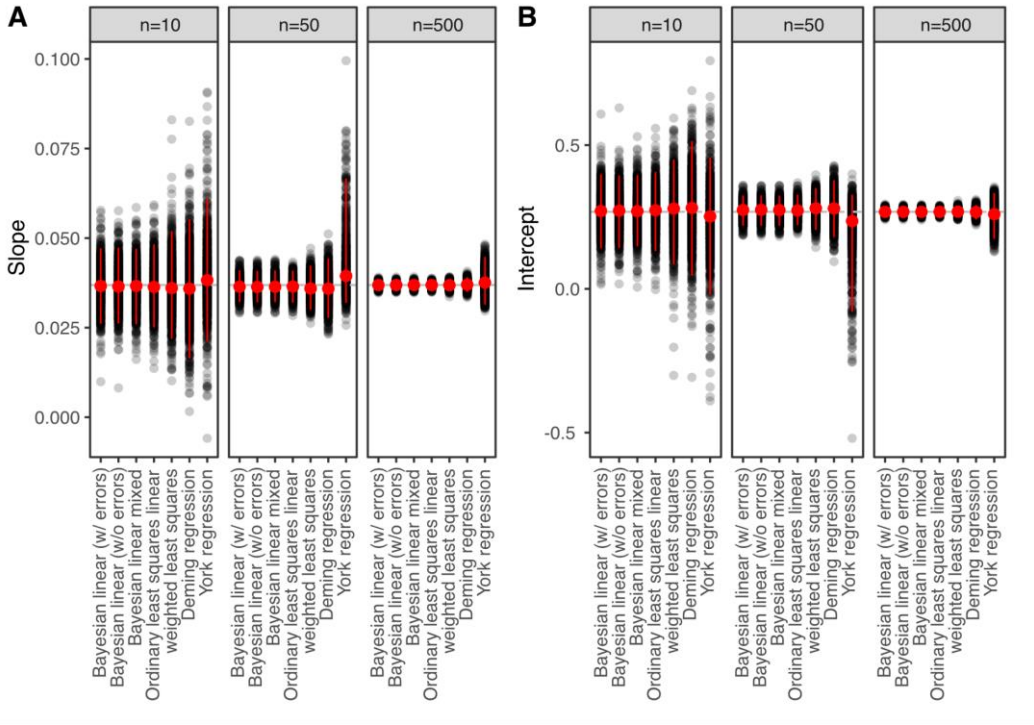

Fig. 5. Patterns of parameter uncertainty across regression models based on calibration datasets of variable size. Results are shown for a low-error scenario (measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C). However, patterns are largely similar to those under intermediate- and high-error scenarios. We present median parameter estimates for the (A) slope and (B) intercept for each given model. We also indicate quantile-based 95% CIs. We fit regression models based on datasets with 10, 50, and 500 data points (n indicated on top of each panel). The raw data used to generate this figure is provided in Data Sets S4–S6, S8, S10, S12, S14, S16, and S18.

3.3 General recommendations for selecting models to calibrate the clumped isotopes paleothermometer

We use our results presented in sections 4.1 and 4.2 to generalize major patterns of model performance between calibration datasets of different sizes. We summarize general recommendations in Table 3. Overall, regardless of the size of the calibration dataset, Bayesian

and ordinary least squares linear models outperform any other regression model in terms of both accuracy and precision. Under low- and high-error scenarios, Bayesian models are more precise and accurate and even slightly better than ordinary least squares linear models. Under intermediate-error scenarios, ordinary least squares linear models are slightly better than Bayesian regression models. Therefore, Bayesian linear models are not consistently better than any of the other models in every single situation. Instead, our results highlight the particular scenarios in which a given regression model generally outperforms alternative regressions. Finally, our analyses indicate that increasing the number of observations in the calibration dataset does not necessarily imply an overall improvement in accuracy for all the parameters. However, the size of the calibration dataset does have important consequences on affecting model precision.

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

We utilize the calibration slopes and intercepts derived using different regression models (Fig. 4; Table S4) to reconstruct temperatures, and evaluate model performance (i.e. the accuracy and precision of reconstructed values). Table 4 summarizes recommendations for methods to utilize for reconstructions. We perform temperature reconstructions both outside of a Bayesian framework and within a Bayesian framework. We note that temperature reconstructions ignoring calibration regression parameter uncertainty are the usual practice in the field. Based on our analyses, non-Bayesian reconstructions that account for error in calibration regression parameters and error in target Δ_{47}show the highest error among all the reconstructions. However, accounting for uncertainty in regression parameters is also not an approach that has been extensively used in previous studies in the field. Our results on the higher uncertainty in reconstructed temperatures when errors in regression parameters are accounted for differ from those presented in a recent study (McClelland et al. 2021). Specifically, the authors indicated that, for calibrations based on certain temperature proxies, uncertainty in reconstructed temperatures has been found to be more strongly affected by regression error than by uncertainty in regression parameters. However, the relative roles of regression versus parameter error seem to be controlled by whether or not the analyzed predictor is the actual variable exerting the strongest control on the response in nature. In the case of clumped isotopes data, there is indeed a strong causal association between the two variables.

