ADVANCING EARTH AND SPACE SCIENCE

Geochemistry, Geophysics, Geosystems

RESEARCH ARTICLE

10.1029/2020GC009592

Special Section:
Clumped Isotope Geochemistry: From Theory to Applications

This article is a companion to Bernasconi et al. (2021), https://doi.org/10.1029/2020GC009588.

Key Points:

- We formulate a quantitative description of Δ_{47} error propagation, fully taking into account standardization errors and their properties
- These standardization errors constitute a non-negligible source of uncertainty affecting samples analyzed together in a correlated manner
- We present a new standardization approach yielding more robust/accurate corrections, and open-source implementations of these error models

Correspondence to:
M. Daëron,
daeron@lsce.ipsl.fr

Citation:
Daëron, M. (2021). Full propagation of analytical uncertainties in Δ_{47} measurements. *Geochemistry, Geophysics, Geosystems*, *22*, e2020GC009592. https://doi.org/10.1029/2020GC009592

Received 14 DEC 2020
Accepted 29 MAR 2021

Full Propagation of Analytical Uncertainties in Δ_{47} Measurements

Mathieu Daëron[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Orme des Merisiers, Gif-sur-Yvette, France

Abstract Clumped-isotope measurements in CO_2 and carbonates (Δ_{47}) present a number of technical challenges and require correcting for various sources of analytical nonlinearity. For now, we lack a formal description of the analytical errors associated with these correction steps, which are not accounted for in most data processing methods currently in use. Here we formulate a quantitative description of Δ_{47} error propagation, fully taking into account standardization errors and their properties. We find that standardization errors are highly sensitive to the isotopic compositions (δ_{47}, Δ_{47}) of unknown samples relative to the standards used for analytical corrections, and in many cases constitute a non-negligible source of uncertainty, causing true measurements errors to exceed traditionally reported error estimates by a factor of 1.5 (typically) to 3.5 (in extreme cases). Using Monte Carlo simulations based on the full InterCarb data set, we find that this model yields accurate error estimates in spite of small non-Gaussian effects which remain entirely negligible in practice. We also describe various standardization strategies, along with the assumptions they rely on, in the context of this model, and propose a new, "pooled" standardization approach designed to yield more robust/accurate corrections. Among other uses, the mathematical framework described here may be helpful to improve standardization protocols (e.g., anchor/unknown ratios) and inform future efforts to define community reference materials. What is more, these models imply that the inter-laboratory scatter ($N = 5{,}329$) observed in the InterCarb exercise (Bernasconi, Daëron, et al., 2021) can be entirely explained as the effects of current standardization procedures. Based on these findings, we recommend that future studies systematically report full analytical uncertainties taking standardization errors into account. In line with this recommendation, we provide user-friendly online resources and an open-source Python library designed to facilitate the use of these error models.

1. Introduction

Clumped-isotope geochemistry is the study of statistical anomalies in the abundance of multiply substituted isotopologues in natural materials (Eiler, 2013; Eiler & Schauble, 2004;). Mass spectrometric measurements of Δ_{47}, quantifying the excess abundance of ^{13}C -^{18}O bonds in CO_2 and, by extension, in carbonate minerals (Ghosh et al., 2006; Schauble et al., 2006), constitute the most widely used branch of this relatively young but rapidly evolving field of research. The appeal of carbonate Δ_{47} measurements is largely based on the fact that the clumped-isotope compositions of natural carbonates directly or indirectly constrain their crystallization temperatures and/or thermal histories, with a broad range of Earth science applications. Establishing a robust calibration of the carbonate Δ_{47} thermometer, however, has long remained a vexing challenge, with inter-laboratory discrepancies equivalent to large uncertainties in reconstructed temperatures, sometimes exceeding 10°C (e.g., Bonifacie et al., 2017; Petersen et al., 2019).

Keeping in mind that "true" calibration differences between certain types of carbonates are not to be excluded *a priori*, various potential causes for these discrepancies have been put forward, such as (a) inconsistent or inaccurate ^{17}O correction parameters (Daëron et al., 2016; Olack & Colman, 2019; Schauer et al., 2016), (b) systematic effects arising from different data processing methods, and (c) poorly corrected analytical biases resulting from instrumental and/or methodological differences between laboratories. Petersen et al. (2019) tested the first two of these hypotheses and found that using unified methods for ^{17}O correction and subsequent data processing reduced inter-laboratory discrepancies without eliminating them. Testing the third hypothesis is one of the goals of the recently completed inter-laboratory comparison exercise "InterCarb", whose results are reported in a companion study (Bernasconi, Daëron, et al., 2021).

