AGU PUBLICATIONS

Geochemistry, Geophysics, Geosystems

RESEARCH ARTICLE
10.1002/2017GC007106

Special Section:
Clumped Isotope Geochemistry: From Theory to Applications

A Reassessment of the Precision of Carbonate Clumped Isotope Measurements: Implications for Calibrations and Paleoclimate Reconstructions

Alvaro Fernandez[1], Inigo A. Müller[1], Laura Rodríguez-Sanz[2], Joep van Dijk[1], Nathan Looser[1], and Stefano M. Bernasconi[1]

[1]Geological Institute, ETH Zürich, Zürich, Switzerland, [2]Research School of Earth Sciences, The Australian National University, Canberra, ACT, Australia

Key Points:
- Highly precise Δ_{47} data can be obtained with multiple replicate analyses
- Disagreements in calibration slopes can be largely explained by poor sample replication, small sample sizes, and narrow temperature ranges
- Uncertainties on Δ_{47} data should be reported as margins of error at a specified confidence level (e.g., 68% or 95% CL)

Correspondence to:
A. Fernandez,
alvaro.bremer@erdw.ethz.ch

Citation:
Fernandez, A., Müller, I. A., Rodríguez-Sanz, L., van Dijk, J., Looser, N., & Bernasconi, S. M. (2017). A reassessment of the precision of carbonate clumped isotope measurements: Implications for calibrations and paleoclimate reconstructions. *Geochemistry, Geophysics, Geosystems*, *18*, 4375–4386. https://doi.org/10.1002/2017GC007106

Received 30 JUN 2017
Accepted 3 NOV 2017
Accepted article online 15 NOV 2017
Published online 7 DEC 2017

Abstract Carbonate clumped isotopes offer a potentially transformational tool to interpret Earth's history, but the proxy is still limited by poor interlaboratory reproducibility. Here, we focus on the uncertainties that result from the analysis of only a few replicate measurements to understand the extent to which unconstrained errors affect calibration relationships and paleoclimate reconstructions. We find that highly precise data can be routinely obtained with multiple replicate analyses, but this is not always done in many laboratories. For instance, using published estimates of external reproducibilities we find that typical clumped isotope measurements (three replicate analyses) have margins of error at the 95% confidence level (CL) that are too large for many applications. These errors, however, can be systematically reduced with more replicate measurements. Second, using a Monte Carlo-type simulation we demonstrate that the degree of disagreement on published calibration slopes is about what we should expect considering the precision of Δ_{47} data, the number of samples and replicate analyses, and the temperature range covered in published calibrations. Finally, we show that the way errors are typically reported in clumped isotope data can be problematic and lead to the impression that data are more precise than warranted. We recommend that uncertainties in Δ_{47} data should no longer be reported as the standard error of a few replicate measurements. Instead, uncertainties should be reported as margins of error at a specified confidence level (e.g., 68% or 95% CL). These error bars are a more realistic indication of the reliability of a measurement.

1. Introduction

Over the past decade, the number of laboratories that routinely measure clumped isotopes as well as the scope of geoscience questions that are amenable to this technique have rapidly expanded. Consequently, a considerable amount of time and effort has been spent improving measurement techniques (Bernasconi et al., 2013; He et al., 2012; Hu et al., 2014; Huntington et al., 2009; Meckler et al., 2014; Petersen & Schrag, 2014; Schmid & Bernasconi, 2010) and data reduction algorithms (Daëron et al., 2016; John & Bowen, 2016; Schauer et al., 2016), expanding calibration data sets (see Kelson et al., 2017 for a review), and standardizing measurements (Dennis et al., 2011; Meckler et al., 2014). This combined effort has set a relatively strong practical foundation for the field of carbonate clumped isotope geochemistry. Nevertheless, the work is not yet complete, and perhaps the most pressing matter is the disagreement among laboratories on the temperature sensitivity of the clumped isotope proxy.

Several workers have attempted to pinpoint the reasons behind the lack of consensus in the slope of clumped isotope calibrations (Defliese et al., 2015; Dennis & Schrag, 2010; Fernandez et al., 2014; Kelson et al., 2017; Murray et al., 2016), with some popular hypotheses being that methodological differences in sample preparation, digestion temperatures, phosphoric acid quality, the precipitation methods of synthetic carbonate samples, ^{17}O correction parameters used in data processing, and/or analytical errors in Δ_{47} data are behind the disagreements.

The role of most of these variables has been mostly discounted by recent publications (e.g., Defliese et al., 2015; Kelson et al., 2017; Murray et al., 2016), and the choice of ^{17}O correction parameters and analytical errors in Δ_{47} data are the only hypotheses that are still considered strong possibilities. For instance, the choice of ^{17}O correction parameters can affect the accuracy of Δ_{47} data, and several workers have shown

that it is likely responsible for some of the disagreement among calibrations (Daëron et al., 2016; Kelson et al., 2017; Schauer et al., 2016). Consequently, there is a large community-wide calibration experiment in progress to recalculate the published calibration data with updated ^{17}O correction parameters.

The role of analytical errors, on the other hand, has only been indirectly addressed as a reason behind the discrepancies between published calibrations. For instance, Kelson et al. (2017) showed that calibration slopes can be easily biased by small number of samples, an observation that implies that published calibrations are not statistically distinct. Similarly, Bonifacie et al. (2017) recognized that the slopes and intercepts of published calibration relationships can be significantly influenced by low numbers of samples and replicate analyses, and by the narrow range of temperatures that were sometimes investigated. Moreover, they demonstrate that when errors at the 95% confidence level (CL) are considered, a weighted linear regression through most of the available data results in a relationship that is consistent with a universal calibration. More recently, Katz et al. (2017) measured the Δ_{47} compositions of coccoliths grown at known temperatures; however, they did not propose a coccolith-specific calibration because they acknowledged that a calibration constructed with their data could be biased due to their small sample size and restricted temperature range. Instead, they use their data to rule out vital effects and suggest that sea surface temperatures from sedimentary coccoliths should be calculated with a calibration that is more robustly constrained, such as the calibration of Bonifacie et al. (2017).

While there is likely an important role for both ^{17}O parameters and analytical errors in slope disagreements, here we focus exclusively on the errors that result due to the low number of replicate measurements that are commonly carried out for clumped isotope studies. We investigate if the way errors in Δ_{47} data are commonly reported can be problematic, and demonstrate that a limited number of replicates do not always produce data with the precision necessary for many paleoclimate reconstructions. Additionally, we show that imprecise data caused by insufficient replication, together with elements of the design of calibration experiments can explain the inconsistencies in published temperature calibrations. Finally, we discuss strategies to increase the precision of clumped isotope measurements and present recommendations for how to best report analytical errors.