Below, we provide a description of temperature reconstruction performance based on Bayesian and non-Bayesian reconstructions for the seven different regression models. Note that our non-Bayesian reconstructions ignore uncertainty in regression parameters. Results of non-Bayesian temperature reconstructions accounting for parameter error are only shown in the supplement (Data Set S19). Our results in the following sections are also supported by extensive and more detailed descriptions presented in sections Text S1–S3.

3.4.1 Reconstructions for low-temperature carbonates (Δ_{47} =0.8‰; T ~ 10 ^{o}C; Fig. 6; Text S1)

Depending on the examined scenario of error, non-Bayesian temperature reconstructions of Bayesian models (low-error scenario), weighted least squares models (intermediate-error scenario), and Bayesian models accounting for error in variables (high-error scenario) recover the best accuracy among the examined models (i.e. median temperature reconstructions were 0.19°C, 0.9°C, and 3.4°C different from true temperature values, respectively; Fig. 8). Bayesian temperature reconstructions consistently outperform all the other models in terms of precision, especially when error in the target Δ_{47} was small (<0.01‰). However, the true temperature is not within the 95% CI under Bayesian reconstructions in three particular situations: when error in target Δ_{47} is 0.005‰ (calibration dataset with low-error and high-error scenarios) or 0.01‰ (calibration dataset with high error scenario). Finally, our results suggest that when Bayesian reconstructions are not able to produce reliable temperature reconstructions (i.e. true temperature is now within the predicted 95% CI), weighted least squares models (intermediate error) and non-Bayesian reconstructions for Bayesian models (high-error) are optimal alternatives for reconstructing temperatures with the highest possible precision and accuracy. Our analyses also indicate that while York models consistently overestimate true temperatures for low-temperature carbonates, other models including Bayesian regressions and weighted least squares models are able to correctly reconstruct true temperature with the best precision and accuracy at different levels of error in both the calibration dataset and in the analyzed Δ_{47}.

3.4.2 Reconstructions for intermediate-temperature carbonates (Δ_{47} = 0.7‰; T ~ 19 °C; Fig. 7; Text S2)

The models showing the highest precision are different between error scenarios (Fig. 9A–C.I). Median temperature reconstruction under Bayesian reconstructions matches perfectly to the second decimal place relative to the true value of temperature (i.e. reconstructions differ from true temperature by only ~0.003°C in the low-error scenario). Bayesian and non-Bayesian ordinary least squares linear models differ by 0.14–0.16°C relative to the true temperature in the intermediate scenarios of error, and Bayesian simple models (non-Bayesian reconstructions) by 0.13°C in the high-error scenario. In terms of precision (Fig. 9A–C.II), Bayesian temperature reconstructions consistently outperform any of the other models by inferring temperatures with the lowest uncertainty. In short, our results show that York models consistently underestimate true temperature and only perform relatively well when error in target Δ_{47} is large (>0.02‰). On the other hand, Bayesian models, either under Bayesian temperature reconstructions or not, generally outperform other methods for reconstructing temperature at Δ_{47} = 0.7‰.

3.4.3 Reconstructions for high-temperature carbonates (Δ_{47}= 0.6‰; T ~ 60 °C; Fig. 8; Text S3)

Temperature reconstructions based on Bayesian models are generally more precise and accurate than those based on any of the other models examined in this study (Fig. 10). However, Bayesian reconstructions fail to recover true temperature at small errors in Δ_{47} for scenarios of intermediate (error in target Δ_{47}=0.005‰) and high error (error in target Δ_{47}=0.005‰ and 0.01‰). We suggest that when Bayesian reconstructions are not able to produce reliable temperature reconstructions (error in target Δ_{47}=0.005‰ [intermediate- and high-error scenarios] and 0.01‰ [high-error scenario]; i.e. true temperature is not within the predicted 95% CI), non-

Bayesian reconstructions for Bayesian models are optimal alternatives to reconstruct temperatures with the highest precision and accuracy among the examined models.