How to accurately estimate the analytical uncertainties affecting Δ_{47} measurements constitutes a distinct but related issue. Compared to most other isotopic or elemental tracers, carbonate clumped isotopes stand out in that analytical uncertainties remain large relative to the range of Δ_{47} values typical of natural samples. Although Fernandez et al. (2017) pointed out that non-robust statistics based on small numbers of observations frequently yield underestimated uncertainties, there is no consensus today as to whether this is the primary cause of inter-laboratory discrepancies. He et al. (2012), Daëron & Blamart (2016) and Kocken et al. (2019) all called attention to the uncertainties associated with Δ_{47} standardization (i.e., conversion of "raw" Δ_{47} measurements to "absolute" values), but for now, we lack an explicit, formal description of this source of analytical error, which most data processing methods do not currently account for. This issue is critical in the context of the InterCarb exercise, which aims to test whether different laboratories, when analyzing a common set of four unknown and three reference carbonate samples, obtain analytically consistent results, i.e., results displaying no more inter-laboratory scatter than expected based on intra-laboratory analytical errors. The present work aims to formulate a comprehensive model of analytical errors in Δ_{47} measurements, including those arising from standardization using carbonate and/or carbon dioxide standards; to describe various standardization strategies along with the assumptions they rely on; and to provide user-friendly data processing tools implementing these error models.

2. Methods

2.1. A Brief Summary of Mass Spectrometric Measurements of Δ_{47} in Carbonates

Clumped-isotope analyses of carbonates are typically performed using dual-inlet gas-source isotope-ratio mass spectrometry. In each analysis, a certain amount of carbonate mineral reacts with pure phosphoric acid at a fixed temperature (usually 25°C, 70°C or 90°C). Each such reaction typically produces between 1 and 100 µmol of CO_2, which is collected in a series of cryogenic traps and carefully purified to eliminate isobaric contaminants (i.e., species with a molecular mass of 47 Da, or compounds liable to produce such species through fragmentation/recombination reactions). Although our primary focus is on measurements of Δ_{47} in carbonate minerals, most aspects of the present study apply just as well to Δ_{47} in CO_2 samples which were not produced by acid digestion of carbonates.

The purified CO_2 is then introduced to the "sample" side of a dual-inlet system and, from there, into a Nier-type ion source. In most systems currently in use, analyte CO_2 is compared with a "working gas" reference CO_2 through the frequent, regular toggling of a change-over valve. The bulk isotopic composition ($\delta^{13}C$, $\delta^{18}O$) and mass-47 to mass-44 abundance ratio of each analyte are determined by comparing ion currents for the analyte and the working gas, averaged over long integration times, typically tens of minutes or longer. These integration times are necessary because counting statistics are one of the primary factors limiting precision when observing rare isotopologues such as $^{16}O^{13}C^{18}O$, which makes up only 46 ppm of natural CO_2 (Huntington et al., 2009).

For the past decade Δ_{47} measurements have been standardized by comparison with specially prepared CO_2 standards with known clumped-isotope compositions and variable bulk isotope compositions (Dennis et al., 2011). Carbonate reference materials have increasingly been also used for standardization, either in addition to or as a replacement for CO_2 standards (Bernasconi, Müller, et al., 2018; Meckler et al., 2014; Schmid & Bernasconi, 2010). Although here we primarily consider standardization using carbonate reference materials, the mathematical framework presented below generally applies as well to CO_2 standards.

2.2. Terminology

We define below, in the context of this work, a number of terms. A **sample** is an amount of presumably homogeneous carbonate material subjected to one or more **analyses** (otherwise known as replicate measurements/observations). Each analysis corresponds to a single acid reaction followed by purification of the evolved CO_2 and by a series of dual-inlet IRMS measurements, yielding **working-gas delta values** (δ_{45} to δ_{49}). These working-gas deltas are then converted to "raw" (non-standardized) values of $\delta^{13}C$, $\delta^{18}O$, and Δ_{47}^{raw}. The specifics of this conversion have been extensively covered elsewhere (e.g., Daëron et al., 2016; Huntington et al., 2009), and are not directly relevant to the topics discussed here. Analyses are generally grouped into **sessions**, each of them usually corresponding to a given time span over which analytical conditions are presumed to have remained stable. One key assumption is that the various analytical/instrumental nonlinearities which affect $\Delta_{47}^{\mathrm{raw}}$ observations remain constant over the duration of each session. These nonlinearities include a **scrambling effect** likely reflecting recombination of isotopologues in the ion source or elsewhere in the sample preparation apparatus (Dennis et al., 2011); a **compositional slope** reflecting small biases in the electrical background of the ion beam measurements (Bernasconi, Hu, Wacker, et al., 2013; He et al., 2012); and a **working gas offset** resulting from the (knowingly inaccurate) assumption that the dual-inlet working gas is stochastic. Within each session, the samples/analyses are divided into two groups: **anchors**, whose Δ_{47} values are assigned *a priori*, and **unknowns**, whose Δ_{47} values are to be determined. Here we define a **standardization model** as any mathematical procedure aiming to estimate these unknown Δ_{47} values by comparing the anchor and unknown analyses, explicitly or implicitly constraining analytical nonlinearities within each session.