2. Methods

Carbonate clumped isotopes refer to the overabundance of the doubly substituted ^{13}C-^{18}O-^{16}O isotopologue of CO_2 released from the phosphoric acid digestion of carbonate minerals relative to a stochastic distribution of isotopes among all possible isotopologues. The excess of this isotopologue is temperature-dependent, and it is reported with the parameter Δ_{47} in ‰ (see Eiler, 2007, for a detailed definition). Clumped isotope measurements are particular compared to other common geochemical analyses in the sense that the variance observed in repeated measurements of a sample is relatively large compared to the total range of natural variations. For instance, typical errors associated with a single measurement (about 15–30 ppm; 1 standard deviation) are approximately 5–10 times larger than the signal expected for a 1°C temperature change (~3 ppm at 25°C). For a proper evaluation of the error associated with the analysis of clumped isotopes it is, therefore, necessary to obtain multiple replicate measurements. In section 2.1 we first examine the effect of sample replication on the estimated precision of a measurement, which can be used to determine the optimal number of replicates to reach a target precision at a given laboratory. In section 2.2 we then evaluate the effect that the number of replicates, the number of samples, and the range of temperatures chosen for a calibration experiments can have on the robustness of the slope of the T – Δ_{47} calibrations.

Table 1
External Reproducibility of Carbonate Standards From Nine Different Clumped Isotope Laboratories Expressed as the Long-Term Standard Deviation (σ) of Multiple Measurements in ppm

Reference	σ ± ppm	Number of analyses (n)	Standards
Tripati et al. (2015)	17	72	Carrara marble
Wacker et al. (2014)	22	152	NBS-19, Arctic islandica
Staudigel and Swart (2016)	31	155	Marble
Zaarur et al. (2013)	36	119	Carrara marble
Kluge and John (2015)	26	74	Carrara marble
Henkes et al. (2013)	15	195	Carrara marble, 102-GC-AZ01
Katz et al. (2017)	14	>300	Carrara marble, 102-GC-AZ01, NBS-19
Müller et al. (2017)	24	428	Carrara marble, ETH 1–4
Kelson et al. (2017)	27	380	Coral, C64
Mean	24		

Note. In some cases, the standard deviations were calculated from the reported s.e.m and n. In all cases, the standard deviations were calculated from populations of individual digestions of carbonate samples (i.e., one replicate equals one acid digestion).

2.1. Sample Replication and Analytical Errors

We explored the effect of poor sample replication using a resampling experiment and published estimates of external reproducibilities across different clumped isotope laboratories (Table 1). Because external reproducibilities can vary over short time periods, we only

considered publications where the long-term standard deviation (several weeks to months) of a large number of replicate analyses (i.e., >30) is reported. Moreover, we only considered cases where standard deviations were reported for populations of individual digestions of CO_2 from carbonate standards (i.e., replicates are defined as single extractions of CO_2), which is the relevant metric that should be used to compare external reproducibilities across different methods and laboratories. In the cases where several standards with the above characteristics were reported, the different standard deviations were averaged to estimate the average reproducibility.

The resampling experiment has two main goals. The first goal is to examine what is the precision of clumped isotope measurements using standard analytical methods. Specifically, we want to determine the average precision at the 95% confidence level (95% CL) of a Δ_{47} value that resulted from averaging together only three Δ_{47} measurements and compare it to the commonly reported standard error of the mean (s.e.m). We selected three replicates for this experiment because it is the typical number of measurements performed for clumped isotope determinations in many laboratories. The second goal is to understand how precision changes with additional analyses. These results can then be used to determine the ideal number of replicates—additional mineral digestions—that should be measured while taking into account the investment of additional analytical time and sample availability constraints.

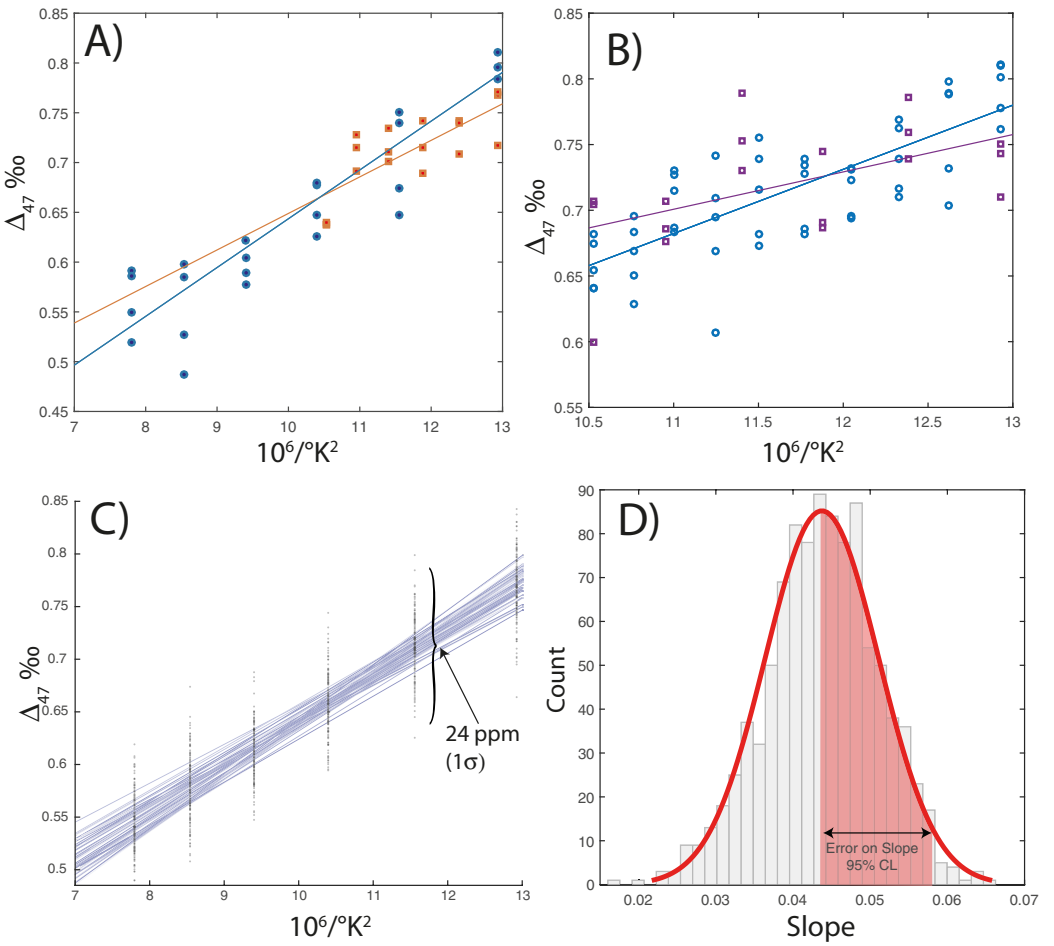

Figure 1. Details of the Monte Carlo simulations used to explore the effect of analytical error and experimental design on calibration uncertainties. (a) Example of Δ_{47} calibration data for two experiments with the same number of samples but with different number of replicate measurements and temperature ranges. (b) Example of Δ_{47} calibration data for two experiments with the same temperature range but with different number of samples and replicate measurements. (c) Δ_{47} data of a calibration experiment with (six samples and three replicates) showing 50 possible linear regressions through the data (for clarity only the first 50 trials are shown). (d) Histogram of 1,000 possible slopes for the experiment in Figure 1c. The errors on the slopes were calculated from these distributions (margins of error at the 95% CL).