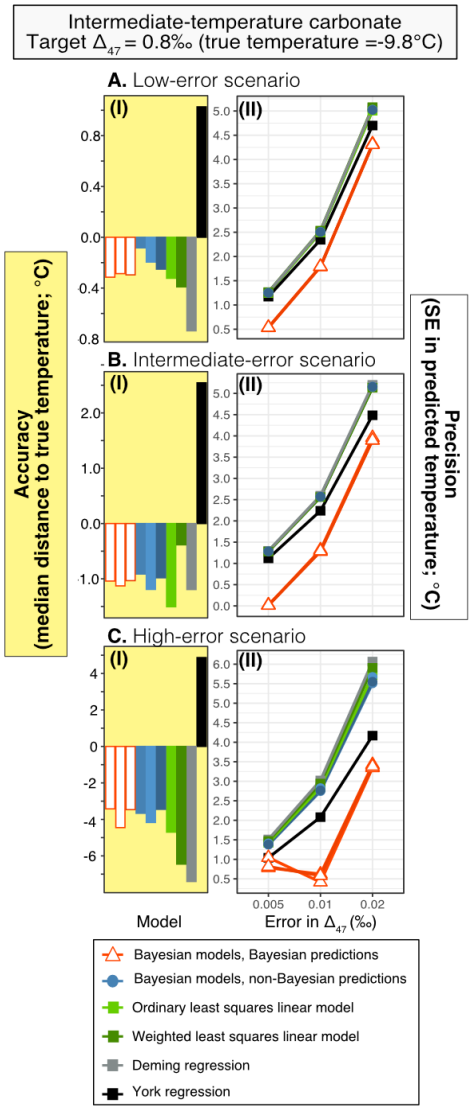

Fig. 6. Comparison of temperature reconstruction performance across regression models. Results are shown in terms of temperature reconstruction accuracy (yellow boxes) and precision (white boxes) for three error scenarios (low-, intermediate, and high-errors), for a single target Δ_{47} (0.8‰), and three errors in target Δ_{47} (0.005‰, 0.01‰, and 0.02‰). We show results for Bayesian and non-Bayesian reconstructions. For Bayesian reconstructions, results are shown for three Bayesian models. For non-Bayesian reconstructions, temperature estimates are shown for seven regression models. These non-Bayesian reconstructions ignore parameter uncertainty. Here, accuracy is summarized as the median distance to true temperature for a target Δ_{47}. Specifically, we subtract the median predicted temperature for a given model to the true temperature. Therefore, negative values of accuracy would indicate that the predicted temperature is lower than the true temperature by a certain number of degrees (in °C). Thus, we

expected a value of zero in accuracy if a model recovered true temperature high accuracy (i.e. no difference between true and predicted value). Next, we summarized the standard error (2 SE based on 95% CI) in predicted temperature to outline the precision of each accuracy. Standard

error is in °C. Note that despite all models being plotted, several overlap. Numerical values shown in this figure are reported in Data Set S19.

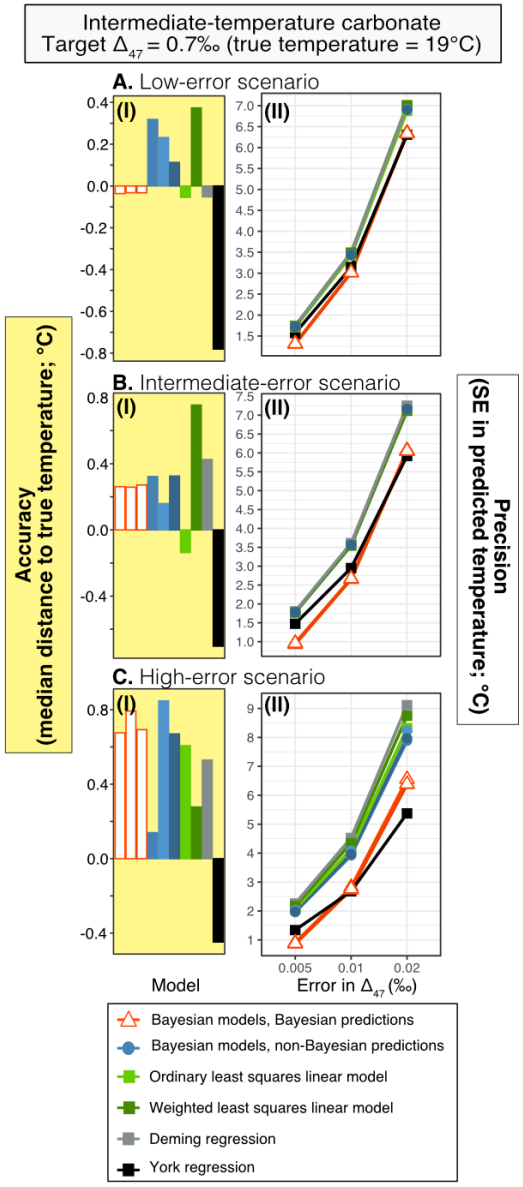

Fig. 7. Comparison of temperature reconstructions performance across regression models. As in Fig. 6 but for a target Δ_{47}=0.7‰.