2.3. Objectives and Strategy

Our aim is to model how random, zero-centered, presumably Gaussian measurement errors on δ_{47} or $\Delta_{47}^{\mathrm{raw}}$ propagate into final, "absolute" Δ_{47} values averaged over a number of analyses/sessions. We do not attempt to account for non-random biases such as those potentially arising, for instance, from errors in the isotopic composition of the working gas, or from assigning inaccurate Δ_{47} values to one or more anchors. The models described here will hopefully provide a framework to report more accurate estimates of the uncertainty associated with clumped-isotope measurements, and inform the choices we make in the laboratory.

We start by describing a general formulation of the standardization function used to compute the "absolute" Δ_{47} value of each analysis. Quantifying the parameters defining this function within a given session is equivalent to constraining the analytical/instrumental nonlinearities mentioned above, and may be treated as a classical least squares minimization problem.

We follow up by estimating the analytical precision of "raw" measurements (before standardization) based, following oft-repeated recommendations, on the pooled external repeatability of a group of standards and/or unknown samples. The general formulation used here then makes it straightforward to propagate the raw measurement errors into the "autogenic" uncertainty of each analysis (that directly arising from the raw errors of this particular analysis) and an independent component of "allogenic" uncertainties arising from the least squares model errors, i.e., from the standardization itself.

We finish by describing the general properties of these two components of error, and briefly discuss several practical standardization approaches applicable to real-world data sets. With non-specialist readers in mind, we attempted, as much as possible, to leave mathematical details out of the main text, but three appendices provide detailed, explicit examples of the calculations underlying our models.

2.4. Standardization to an "Absolute" Δ_{47} Reference Frame Within a Single Session

Computing "absolute" Δ_{47} values traditionally involves two consecutive affine transformations designed to correct for known instrumental nonlinearities (Equations 5 and 6 of Dennis et al. (2011), using the original notation):

$$\Delta_{47\text{-[SGvsWG]}0} = \Delta_{47\text{-[SGvsWG]}} - \delta^{47}_{\text{[SGvsWG]}} \times \mathrm{Slope}_{\mathrm{EGL}} \tag{1}$$

$$\Delta_{47\text{-RF}} = \Delta_{47\text{-[SGvsWG]}0} \times \mathrm{Slope}_{\mathrm{ETF}} + \mathrm{Intercept}_{\mathrm{ETF}} \tag{2}$$

This is mathematically equivalent to the following formulation:

$$\Delta_{47}^{\mathrm{raw}} = a\,\Delta_{47} + b\,\delta_{47} + c \tag{3}$$

In this equation, the parameters (a, b, c) respectively account for scrambling effects, the compositional slope, and the working gas offset. To estimate these parameters, a natural approach is to use classical least squares minimization methods, treating $\Delta_{47}^{\mathrm{raw}}$ as the response/dependent variable and (Δ_{47}, δ_{47}) as explanatory variables. Despite uncertainties on δ_{47} usually being as large as those on $\Delta_{47}^{\mathrm{raw}}$, the former may safely be treated as an explanatory variable because *b* is typically small enough (10^{-2} or less) for errors on δ_{47} to have a negligible impact. As an aside, even in cases where $|b|$ is so small as to be indistinguishable from zero, it remains important, as argued below, to quantify the precision of this estimate. The models discussed here are thus fully consistent with background correction procedures such as the "pressure baseline correction" of He et al. (2012), and neither approach should preclude the other.