Both of these questions were investigated by randomly drawing N subsamples (N = 3–15) from three synthetic data sets. The synthetic data sets (N = 10,000) were created using random numbers assuming a normal distribution with standard deviations equal to the average (25 ppm), maximum (36 ppm), and minimum (14 ppm) values observed in the laboratories shown in Table 1, and a mean Δ_{47} value of 0.7 ‰. For each set of subsamples, the means ($\bar{x}$), standard deviations (s), s.e.m, and the margins of error (w) at the 95% confidence level were calculated. The margins of error were calculated using the s.e.m. and the critical value from the T-distribution at the 97.5th percentile ($\alpha = 0.05$; two-tailed) with n − 1 degrees of freedom (IUPAC, 2006), using equation (1). This procedure was repeated 10,000 times. The mean length of the margins of error for a given number of replicate analyses was calculated as the median (50th percentile) of the 10,000 simulations.

$$w = Tcrit_{(97.5,\,n-1)} \times \text{s.e.m.} \quad (1)$$

2.2. Analytical Errors and Calibrations

An additional Monte Carlo simulation experiment was performed to investigate how the design of a calibration experiment coupled with a constant source of analytical error can bias the calculated temperature sensitivity of the clumped isotope proxy. Specifically, we examined the effect of (1) the number of samples, (2) the temperature range that the calibration covers, and (3) the number of replicate measurements. This experiment consisted of two separate sets of Monte Carlo simulations where one variable was held constant and the two others were allowed to change. In one set of simulations, we varied the number of replicate measurements and the range of temperatures from which six equally spaced samples were drawn (Figure 1a). In another set of simulations, we changed both the number of samples and the number of replicate measurements and left the temperature range constant (Figure 1b). In all cases, we generated synthetic calibration data using the mean slope and mean intercept of the published calibration relationships (Table 2) and added analytical noise assuming a Gaussian distribution of errors of 24 ppm per replicate measurement, which is the mean external reproducibility across multiple laboratories (Table 1). A linear ordinary least squares (OLS) regression was then fitted to the data to obtain the slope (Figure 1c). This procedure was repeated 1,000 times. The 97.5th percentile of the 1,000 simulated slopes was used as the margin of error for the slope (Figure 1d).

Table 2
Published Clumped Isotope Calibrations

Publication	Slope	Absolute offset	Model error prediction	Samples	Replicates	T °C range
Henkes et al. (2013)	0.033	0.009	0.006	14	10	29
Wacker et al. (2014)	0.033	0.009	0.007	6	12	29
Katz et al. (2017)	0.033	0.009	0.014	5	8	18
Eagle et al. (2013)	0.034	0.008	0.006	11	12	30
Defliese et al. (2015)	0.035	0.007	0.005	4	8	65
Dennis and Schrag (2010)	0.036	0.006	0.005	11	4	70
Winkelstern et al. (2016)	0.037	0.005	0.003	8	4	229
Tang et al. (2014)	0.039	0.003	0.005	3	25	35
Kluge and John (2015)	0.040	0.002	0.004	8	7	68
Kluge et al. (2015)	0.040	0.002	0.002	10	5	225
Kelson et al. (2017)	0.042	0.000	0.002	11	15	74
Bonifacie et al. (2017)	0.043	0.001	0.002	11	6	327
Kele et al. (2015)	0.044	0.002	0.001	16	43	89
Came et al. (2014)	0.048	0.006	0.010	11	3	30
Affek and Zaarur (2014)	0.048	0.006	0.006	10	3	62
Tripati et al. (2015)	0.051	0.009	0.011	6	2	50
Zaarur et al. (2013)	0.053	0.011	0.007	7	3	60
Ghosh et al. (2006)	0.064	0.022	0.016	5	1	49
Mean slope	0.042 ± 0.01 (1σ)					

Note. Absolute offset is the absolute difference between a particular calibration slope and the mean slope. Samples are the average number of samples of distinct temperatures. Samples with temperatures within ±1°C are counted as replicate analyses. Replicates are the average numbers in the respective publications. Model prediction is the error on the slope for an experiment with the characteristics shown in the table.

We chose to limit our analysis to the slope because disagreements in the intercept can potentially be explained by other variables that would not affect the slope. For instance, the choice of incorrect acid fractionation corrections for Δ_{47} values can create a systematic offset between laboratories that do not affect the temperature sensitivity of clumped isotopes. Additionally, there may be other unrecognized interlaboratory biases that may not disappear unless samples are normalized to carbonate standards with agreed upon Δ_{47} values. This was recently highlighted by Spooner et al. (2016) who found agreement between Δ_{47} data produced in two different laboratories only after the data were normalized to carbonate standards and concluded that carbonate normalization may be a useful way to remove systematic interlaboratory biases. This was not done for the vast majority of published calibration data.

3. Results and Discussions

3.1. Sample Replication and Analytical Uncertainties

In Figure 2a, we show the first 100 confidence intervals from the resampling experiment using the "average sigma data set" (i.e., the data set with a standard deviation equal to the average reproducibility in Table 1). The mean values in the figure (blue and red dots) were calculated out of three replicates. We observe that for these samples the mean margin of error at the 95% CL is ±51 ppm. This corresponds to an error of ±17°C assuming a linear temperature sensitivity of 3 ppm per 1°C (Kele et al., 2015), which is only strictly true at earth surface temperatures. As expected, the 95% confidence intervals fail to capture the mean 5 times (Figure 2a red bars) when they are calculated using the critical value from the T-distribution (see methods).

In Figure 2b, we show how different external reproducibilities affect measurement precision at a given number of replicate analyses. This is done by plotting the margins of error at the 95% CL with increasing number of replicates using all three synthetic data sets. In this case, for three replicate analyses, the margins of error at the 95% CL range from ±29 ppm for the "minimum sigma data set" (i.e., the data set with a standard deviation equal to the best reproducibility in Table 1) and up to ±74 ppm for the "maximum sigma data set" (i.e., the data set with a standard deviation equal to the worst reproducibility in Table 1).

Figure 2b also shows that more precise Δ_{47} values can clearly be obtained with additional analyses. This is true for all three data sets where the precision rapidly increases for the first 10 replicates and then increases at a much slower rate. The initial fast increase in precision demonstrates that it is always worthwhile to obtain more than three replicates. However, this is not always feasible due to the expense of additional analytical time and in many cases because of the lack of enough sample material. In any case, these results indicate that clumped isotope measurement procedures where means are reported for three replicate

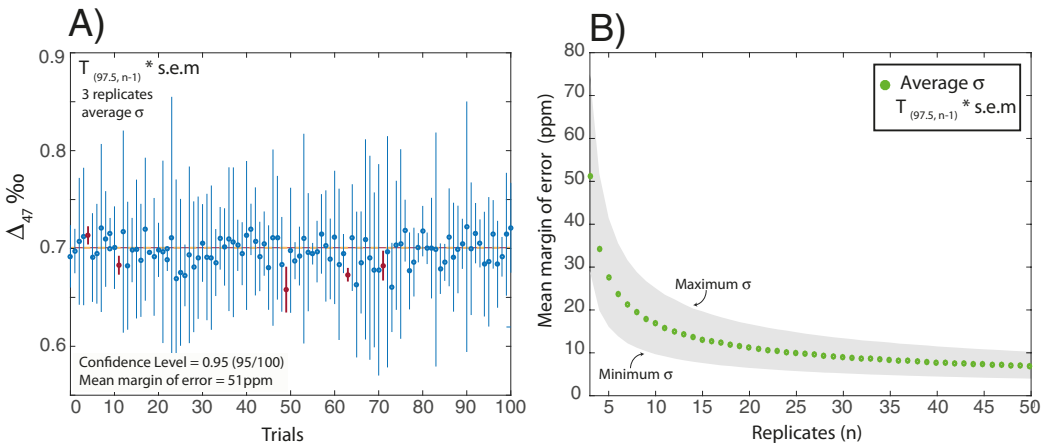

Figure 2. Results from the resampling experiment using three synthetic data sets. (a) Results of 100 trials from the "average sigma data set" with confidence intervals calculated with the s.e.m and the critical value from T-distribution at the 97.5 percentile and 2 degrees of freedom ($T_{(97.5,\ 2)} = 4.3$). Mean margin of error at the 95% CL is 51 ppm. The trials that do not capture the mean of the sample are highlighted in red. (b) Mean margins of error at the 95% CL versus number of replicate measurements for all three synthetic data sets.