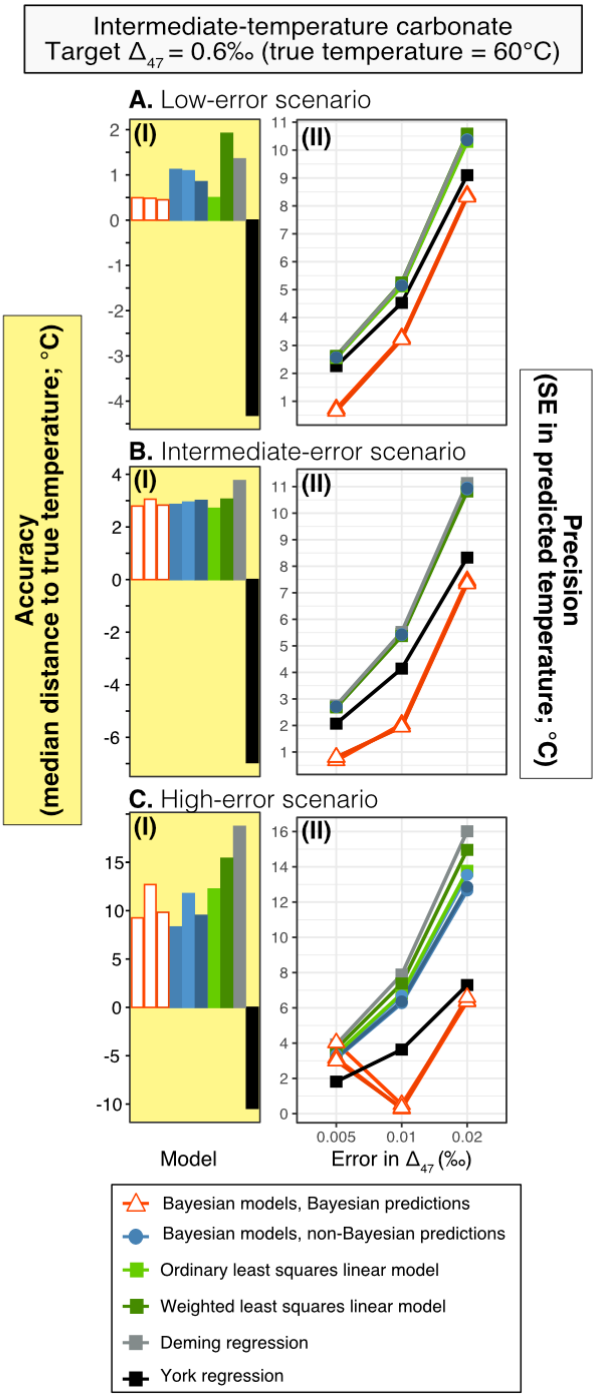

Fig. 8. Comparison of temperature reconstructions performance across regression models. As in Fig. 6 but for a target Δ_{47}=0.6‰.

4.5 BayClump

To support the use of Bayesian models 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. 9). 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), and different datasets will be added 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 the Petersen and Anderson datasets to create a larger calibration set if desired. Any data 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 the values 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. Finally, BayClump provides the ability for the user to download full calibration regression model output and any or all 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 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, 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.

4.5.1 Re-analyzing 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 the relevant study, the authors provided estimates of the slope and intercept using a Monte Carlo least squares regression based on 10,000 replicates (Table 5). Here, we analyze the same dataset using all seven regression models analyzed in this study. We perform a total of 10,000 replicates of each model implemented in BayClump. Finally, we compare regression coefficients and their associated uncertainties between models and relative to Petersen et al. (2019).

Overall, we found that parameter estimates under the newly implemented Bayesian models differed from parameter estimates in Petersen et al. (2019) by less than 2% (Table 5). Therefore, both approaches (Bayesian linear regressions and Monte Carlo least squares regression in Petersen et al. [2021]) provided similar value for the slope and intercept. Parameter estimates under the OLS, weighted OLS, York and Deming regression models (all run inside BayClump) differed from parameter estimates in Petersen et al. (2019) by only 2%, 1%, 1–2%, and 2–4%, respectively. Therefore, point estimates for the slope and intercept were similar across the analyzed models.

Next, we found that uncertainty (standard error, SE) in regression parameters was the main factor that differed between regression models. First, Bayesian models recovered 7% less uncertainty in the slope and 3% less uncertainty in the intercept relative to the model published in Petersen et al. (2021; average of SE across Bayesian models). Second, the OLS recovered 6% less uncertainty in the slope and only 1% less uncertainty in the intercept. Third, York and weighted models recovered ~100% more uncertainty in both the slope and intercept relative to the model in Petersen et al. (2019). Fourth, uncertainty in the slope and intercept based on the Deming regression model was ~480% higher than in Petersen et al. (2019). Overall, our results are consistent with results from the synthetic datasets that suggest parameter optimization under least squares (e.g. Monte Carlo as in Petersen et al. [2019]) shows a similar performance to Bayesian regression models. However, Bayesian models are able to recover similar regression parameters and slightly decrease parameter uncertainty.

In summary, overall, based on our analyses for the Petersen et al. (2019) dataset, we show that while differences in point estimates for regression parameters are largely similar between Bayesian and the Monte Carlo least squares regression models (*sensu* Petersen et al. [2019]), Bayesian regression models implemented in this study are able to reduce uncertainty in parameters estimates by up to 7% relative to Petersen et al. (2019).