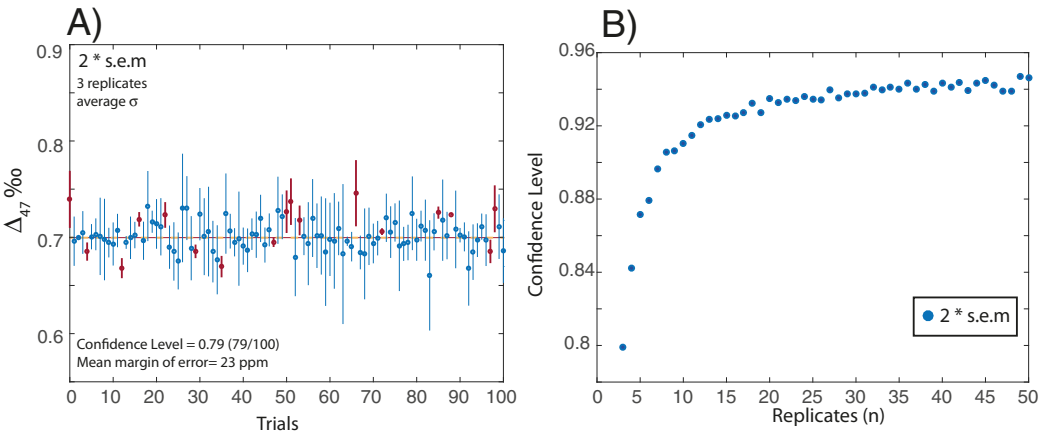

Figure 3. Results from the resampling experiment using the "average sigma data set." (a) Results of 100 trials with confidence intervals calculated as 2 times the s.e.m. Confidence level is 79% and mean length of the margin of error is 23 ppm. The trials that do not capture the mean of the sample (21) are highlighted in red. (b) Confidence level when the standard error is doubled versus number of replicate measurements.

measurements are unlikely to produce Δ_{47} estimates that are precise enough for many applications that require margins of error of $\pm$1–2°C.

Our observation on the imprecise nature of clumped isotope data also highlights how the way we report errors can be problematic. Typically, errors are reported in figures and tables as the s.e.m of three or at most five replicate measurements. This practice leads to the impression that if you double the s.e.m you can estimate the precision of a measurement at the 95% CL. This is not always correct, and it is in fact a common misconception that is frequently made across many research fields (Cumming & Finch, 2005; Huck, 2009). To illustrate this point, we calculated the confidence level when the s.e.m. is doubled using random subsamples of the "mean sigma data set" (Figure 3). For three replicate analyses, these error bars have an average margin of error of 23 ppm, which is much smaller than the 95% CL length of 51 ppm (Figure 2a). As a result, these error bars fail to capture the mean 21 times out of 100 trials (Figure 3a, red bars) rather than 5 times, as expected for a 95% CL (Figure 2a, red bars) (Figure 3a). In other words, these error bars have low confidence levels ($\sim$79% CL), which do not captures the true mean of the sample one out of 5 times. In fact, the s.e.m.'s and the margins of error at the 95% CL are related to each other by the *t* statistic, which depends on sample size (Blainey et al., 2014; IUPAC, 2006); consequently, margins of error constructed by doubling the s.e.m. only begin to approach the 95% CL when a large number of replicates (over 30) are measured (Figure 3b).

One important example where analytical errors may be relatively large is in the efforts made to calibrate the clumped isotope proxy, where in the majority of cases carbonate samples were measured only a few times. In the following section, we evaluate how a constant source of analytical error coupled to the number of replicates, the number of samples, and the range of temperatures chosen for a calibration experiment affect the robustness of the slope of the T $-$ Δ_{47} calibrations.

3.2. Calibration Experiments and Analytical Errors

The results from our simulation experiments are presented in Figure 4. In Figure 4a, we show the effect of the number of replicate measurements on the margins of error of the slope of a calibration with 6, 12, or 18 samples when the calibration data cover a temperature range of 30°C. This is a typical range of temperatures for many published calibrations. In Figure 4b, we show the effect of expanding the temperature range on the robustness of the slope. As expected, the three variables—replicates, sample number, and temperature range—exert a strong control on the precision of a clumped isotope calibration. At a given temperature range and number of samples, the largest margins of error were generated with the least precise Δ_{47} data (few replicates). Similarly, at a given temperature range and number of replicate analyses, experiments where more samples are measured always result in smaller errors (Figure 4a).

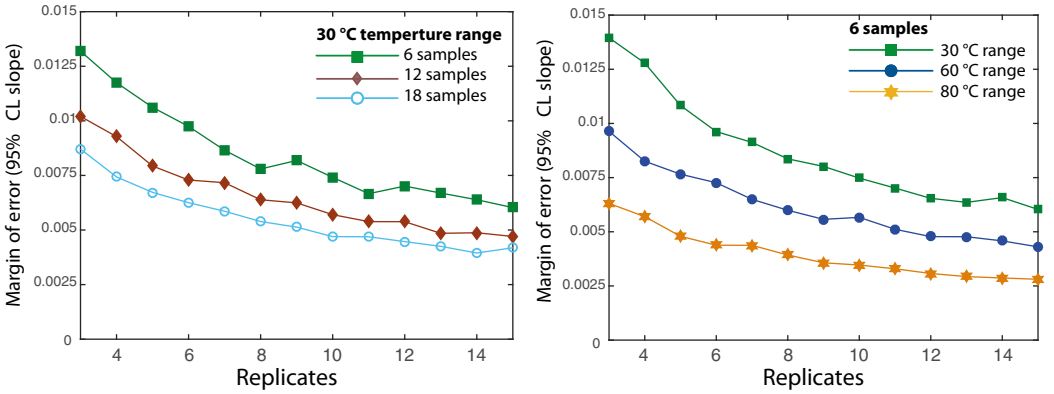

Figure 4. Results from the Monte Carlo simulation showing the effect of a constant source of analytical error (24 ppm), the number of samples, sample replication, and the temperature ranges on the uncertainties of the slope in a calibration experiment. The error on the slope is given as the margin of error at the 95% CL. (a) Effect of number of samples and number of replicates at a constant temperature range. (b) Effect of temperature range and number of replicates at a constant number of samples.

The total range of temperatures from where samples are drawn also has a large influence on the magnitude of errors (Figure 4b) with a large range strongly reducing the uncertainties. It may in fact, have the largest control on the uncertainties; although, all three variables exert some control. This is perhaps a key observation from our model, and it has important implications for the calibration of biogenic materials where the range of temperatures is necessarily small. For instance, our model suggests that foraminifera, coccolithophore, mollusc, etc., calibrations will always have the largest associated uncertainties, due to the limited temperature range of living organisms, unless a strategy to reduce the analytical error is implemented. One possible strategy is the approach taken by Katz et al. (2017), who used coccolith samples to investigate the potential occurrence of vital effects rather than to construct a calibration relationship. We believe that a similar approach should be taken for the calibration of other biogenic materials.

From these results we can conclude that clumped isotope calibrations cannot be expected to be particularly accurate if they are generated from a small number of samples taken within a narrow temperature range and with only a few replicate analyses. Moreover, these observations can be used to provide a first-order constraint on the role of analytical error in the calibration problem. For example, does the model support the hypothesis that there is a single calibration and that the published relationships are simply noise around the universal slope? If that is the case, we would expect that calibrations that are far away from the mean slope to be relatively imprecise and to have a combination of few replicates, a small number of samples, and a restricted temperature range. Conversely, we expect the opposite for relationships that are close to the mean value.

This is generally what we see in the literature (Table 2). For example, the calibrations of Ghosh et al. (2006) and Zaarur et al. (2013) are the two calibrations with the steeper slopes, and they have few samples (5–7) that were poorly replicated (1–3) within a relatively narrow temperature range (50–60°C). We should mention that the Ghosh et al. (2006) calibration may be particularly imprecise because samples were measured only once. Similarly, the calibrations of Wacker et al. (2014), Henkes et al. (2013), and Katz et al. (2017) have the shallowest slopes and meet some of the criteria that can lead to imprecise calibrations. For example, the calibrations of Wacker et al. (2014) and Katz et al. (2017) have few samples (5–6) that, although in some cases are well replicated (8–12), were obtained from a narrow temperature range (18–30°C). In the same manner, the Henkes et al. (2013) calibration has a large number of samples (14), which are in some cases well replicated (10), but with a small temperature range (30°C). On the other hand, the four calibrations that are closest to the mean—Kele et al. (2015), Kluge et al. (2015), Kelson et al. (2017), and Bonifacie et al. (2017)—meet the criteria that produces the most precise calibrations: they have many samples, in a large temperature range, and are among the ones with the best replicated samples (Table 2).

To illustrate this argument in a different way, we modified our calibration model to incorporate the characteristics of the individual calibrations experiments. This makes it possible to obtain model predictions for

the margins of error of the slope for the different calibration experiments (Table 2; model predictions). We observe a significant correlation between these predictions and how far away a particular slope is from the mean slope of the calibration experiments (Table 2, model predictions versus absolute offset; $R^2 = 0.69$, $p < 0.0001$). Again, this is what we would expect if the span of calibration slopes in the literature can be explained by analytical errors and the way calibration experiments were carried out. In other words, we see a positive correlation because the least precise calibrations also tend to be the least accurate, and because the calibrations closest to the mean are the most precise and more accurately capture the true temperature sensitivity of the clumped isotope proxy.

We should point out that our model only considers how random sources of error interact with the three variables we identified above and does not take into account the contribution that other variables such as the incorrect choice of ^{17}O correction parameters or errors in the temperatures of calibration samples may have on the slope problem. As was pointed out earlier, the ^{17}O correction parameters likely explain some of the disagreements between laboratories. For instance, Schauer et al. (2016) showed that the calibration data of Kelson et al. (2017) collapses into a single linear relationship when it is processed with a different set of parameters. However, it is unlikely that an incorrect choice of parameters can by itself explain the large range of slopes in the literature. This was pointed out by Daëron et al. (2016) who were unable to reconcile the calibration data from two different laboratories that have very different calibrations slopes after the data were recalculated with a different set of correction parameters. A similar argument was made by Katz et al., (2017) who showed that the slope of their coccolithophore calibration does not change significantly with the choice of ^{17}O correction parameters.

We believe that these observations suggest that the range of slopes reported in the literature can be largely explained by poor replication, small sample sizes, and narrow temperature ranges, and that the choice of ^{17}O parameters likely explains the remaining variability. As was pointed out in the introduction, this interpretation is not new. The potential role of these variables was initially recognized by Kelson et al. (2017), Bonifacie et al. (2017), and Katz et al. (2017). Our results here provide quantitative support for the role of these variables and allow us to identify the more (and least) statically robust slopes. For instance, we can use our model to explain why the coccolith-specific calibration of Katz et al. (2017) is so different from the Bonifacie et al. (2017) calibration that was produced in the same laboratory (Table 2). The Katz et al. (2017) calibration has few samples (five) from a very restricted temperature range (18°C) and consequently its slope is probably not very accurate (Table 2).

Although our model can be used to identify which calibrations have the slopes that are more statistically robust, at this point it is not possible to give an opinion on which calibration curve is more appropriate because that also heavily depends on the choice of an intercept. As was pointed out earlier, we purposely excluded the intercept from our analysis because we believe that there are likely systematic biases between laboratories that can potentially explain disagreements between intercepts. Taking this into account, it is not clear to what extent it is preferable for a laboratory to use their in-house calibration or a calibration with a slope that is more statistically robust. However, it should be pointed that some calibrations have intercepts that are better constrained because the three variables that we have identified here probably also affect the precision of the intercepts. In that regard, it may be advantageous to use one of the calibrations with better constrained parameters but it should be recognized that they may produce inaccurate absolute temperatures if there are large biases between laboratories.

3.3. The Role of Gas Integration Times in External Precision

Most laboratories measure gases for very long times in order to reach the shot-noise limit of an instrument (e.g., Huntington et al., 2009; Zaarur et al., 2013). Long integration times result in good internal precisions for single Δ_{47} measurements, but do not necessarily ensure good external reproducibilities (Figure 2b). This is because the majority of noise is likely added by factors outside of the precision of a single measurement (i.e., the nonpoison errors identified by Zaarur et al., 2013). Therefore, we suggest that one possible way to improve the of Δ_{47} data is to measure more replicate analyses at the expense of long ion-counting times.

The recent study of Müller et al. (2017) supports this. For instance, they obtained external reproducibilities that are comparable to laboratories that measure gases roughly 3 times longer using a single reference-gas versus sample-gas comparison with a total integration time of only 1,200 s for both sample and reference

gases. The average integration time for the other laboratories is shown in Table 1 is 3,400 s, and these times do not include the time necessary for pressure balancing and signal stabilization, which are significantly longer in the other laboratories. Because their measurements are much faster and the amount of sample needed for a replicate is only 100–120 μg as opposed to 3–10 mg, Müller et al. (2017) are able to routinely measure the 10–15 replicates that are needed for margins of error of ±10–15 ppm at the 95% CL in their system. It is important to emphasize that the comparisons we make between the external reproducibility reported by Müller et al. (2017) and the other laboratories are made on standard deviations calculated from individual digestions of CO_2, so the comparisons are made at the same scale.