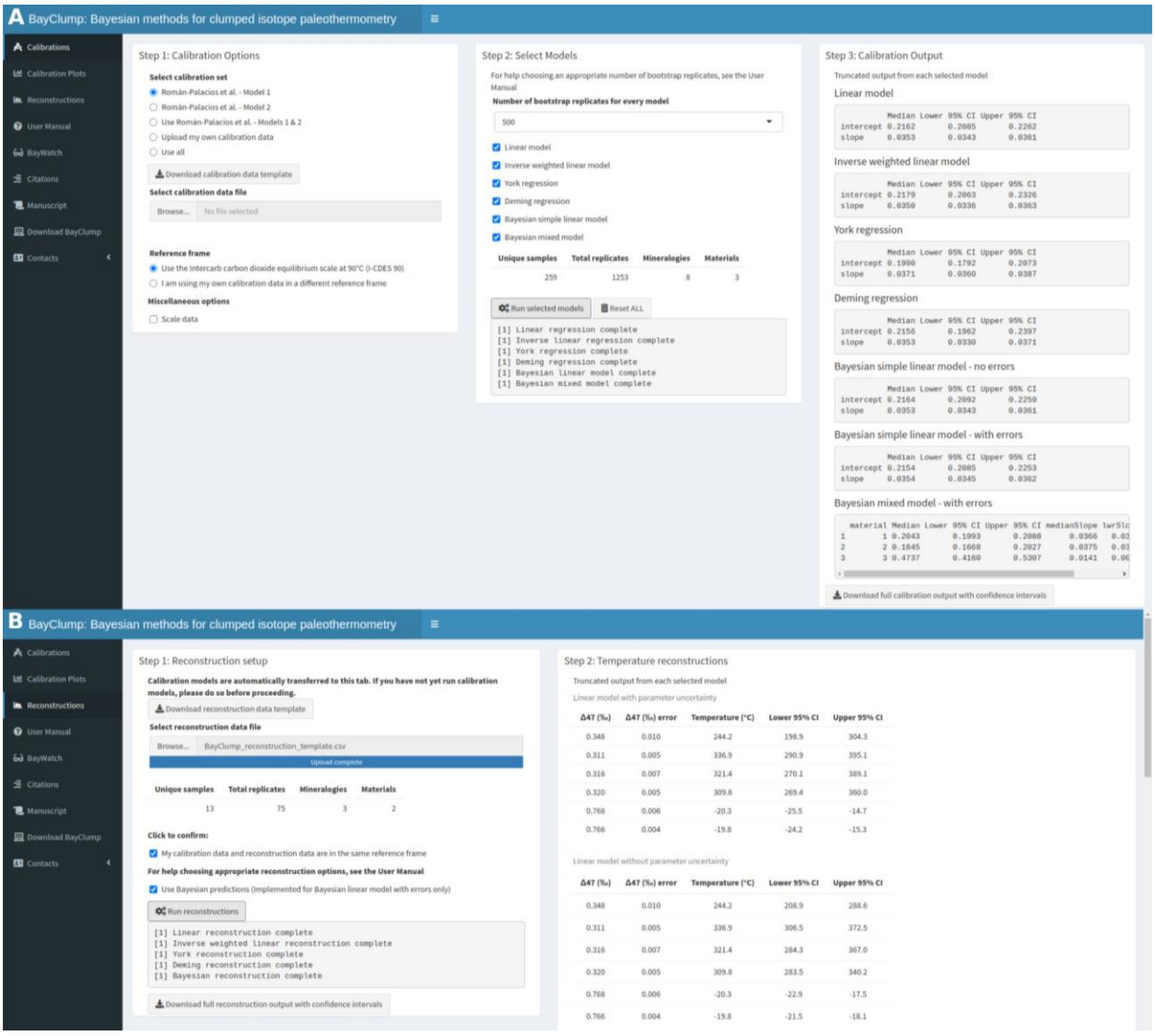

Fig. 9. BayClump application screenshots for proxy calibration and for temperature reconstruction. Top: Calibration options are chosen in the first tab of the application. Multiple preloaded datasets are included, or the user may opt to upload their own calibration data. Summary information is provided for each selected regression model upon run completion. Bottom: 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, labeled Excel spreadsheets. Not shown: Plots of calibration data and each

selected regression model are provided in the Plots tab. Plots are fully interactive and downloadable.

5 Conclusions

We examined the performance of seven regression models for calibrating the clumped isotope paleothermometer, including the first Bayesian implementations of regression models. Additionally, we implemented a Bayesian linear mixed model that can accommodate for differences in regression parameters between groups (e.g. materials can have different slopes and/or intercepts). Using simulated calibration datasets with variable number of observations (10 to 500 samples) and degrees of error in clumped isotope measurements and temperature, we found that Bayesian and non-Bayesian ordinary least squares linear models consistently outperformed other regression models in terms of accuracy and precision under most synthetic scenarios.

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‰). Under higher error in Δ_{47} (>0.01‰), York models were able to recover a smaller uncertainty in predicted temperatures. 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 the calibration synthesis dataset from Petersen et al. (2019) yields similar values for the slope and intercept to those reported in the original study, but with reduced uncertainty. The analytical tools developed in this study are available in BayClump, a Shiny dashboard with data templates that facilitates the use of Bayesian and non-Bayesian methods for both calibration and temperature reconstruction.

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. 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 DOE BES DE-SC0010288, NSF EAR-1936715, 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 supported through a postdoctoral fellowship by the Institutional Research and Academic Career Development Awards (IRACDA) program at UCLA (Award # K12 GM106996).