Our suggestion that other laboratories should be able to shorten integration times and see a negligible change in external reproducibilities can be tested with data that is already available. For instance, it is easy to exclude aquisitions from existing data sets and then look at their effect on external precisions. This can in fact be readily done in the many laboratories that utilize Easotope, an open source clumped isotope data software package (John & Bowen, 2016). To illustrate how this can work, we excluded the last 150 s of integration time of the Müller et al. (2017) data set for both sample and reference gases (25% reduction in ion-counting times). We observe only a small change in external reproducibilities for the five carbonate standards that were measured by these authors (mean external reproducibility 26 versus 24 ppm). Note that the counting times of Müller et al. (2017) are already relatively low, so it is likely that other laboratories, which integrate for much longer, will see an insignificant change with a similar—or even larger—reduction in integration time. The instrument time gained could then be spent measuring additional replicate analyses.

3.4. How Should Errors Be Reported?

Based on the above discussion, we consider that the way in which errors on clumped isotope data are typically reported in the literature can be problematic. For instance, in some cases data may be judged to be more precise than justified, as may have occurred for calibration Δ_{47} data. In other cases, it is impossible to judge the statistical significance of error estimates because the number of replicates are not reported along with standard errors. It is not uncommon to find much smaller errors reported for samples that were measured 1 or 2 times than in samples that were measured many more times. When errors like this are included in figures it is especially difficult to judge the precision of individual data points. For these reasons, we believe that it is not statistically justifiable to report uncertainties as the standard errors of a few ($<$10) replicate measurements. These error bars are not robust, and a better alternative is to present errors as interval estimates at a specified confidence level (Blainey et al., 2014; Cumming & Finch, 2005). Error bars like this plainly show the reliability of a measurement, and they reduce the chance that uncertainties are misinterpreted because they clearly indicate precision.

We suggest that uncertainties should be reported as confidence intervals at the 95% confidence level (or their associated ± margins of error at the 95% CL); although, margins of error at the 68% CL may be preferable in some cases because this is the confidence level that is more commonly used for other geochemical data. For samples where the external reproducibility is not known, as is the case for geologic samples, CIs can be calculated using the critical value from the T-distribution (see Müller et al., 2017 for examples in geologic materials). The margin error can then be propagated into a calibration relationship (see Bonifacie et al., 2017 for a recent discussion of error propagation). This is the most conservative approach, and it can be routinely implemented with existing technology (Müller et al., 2017). However, this approach can result in large error bars if only a few replicate measurements are available, and there are other valid alternative approaches that have been successfully utilized in the clumped isotope literature.

One alternative approach was proposed by Zaarur et al. (2013), who used a set of closely related samples to estimate external reproducibilities. Another potential approach is to use the long-term external reproducibility of carbonate standards as a proxy for the variance of all carbonate measurements in a laboratory (generalized variance). This, however, assumes that samples are as reproducible, homogeneous, and free of contaminants, which can cause isobaric interferences on *m/z* 47, as the standards. And, this is probably not true for many geologic samples. Importantly, because these approaches assume that the variance is known, margins of error can be calculated using the critical values from the standard normal distribution (which do not depend on sample size; IUPAC, 2006; Moore et al., 2009) and the following equation:

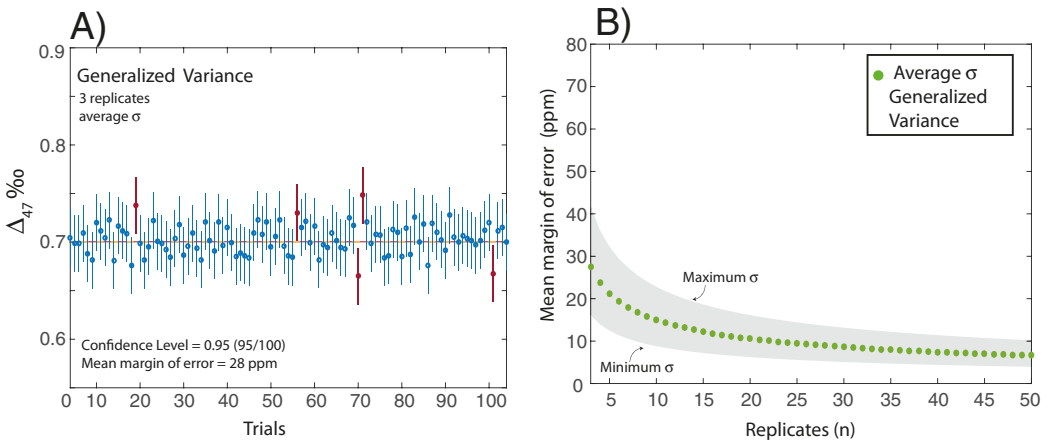

Figure 5. Results from the resampling experiment using three synthetic data sets. (a) Results of the first 100 trials from the "average sigma data set" with confidence intervals calculated using the standard deviation of the entire data set (i.e., generalized variance). Margin of error at the 95% CL is 28 ppm. The trials that do not capture the mean of the sample are highlighted in red. (b) Margins of error at the 95% CL versus number of replicate measurements for all synthetic data sets three data sets. For three replicate analyses, the margins of error range from 16 to 41 ppm.

$$w = z \times {}^{\sigma}/_{\sqrt{n}}, \tag{2}$$

where z is the critical value from the standard normal distribution, which equals 1.96 for a margin of error at the 95% CL (Moore et al., 2009), σ is the population standard deviation (e.g., the standard deviation of carbonate standards), and n is the number of replicate analyses.

To illustrate how this can work in practice, we calculated the margins of error at 95% CL using equation (2) for the means of random subsamples of the "average sigma data set" data set. The CIs were calculated with a standard deviation of 24 ppm, which is the standard deviation of the mean data set. As expected, these CIs cover the mean of the sample 95% of the time (Figure 5a), and they are smaller than the margins of error calculated using the critical values from the T-distribution (Figure 2a). Additionally, we calculated the margins of error as a function of more replicates to understand how margins of error change with more measurements. The errors were calculated for all three data sets and are shown in Figure 5b. Note that although for the first few replicates the margins of error are smaller than the errors calculated using the T-distribution (Figure 2a), this does not change the observation that more than three replicates are needed for applications that require margins of error of ±1–2°C for meaningful interpretations.

3.5. Can We Detect Changes of Just a Few Degrees Celsius in the Geological Record?

With existing technology and a reasonable amount of replicate analyses, we can expect margins of error of approximately ±3–5°C for earth surface temperature samples at the 95% CL per sample. These error bars may seem to be too large if the goal is to use two samples (one before and one after a climate transition) to detect small temperature changes in the geological record. However, this inference is not correct and, even though these error bars are still relatively large, they do not imply that we cannot detect temperature changes less than 3–5°C. The reason behind this lies in the correct interpretation behind confidence intervals. For instance, it is correct that when the margins of error (at the 95% CL) of two independent samples do not touch the difference between these samples is statistically significant. The difference is, in fact, highly significant with a statistical significance much larger than 95% (Blainey et al., 2014; Cumming & Finch, 2005). The opposite, however, is not true. When the margins of error of two samples overlap, the difference between them can still have a significance larger than 95%. In reality, for two samples with a large number of replicate analyses (i.e., more than 10), the margins of error can overlap by up to 50% and the statistical significance can still be larger than 95% (Cumming & Finch, 2005). A full explanation of how to interpret confidence intervals is beyond the scope of this manuscript, but we refer interested readers to Cumming and Finch (2005), who provide useful "rules of thumb" that can aid in their interpration.