Data Availability Statement

Synthetic datasets are provided in the Supplement and code used to run the models is available through BayClump's GitHub repository (https://github.com/Tripati-Lab/BayClump). All data and code will also be available on FigShare upon publication. The Petersen et al. (2019) calibration datasets are available through EarthChem data repository, as detailed in the original paper.

References

Anand, P., Elderfield, H. and M. H. Conte (2003), Calibration of Mg/Ca thermometry in planktonic foraminifera from a sediment trap time series, *Paleoceanography*, 18(2), 1050, doi:10.1029/2002pa000846.

Bard, E., Raisbeck, G., Yiou, F. and J. Jouzel (2000), Solar irradiance during the last 1200 years based on cosmogenic nuclides, *Tellus B*, 52(3), 985–992, doi:10.1034/j.1600-0889.2000.d01-7.x, 2000.

Bard, E., Rostek, F. and C. Sonzogni (1997) Interhemispheric synchrony of the last deglaciation inferred from alkenone palaeothermometry, *Nature*, 385(6618), 707–710, doi:10.1038/385707a0.

Barker, S., Cacho, I., Benway, H. and K. Tachikawa (2005), Planktonic foraminiferal Mg/Ca as a proxy for past oceanic temperatures: a methodological overview and data compilation for the Last Glacial Maximum, *Quat. Sci. Rev.*, 24(7–9), 821–834, doi:10.1016/j.quascirev.2004.07.016.

Bernasconi, S., Daëron, M., Bergmann, K. D., Bonifacie, M. and A. N. Meckler (2020), 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.

Bian, N. and P. A. Martin (2010), Investigating the fidelity of Mg/Ca and other elemental data from reductively cleaned planktonic foraminifera, *Paleoceanography*, 25(2), PA2215, doi:10.1029/2009pa001796.

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.

Conte, M., Weber, J., King, L. and S. Wakeham (2001), The alkenone temperature signal in western North Atlantic surface waters, *Geochim. Cosmochim. Acta*, 65(23), 4275–4287, doi:10.1016/s0016-7037(01)00718-9, 2001.

Conte, M. H., Sicre, M.-A., Rühlemann, C., Weber, J. C., Schulte, S., Schulz-Bull, D. and T. Blanz (2006), Global temperature calibration of the alkenone unsaturation index (UK′37) in surface waters and comparison with surface sediments, *Geochem. Geophys.*, 7(2), Q02005, doi:10.1029/2005gc001054.

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., 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.

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.

Elderfield, H. and G. Ganssen (2000), Past temperature and δ18O of surface ocean waters inferred from foraminiferal Mg/Ca ratios, *Nature*, 405(6785), 442–445, doi:10.1038/35013033.

Garidel-Thoron, T. D., Rosenthal, Y., Bassinot, F. and L. Beaufort (2005), Stable sea surface temperatures in the western Pacific warm pool over the past 1.75 million years, *Nature*, 433(7023), 294–298, doi:10.1038/nature03189.

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.

Grauel, A.-L., Schmid, T. W., Hu, B., Bergami, C., Capotondi, L., Zhou, L. and S. M. Bernasconi (2013) Calibration and application of the 'clumped isotope' thermometer to foraminifera for high-resolution climate reconstructions, *Geochim. Cosmochim. Acta*, 108, 125–140, doi:10.1016/j.gca.2012.12.049.

Greaves, M., Barker, S., Daunt, C. and H. Elderfield (2005), Accuracy, standardization, and interlaboratory calibration standards for foraminiferal Mg/Ca thermometry, Geochem. Geophys., 6(2), Q02D13, doi:10.1029/2004gc000790.

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.

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.

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

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.

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.

Kim, J.-H., Schouten, S., Hopmans, E. C., Donner, B. and J. S. S. Damsté (2008), Global sediment core-top calibration of the TEX86 paleothermometer in the ocean, *Geochim. Cosmochim. Acta*, 72(4), 1154–1173, doi:10.1016/j.gca.2007.12.010.

Koutavas, A. (2002), El Nino-Like Pattern in Ice Age Tropical Pacific Sea Surface Temperature, *Science*, 297(5579), 226–230, doi:10.1126/science.1072376, 2002.

Lea, D. W., Martin, P. A., Pak, D. K. and H. J. Spero (2002), Reconstructing a 350ky history of sea level using planktonic Mg/Ca and oxygen isotope records from a Cocos Ridge core, Quat. Sci. Rev. 21(1-3), 283–293, doi:10.1016/s0277-3791(01)00081-6.

Lea, D. W., Mashiotta, T. A. and H. J. Spero (1999), Controls on magnesium and strontium uptake in planktonic foraminifera determined by live culturing, *Geochim. Cosmochim. Acta*, 63(16), 2369–2379, doi:10.1016/s0016-7037(99)00197-0, 1999.

Leider, A., Hinrichs, K.-U., Mollenhauer, G. and G. J. Versteegh (2010), Core-top calibration of the lipid-based U37K′ and TEX86 temperature proxies on the southern Italian shelf (SW Adriatic Sea, Gulf of Taranto), *Earth Planet. Sci. Lett.*, 300(1-2), 112–124, doi:10.1016/j.epsl.2010.09.042.