Moreover, the previous paragraph deals only with temperature differences between two samples. In practice, however, a paleoclimate reconstruction is never done with just two samples and typically multiple

samples are collected in a stratigraphic section as a time series across a climate transition. For this reason, the feasibility of resolving small temperature changes depends not only on the error of a single Δ_{47} data point, but also on means and errors of neighboring samples. Hence, even if a single Δ_{47} data point has an error of $\pm$3–5°C at the 95% CL, it can still be possible to reconstruct much smaller temperature changes by combining several neighboring samples (i.e., before and after a transition) and evaluating the combined Δ_{47} data in a probabilistic manner.

4. Conclusions

We have shown that Δ_{47} data that results from averaging few replicate measurements are not precise enough for many applications. For instance, based on published external reproducibilities, when samples are replicated only 3 times the Δ_{47} margins of error range from $\pm$29 to 74 ppm at the 95% CL. Without considering uncertainties in the calibration relationship, these Δ_{47} uncertainties translate into margins of error of $\pm$10 to 25°C at 20°C. These errors are too large to resolve many of the climatic changes that occurred through earth's history. However, much more precise estimates—less than $\pm$10 ppm—can be routinely obtained with additional analyses.

Our results suggest that the degree of disagreement on calibration slopes is about what we should expect given the precision of Δ_{47} data and how calibration experiments were carried out. For instance, our simulations show that the precisions of calibrations are largely controlled by the number of replicate analyses, the number of samples, and the total range in temperatures from where samples are drawn. We find that the steeper and shallowest slopes in the literature were obtained in calibrations that have a combination of few replicates, a small number of samples, and a restricted temperature range. On the other hand, the opposite is true for calibrations that have slopes similar to the mean slope of all published calibrations. We believe these calibrations more accurately describe the true temperature sensitivity of the clumped isotope proxy.

We recommend that uncertainties in Δ_{47} data should no longer be reported as the standard error of a few replicate measurements. These error bars are difficult to interpret and lead to the impression that data are more accurate than warranted. Instead, we suggest that uncertainties should be reported as margins of error at a specified confidence level. For instance, margins of error at the 95 or 68% confidence level. These error bars clearly indicate the reliability of a measurement.

Acknowledgments
We would like to thank Madalina Jaggi and Stewart Bishop for sample preparation and for the installation and maintenance of our mass spectrometers. We would also like to thank Sebastian Breitenbach, Brad Rosenheim, and Ryan Venturelli for helpful discussions. We also acknowledge the helpful comments of Cedric John and two anonymous reviewers, which greatly improved our manuscript. This study was funded by the Swiss National Science Foundation project 200020_160046, 200021_143485, 200021_169849, and IZK022–160377, ETH research project ETH-33 14-1, and by Australian Research Council Australian Laureate Fellowship FL120100050. The data used in this contribution are available in the main text and in the references cited in this manuscript.

References

Affek, H. P., & Zaarur, S. (2014). Kinetic isotope effect in CO2 degassing: Insight from clumped and oxygen isotopes in laboratory precipitation experiments. *Geochimica et Cosmochimica Acta*, *143*, 319–330. https://doi.org/10.1016/j.gca.2014.08.005

Bernasconi, S. M., Hu, B., Wacker, U., Fiebig, J., Breitenbach, S. F. M., & Rutz, T. (2013). Background effects on Faraday collectors in gas-source mass spectrometry and implications for clumped isotope measurements. *Rapid Communications in Mass Spectrometry*, *27*, 603–612. https://doi.org/10.1002/rcm.6490

Blainey, P., Krzywinski, M., & Altman, N. (2014). Points of significance: Replication. *Nature Methods*, *11*, 879–880. https://doi.org/10.1038/nmeth.3091

Bonifacie, M., Calmels, D., Eiler, J. M., Horita, J., Chaduteau, C., Vasconcelos, C., . . . Bourrand, J. (2017). Calibration of the dolomite clumped isotope thermometer from 25 to 350°C, and implications for a universal calibration for all (Ca, Mg, Fe) CO_3 carbonates. *Geochimica et Cosmochimica Acta*, *200*, 255–279. https://doi.org/10.1016/j.gca.2016.11.028

Came, R. E., Brand, U., & Affek, H. P. (2014). Clumped isotope signatures in modern brachiopod carbonate. *Chemical Geology*, *377*, 20–30. https://doi.org/10.1016/j.chemgeo.2014.04.004

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170–180. https://doi.org/10.1037/0003-066X.60.2.170

Daëron, M., Blamart, D., Peral, M., & Affek, H. P. (2016). Absolute isotopic abundance ratios and the accuracy of Δ_{47} measurements, *Chemical Geology*, *442*, 83–96. https://doi.org/10.1016/j.chemgeo.2016.08.014

Defliese, W. F., Hren, M. T., & Lohmann, K. C. (2015). Compositional and temperature effects of phosphoric acid fractionation on Δ_{47} analysis and implications for discrepant calibrations. *Chemical Geology*, *396*, 51–60. https://doi.org/10.1016/j.chemgeo.2014.12.018

Dennis, K. J., Affek, H. P., Passey, B. H., Schrag, D. P., & Eiler, J. M. (2011). Defining an absolute reference frame for "clumped" isotope studies of CO_2. *Geochimica et Cosmochimica Acta*, *75*, 7117–7131. https://doi.org/10.1016/j.gca.2011.09.025, accessed 22 May 2013.

Dennis, K. J., & Schrag, D. P. (2010). Clumped isotope thermometry of carbonatites as an indicator of diagenetic alteration. *Geochimica et Cosmochimica Acta*, *74*, 4110–4122. https://doi.org/10.1016/j.gca.2010.04.005

Eagle, R. A., Eiler, J. M., Tripati, A. K., Ries, J. B., Freitas, P. S., Hiebenthal, C., . . . Roy, K. (2013). The influence of temperature and seawater carbonate saturation state on ^{13}C-^{18}O bond ordering in bivalve mollusks. *Biogeosciences*, *10*, 4591–4606. https://doi.org/10.5194/bg-10-4591-2013

Eiler, J. M. (2007). Clumped-isotope" geochemistry—The study of naturally-occurring, multiply-substituted isotopologues. *Earth and Planetary Science Letters*, *262*, 309–327. https://doi.org/10.1016/j.epsl.2007.08.020, accessed 30 August 2013

Fernandez, A., Tang, J., & Rosenheim, B. E. (2014). Siderite "clumped" isotope thermometry: A new paleoclimate proxy for humid continental environments. *Geochimica et Cosmochimica Acta*, *126*, 411–421. https://doi.org/10.1016/j.gca.2013.11.006, accessed 24 July 2014

Ghosh, P., Adkins, J., Affek, H., Balta, B., Guo, W., Schauble, E. A., . . . Eiler, J. M. (2006). $^{13}C^{18}O$ bonds in carbonate minerals: A new kind of paleothermometer. *Geochimica et Cosmochimica Acta*, *70*, 1439–1456. http://dx.doi.org/10.1016/j.gca.2005.11.014

He, B., Olack, G. A., & Colman, A. S. (2012). Pressure baseline correction and high-precision CO_2 clumped-isotope (Δ_{47}) measurements in bellows and micro-volume modes. *Rapid Communications in Mass Spectrometry*, *26*, 2837–2853. https://doi.org/10.1002/rcm.6436

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

Hu, B., Radke, J., Schlüter, H. J., Heine, F. T., Zhou, L., & Bernasconi, S. M. (2014). A modified procedure for gas-source isotope ratio mass spectrometry: The long-integration dual-inlet (LIDI) methodology and implications for clumped isotope measurements. *Rapid Communications in Mass Spectrometry*, *28*, 1413–1425. https://doi.org/10.1002/rcm.6909

Huck, S. W. (2009). *Statistical misconceptions* (1st ed.). Abingdon, UK: Routledge.