Martin, P. A., Lea, D. W., Rosenthal, Y., Shackleton, N. J., Sarnthein, M. and T. Papenfuss (2002), Quaternary deep sea temperature histories derived from benthic foraminiferal Mg/Ca, *Earth Planet. Sci. Lett.*, 198(1–2), 193–209, doi:10.1016/s0012-821x(02)00472-7.

Martin, P. A. and W. D. Lea (2002), A simple evaluation of cleaning procedures on fossil benthic foraminiferal Mg/Ca, *Geochem. Geophys.*, 3(10), 1–8, doi:10.1029/2001gc000280.

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.

Mashiotta, T. A., Lea, D. W. and H. J. Spero (1999), Glacial–interglacial changes in Subantarctic sea surface temperature and δ18O-water using foraminiferal Mg, *Earth Planet. Sci. Lett.*, 170(4), 417–432, doi:10.1016/s0012-821x(99)00116-8.

McClelland, H. L., Halevy, I., Wolf-Gladrow, D. A., Evans, D., and A. S. Bradley (2021), Statistical uncertainty in paleoclimate proxy reconstructions, *Geophys. Res. Lett.*, e2021GL092773.

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.

Müller, P. J., Kirst, G., Ruhland, G., Storch, I. V. and A. Rosell-Melé (1998), Calibration of the alkenone paleotemperature index U37K′ based on core-tops from the eastern South Atlantic and the global ocean (60°N-60°S), *Geochim. Cosmochim. Acta*, 62(10), 1757–1772, doi:10.1016/s0016-7037(98)00097-0.

Nürnberg, D., Bijma, J. and C. Hemleben, (1996), Assessing the reliability of magnesium in foraminiferal calcite as a proxy for water mass temperatures, *Geochim. Cosmochim. Acta*, 60(5), 803–814, doi:10.1016/0016-7037(95)00446-7.

Pak, D. K., Lea, D. W. and J. P. Kennett (2004), Seasonal and interannual variation in Santa Barbara Basin water temperatures observed in sediment trap foraminiferal Mg/Ca, *Geochem. Geophys.*, 5(12), doi:10.1029/2004gc000760.

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.

Powers, L., Werne, J. P., Vanderwoude, A. J., Damsté, J. S. S., Hopmans, E. C. and S. Schouten (2010), Applicability and calibration of the TEX86 paleothermometer in lakes, *Org. Geochem.*, 41(4), 404–413, doi:10.1016/j.orggeochem.2009.11.009.

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/

Rosenthal, Y., Perron-Cashman, S., Lear, C. H., Bard, E., Barker, S., Billups, K., Bryan, M., Delaney, M. L., Demenocal, P. B., Dwyer, G. S., Elderfield, H., German, C. R., Greaves, M., Lea, D. W., Marchitto, T. M., Pak, D. K., Paradis, G. L., Russell, A. D., Schneider, R. R., Scheiderich, K., Stott, L., Tachikawa, K., Tappa, E., Thunell, R., Wara, M., Weldeab, S. and P. A. Wilson (2004), Interlaboratory comparison study of Mg/Ca and Sr/Ca measurements in planktonic foraminifera for paleoceanographic research, *Geochem. Geophys.*, 5(4), doi:10.1029/2003gc000650.

Russell, A. D., Hönisch, B., Spero, H. J., and D. W. Lea (2004), Effects of seawater carbonate ion concentration and temperature on shell U, Mg, and Sr in cultured planktonic foraminifera, *Geochim. Cosmochim. Acta*, 68(21), 4347–4361, doi:10.1016/j.gca.2004.03.013.

Sachs, J. P., Schneider, R. R., Eglinton, T. I., Freeman, K. H., Ganssen, G., Mcmanus, J. F. and D. W. Oppo (2000), Alkenones as paleoceanographic proxies, *Geochem. Geophys.*, 1(11), doi:10.1029/2000gc000059.

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.

Schouten, S., Huguet, C., Hopmans, E. C., Kienhuis, M. V. M. and J. S. S. Damsté (2007), Analytical Methodology for TEX86Paleothermometry by High-Performance Liquid Chromatography/Atmospheric Pressure Chemical Ionization-Mass Spectrometry, *Anal. Chem.*, 79(7), 2940–2944, doi:10.1021/ac062339v.

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

Thiagarajan, N., Adkins, J. and J. Eiler (2011), Carbonate clumped isotope thermometry of deep-sea corals and implications for vital effects, *Geochim. Cosmochim. Acta*, 75(16), 4416–4425, doi:10.1016/j.gca.2011.05.004.

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.

Visser, K., Thunell, R. and L. Stott (2003), Magnitude and timing of temperature change in the Indo-Pacific warm pool during deglaciation, *Nature*, 421(6919), 152–155, doi:10.1038/nature01297.

Zachos, J. (2001), Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present, *Science*, 292(5517), 686–693, doi:10.1126/science.1059412.

Zachos, J. C., Dickens, G. R. and R. E. Zeebe (2008), An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics, *Nature*, 451(7176), 279–283, doi:10.1038/nature06588.

Zachos, J. C., Stott, L. D. and K. C. Lohmann (1994), Evolution of Early Cenozoic marine temperatures, *Paleoceanography*, 9(2), 353–387, doi:10.1029/93pa03266.

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 on temperatures that was used to inform the design of simulated 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 degrees (°C) are transformed into $10^6/T^2$, with T in K, for calibration purposes (Ghosh et al., 2006).

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	9.88	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
Material types		Synthetic carbonates, benthic foraminifera		Planktic foraminifera, lake carbonates		Some terrestrial carbonates		Natural dolomites

Table 2. Distribution of measurement error on Δ_{47} used to design our simulated datasets. In table, measurement error in Δ_{47} was estimated using the distribution of reported Δ_{47} errors in a recent 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, which was used to design simulations shown here.

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	High
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	
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	Very High
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	
Average:	0.01	Average:	0.0086	Average:	0.0133	Average:	0.0086	Average:	0.0105	

Table 3. Summary of model recommendations for calibrations based on the 'clumped isotopes' paleothermometer. We present a list of the regression models yielding the best performance under different scenarios of error in synthetic datasets (low [measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C], intermediate [measurement error in Δ_{47}= 0.0075‰, instrument error Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2°C], and high errors [measurement error in Δ_{47}= 0.0125‰, instrument error Δ_{47}= 0.0275‰, measurement error in $10^6/T^2$= 5°C]), target Δ_{47} (0.6‰, 0.7‰, and 0.8‰) and errors in target Δ_{47} (0.005‰, 0.01‰, 0.05‰). We also provide guidance on models that yield the best performance for different calibration dataset sizes (n=10, 50, 500).

Error scenario in the calibration dataset	Calibration dataset (n)	Recommended calibration model
Low (measurement error in Δ_{47}= 0.0025‰, error in true Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C)	10	Bayesian and simple linear models
	50	Bayesian and simple linear models
	500	Bayesian and simple linear models
Intermediate (measurement error in Δ_{47}= 0.0075‰, error in true Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2°C)	10	Bayesian and simple linear models
	50	Bayesian and simple linear models
	500	Bayesian and simple linear models
High (measurement error in Δ_{47}= 0.0125‰, error in true Δ_{47}= 0.0275‰, measurement error in $10^6/T^2$= 5°C)	10	Bayesian and simple linear models
	50	Bayesian and simple linear models
	500	Bayesian and simple linear models

Table 4. Summary of model recommendations for reconstructions using the 'clumped isotopes' paleothermometer. We present a list of the regression models yielding the best reconstruction performance under different scenarios of error in the synthetic dataset (low [measurement error in Δ_{47}= 0.0025‰, instrument error Δ_{47}= 0.0125‰, measurement error in $10^6/T^2$= 0.25°C], intermediate [measurement error in Δ_{47}= 0.0075‰, instrument error Δ_{47}= 0.0225‰, measurement error in $10^6/T^2$= 2°C], and high errors [measurement error in Δ_{47}= 0.0125‰, instrument error Δ_{47}= 0.0275‰, measurement error in $10^6/T^2$= 5°C]), target Δ_{47} (0.6‰, 0.7‰, and 0.8‰) and errors in target Δ_{47} (0.005‰, 0.01‰, 0.05‰).

Reconstruction target Δ_{47} (‰)	Error in Δ_{47} (‰)	Recommended reconstruction model
0.6 (high-temperature carbonates)	0.005 (low)	Bayesian reconstructions
	0.01 (intermediate)	Bayesian reconstructions
	0.05 (high)	Bayesian reconstructions
0.7 (intermediate-temperature carbonates)	0.005 (low)	Bayesian reconstructions
	0.01 (intermediate)	Bayesian reconstructions
	0.05 (high)	Bayesian reconstructions
0.8 (low-temperature carbonates)	0.005 (low)	Bayesian reconstructions and non-Bayesian reconstructions for Bayesian models
	0.01 (intermediate)	Bayesian reconstructions and non-Bayesian reconstructions for Bayesian models
	0.05 (high)	Bayesian reconstructions and non-Bayesian reconstructions for Bayesian models

Table 5. 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 present the mean and SE for each parameter.

Regression model	Slope	SE	Intercept	SE
Petersen et al. (2021)	0.0383	1.70E-06	0.258	1.70E-05
Linear model	0.0376	1.60E-06	0.262	1.68E-05
Inverse weighted linear model	0.0379	3.71E-06	0.260	3.72E-05
York regression	0.0386	3.65E-06	0.252	3.68E-05
Deming regression	0.0391	1.00E-05	0.248	9.95E-05
Bayesian simple linear model (w/o errors)	**0.0376**	**1.62E-06**	**0.262**	**1.69E-05**
Bayesian simple linear model (w/ errors)	**0.0377**	**1.57E-06**	**0.262**	**1.65E-05**
Bayesian mixed model	**0.0377**	**1.54E-06**	**0.262**	**1.63E-05**