Huntington, K. W., Eiler, J. M., Affek, H. P., Guo, W., Bonifacie, M., Yeung, L. Y., . . . Came, R. (2009). Methods and limitations of "clumped" CO_2 isotope (Δ_{47}) analysis by gas-source isotope ratio mass spectrometry. *Journal of Mass Spectrometry*, *44*, 1318–1329. https://doi.org/10.1002/jms.1614

IUPAC (2006). *UPAC. Compendium of chemical terminology, (the "Gold Book")* (2nd ed.). Oxford, UK: Blackwell Science Publishing. Retrieved from: http://goldbook.iupac.org

John, C. M., & Bowen, D. (2016). Community software for challenging isotope analysis: First applications of "Easotope" to clumped isotopes. *Rapid Communications in Mass Spectrometry*, *30*, 2285–2300. https://doi.org/10.1002/rcm.7720

Katz, A., Bonifacie, M., Hermoso, M., Cartigny, P., & Calmels, D. (2017). Laboratory-grown coccoliths exhibit no vital effect in clumped isotope (Δ_{47}) composition on a range of geologically relevant temperatures. *Geochimica et Cosmochimica Acta*, *208*, 335–353. https://doi.org/10.1016/j.gca.2017.02.025

Kele, S., Breitenbach, S. F. M., Capezzuoli, E., Meckler, A. N., Ziegler, M., Millan, I. M., . . . Bernasconi, S. M. (2015). Temperature dependence of oxygen- and clumped isotope fractionation in carbonates: A study of travertines and tufas in the 6–95°C temperature range. *Geochimica et Cosmochimica Acta*, *168*, 172–192. https://doi.org/10.1016/j.gca.2015.06.032

Kelson, R., Huntington, K. W., Schauer, A. J., Saenger, C., & Lechler, A. R. (2017). Toward a universal carbonate clumped isotope calibration: Diverse synthesis and preparatory methods suggest a single temperature relationship. *Geochimica et Cosmochimica Acta*, *197*, 104–131. https://doi.org/10.1016/j.gca.2016.10.010

Kluge, T., & John, C. M. (2015). Effects of brine chemistry and polymorphism on clumped isotopes revealed by laboratory precipitation of mono- and multiphase calcium carbonates. *Geochimica et Cosmochimica Acta*, *160*, 155–168. https://doi.org/10.1016/j.gca.2015.03.031

Kluge, T., John, C. M., Jourdan, A.-L., Davis, S., & Crawshaw, J. (2015). Laboratory calibration of the calcium carbonate clumped isotope thermometer in the 25–250°C temperature range. *Geochimica et Cosmochimica Acta*, *157*, 213–227. https://doi.org/10.1016/j.gca.2015.02.028

Meckler, A. N., Ziegler, M., Millán, M. I., Breitenbach, S. F. M., & Bernasconi, S. M. (2014). Long-term performance of the Kiel carbonate device with a new correction scheme for clumped isotope measurements. *Rapid Communications in Mass Spectrometry*, *28*, 1705–1715. https://doi.org/10.1002/rcm.6949

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

Müller, I. A., Fernandez, A., Radke, J., van Dijk, J., Bowen, D., Schwieters, J., & Bernasconi, S. M. (2017). Carbonate clumped isotope analyses with the long- integration dual-inlet (LIDI) workflow: Scratching at the lower sample weight boundaries. *Rapid Communications in Mass Spectrometry*, *31*, 1057–1066. https://doi.org/10.1002/rcm.7878

Murray, S. T., Arienzo, M. M., & Swart, P. K. (2016). Determining the Δ_{47} acid fractionation in dolomites. *Geochimica et Cosmochimica Acta*, *174*, 42–53. https://doi.org/10.1016/j.gca.2015.10.029

Petersen, S. V., & Schrag, D. P. (2014). Clumped isotope measurements of small carbonate samples using a high-efficiency dual-reservoir technique. *Rapid Communications in Mass Spectrometry*, *28*, 2371–2381. https://doi.org/10.1002/rcm.7022

Schauer, A. J., Kelson, J., Saenger, C., & Huntington, K. W. (2016). Choice of ^{17}O correction affects clumped isotope (Δ_{47}) values of CO_2 measured with mass spectrometry. *Rapid Communications in Mass Spectrometry*, *30*, 2607–2616. https://doi.org/10.1002/rcm.7743

Schmid, T., & Bernasconi, S. (2010). An automated method for "clumped-isotope" measurements on small carbonate samples. *Rapid Communications in Mass Spectrometry*, *24*, 1955–1963. https://doi.org/10.1002/rcm.4598

Spooner, P. T., Guo, W., Robinson, L. F., Thiagarajan, N., Hendry, K. R., Rosenheim, B. E., & Leng, M. J. (2016). Clumped isotope composition of cold-water corals: A role for vital effects? *Geochimica et Cosmochimica Acta*, *179*, 123–141. https://doi.org/10.1016/j.gca.2016.01.023

Staudigel, P. T., & Swart, P. K. (2016). Isotopic behavior during the aragonite-calcite transition: Implications for sample preparation and proxy interpretation. *Chemical Geology*, *442*, 130–138. https://doi.org/10.1016/j.chemgeo.2016.09.013

Tang, J., Dietzel, M., Fernandez, A., Tripati, A. K., & Rosenheim, B. E. (2014). Evaluation of kinetic effects on clumped isotope fractionation (Δ_{47}) during inorganic calcite precipitation. *Geochimica et Cosmochimica Acta*, *134*, 120–136. https://doi.org/10.1016/j.gca.2014.03.005

Tripati, A. K., Hill, P. S., Eagle, R. A., Mosenfelder, J. L., Tang, J., Schauble, E. A., . . . Henry, D. (2015). Beyond temperature: Clumped isotope signatures in dissolved inorganic carbon species and the influence of solution chemistry on carbonate mineral composition. *Geochimica et Cosmochimica Acta*, *166*, 344–371. https://doi.org/10.1016/j.gca.2015.06.021

Wacker, U., Fiebig, J., Tödter, J., Schöne, B. R., Bahr, A., Friedrich, O., . . . Joachimski, M. M. (2014). Empirical calibration of the clumped isotope paleothermometer using calcites of various origins. *Geochimica et Cosmochimica Acta*, *141*, 127–144. https://doi.org/10.1016/j.gca.2014.06.004

Winkelstern, I. Z., Kaczmarek, S. E., Lohmann, K. C., & Humphrey, J. D. (2016). Calibration of dolomite clumped isotope thermometry. *Chemical Geology*, *443*, 32–38. https://doi.org/10.1016/j.chemgeo.2016.09.021

Zaarur, S., Affek, H. P., & Brandon, M. T. (2013). A revised calibration of the clumped isotope thermometer. *Earth and Planetary Science Letters*, *382*, 47–57. https://doi.org/10.1016/j.epsl.2013.07.026