Without compelling reasons to do otherwise, we assign equal weights to all measurements belonging to the same session. The best fit standardization parameters (a, b, c) for any given session are thus those minimizing the following χ^2 statistic, summed over all anchor analyses within that session (unknown analyses are not considered here because their Δ_{47} values are not known *a priori*):

$$\chi^2 = \sum\left(\Delta_{47}^{\mathrm{raw}} - a\,\Delta_{47} - b\,\delta_{47} - c\right)^2 \tag{4}$$

This computation, whose underlying mathematical steps and details are summarized in Appendix A, yields a triplet of best fit values for (a, b, c), thus defining the standardization function of Equation 3 for this session. It also yields a covariance matrix V_0 for the best fit values of (a, b, c). At this stage, the covariance matrix is unscaled, meaning that it only constrains the *relative* scaling between model standard errors and covariances in (a, b, c). The additional piece of information needed to scale these model errors is the uncertainty assigned to each observation, i.e., the analytical precision of individual $\Delta_{47}^{\mathrm{raw}}$ measurements.

2.5. Estimating the Analytical Precision of Raw Measurements

The uncertainty assigned to individual $\Delta_{47}^{\mathrm{raw}}$ measurements, noted $\sigma_{47}^{\mathrm{raw}}$, may be quantified in various ways, but always keeping in mind that over-reliance on the statistics of small numbers is problematic (Fernandez et al., 2017). We propose that in most cases a robust estimate of $\sigma_{47}^{\mathrm{raw}}$ can be obtained by considering carbonate samples deemed free of contaminants and isotopically homogeneous, be them anchors, unknown samples, or carbonate standards treated as unknowns, and computing the pooled Δ_{47} repeatability of analyses within this group:

$$\sigma_{47}^2 = \frac{1}{\mathrm{N_a} - \mathrm{N_S}}\sum\left(\Delta_{47} - \overline{\Delta_{47}}\right)^2 \tag{5}$$

where $\overline{\Delta_{47}}$ is the average Δ_{47} value for the sample considered, N_{a} the total number of analyses considered, and N_{S} the number of different samples considered. The $\Delta_{47}^{\mathrm{raw}}$ repeatability of analyses is then:

$$\left(\sigma_{47}^{\mathrm{raw}}\right)^2 = \left(a\,\sigma_{47}\right)^2 = \frac{1}{\mathrm{N_a} - \mathrm{N_S}}\sum\left(\Delta_{47}^{raw} - a\,\overline{\Delta_{47}} - b\,\delta_{47} - c\right)^2 \tag{6}$$

It bears noting that if the group of samples used to estimate $\sigma_{47}^{\mathrm{raw}}$ within a single session only comprises three anchors, then $\left(\sigma_{47}^{\mathrm{raw}}\right)^2$ is equal to the reduced chi-square statistic $\chi^2/(N_{\mathrm{a}} - 3)$ for that session. In such case, scaling the standardization errors by $\sigma_{47}^{\mathrm{raw}}$ is equivalent to the common practice of estimating least squares model errors based on the scatter/variance of residuals. Taking additional samples into account increases confidence in our estimate by virtue of increasing the statistical degrees of freedom ($N_{\mathrm{f}} = N_{\mathrm{a}} - N_{\mathrm{S}}$), on the condition that the replicability of these additional samples is equal to (or indistinguishable from) that of carbonate standards. In our experience, this condition is frequently met when samples are well-mixed, finely ground, relatively pure carbonate powders.

2.6. Propagation of Standardization Errors Within a Single Session

Regardless of its estimation method, $\sigma_{47}^{\mathrm{raw}}$ may now be used to quantify the standard model errors (σ_a, σ_b, σ_c) on the best fit standardization parameters and their covariances (c_{ab}, c_{bc}, c_{bc}):

$$\begin{bmatrix} \sigma_a^2 & c_{ab} & c_{ac} \\ c_{ab} & \sigma_b^2 & c_{bc} \\ c_{ac} & c_{bc} & \sigma_c^2 \end{bmatrix} = \left(\sigma_{47}^{\mathrm{raw}}\right)^2 \mathrm{V}_0 \tag{7}$$

Here, the σ values quantify the precision of the constraints obtained on each of the best fit model parameters considered independently, while the covariances indicate the statistical correlation between these model errors (e.g., c_{ab} is the product of $\sigma_a\sigma_b$ and the dimensionless correlation coefficient between best fit values of a and b). These model errors and covariances fully describe the standardization uncertainty associated with anchor measurement errors, and can now be propagated explicitly using classic propagation methods (e.g., Tellinghuisen, 2001) to the session average Δ_{47} value of a given unknown sample, noted $\overline{\Delta_{47}}$:

$$\overline{\Delta_{47}} = \frac{1}{a}\overline{\Delta_{47}^{raw}} - \frac{b}{a}\overline{\delta_{47}} - \frac{c}{a} \quad \Rightarrow \quad \sigma\left(\overline{\Delta_{47}}\right)^2 = \mathrm{J} \times \mathrm{C} \times \mathrm{J}^{\mathrm{T}} \tag{8}$$

with $\overline{\Delta_{47}^{\mathrm{raw}}}$ and $\overline{\delta_{47}}$ being the session average values of $\Delta_{47}^{\mathrm{raw}}$ and δ_{47}, respectively; J the Jacobian matrix of $\overline{\Delta_{47}}$ (i.e., the matrix of all partial derivatives of $\overline{\Delta_{47}}$); J^{T} the transpose of J; and C the covariance matrix of $\left(\overline{\Delta_{47}^{\mathrm{raw}}}, a, b, c\right)$:

$$\mathrm{J} = \left[\frac{\partial\overline{\Delta_{47}}}{\partial\overline{\Delta_{47}^{\mathrm{raw}}}}, \frac{\partial\overline{\Delta_{47}}}{\partial a}, \frac{\partial\overline{\Delta_{47}}}{\partial b}, \frac{\partial\overline{\Delta_{47}}}{\partial c} \right] = \frac{1}{a}\left[1, -\overline{\Delta_{47}}, -\overline{\delta_{47}}, -1 \right] \tag{9}$$

$$\mathrm{C} = \begin{bmatrix} \sigma\left(\overline{\Delta_{47}^{\mathrm{raw}}}\right)^2 & 0 & 0 & 0 \\ 0 & \sigma_a^2 & c_{ab} & c_{ac} \\ 0 & c_{ab} & \sigma_b^2 & c_{bc} \\ 0 & c_{ac} & c_{bc} & \sigma_c^2 \end{bmatrix} = \begin{bmatrix} \left(\sigma_{47}^{\mathrm{raw}}\right)^2 / \mathrm{N_a} & 0 & 0 & 0 \\ 0 & \sigma_a^2 & c_{ab} & c_{ac} \\ 0 & c_{ab} & \sigma_b^2 & c_{bc} \\ 0 & c_{ac} & c_{bc} & \sigma_c^2 \end{bmatrix} \tag{10}$$

The structure of the above covariance matrix makes it clear that $\sigma\left(\overline{\Delta_{47}}\right)^2$ for an unknown sample is the sum of two statistically independent sources of error: an "autogenic" component σ_{u} reflecting uncertainties in $\Delta_{47}^{\mathrm{raw}}$ measurements for that sample, and an "allogenic" component σ_{s} reflecting uncertainties in the standardization model used to convert $\Delta_{47}^{\mathrm{raw}}$ to final Δ_{47} values:

$$\sigma\left(\overline{\Delta_{47}}\right)^2 = \sigma_{\mathrm{u}}^2 + \sigma_{\mathrm{s}}^2 \tag{11}$$

$$\sigma_{\mathrm{u}}^2 = \sigma_{47}^2 / \mathrm{N_a} \tag{12}$$

$$\sigma_{\mathrm{s}}^2 = \frac{1}{a^2}\left(\overline{\Delta_{47}}^2 \sigma_a^2 + \overline{\delta_{47}}^2 \sigma_b^2 + \sigma_c^2 + 2\left(\overline{\Delta_{47}}\,\overline{\delta_{47}}\, c_{ab} + \overline{\Delta_{47}}\, c_{ac} + \overline{\delta_{47}}\, c_{bc} \right) \right) \tag{13}$$

2.7. Combining Data From Several Independent Sessions

As long as the standardization of each session only takes into accounts analyses from that session, the values of $\overline{\Delta_{47}}$ computed as above within each session are statistically independent from each other. The final Δ_{47} value for a given unknown sample may thus be simply computed as the weighted average of $\overline{\Delta_{47}}$ from different sessions (with weights noted ω). Using a weighted average for this last step is necessary to account for inter-session differences in the number of analyses of that sample, and also potentially in raw analytical repeatability (unless, for instance, a deliberate choice is made to use a single estimate of σ_{47}^{raw} constrained by all sessions):

$$\Delta_{47}^{\text{final}} = \sum_i \omega_i \left(\overline{\Delta_{47}}\right)_i \quad \text{with sessions noted as } i \tag{14}$$

$$\omega_i = \sigma\left(\overline{\Delta_{47}}\right)_i^{-2} \Big/ \sum_i \sigma\left(\overline{\Delta_{47}}\right)_i^{-2} \tag{15}$$

$$\sigma\left(\Delta_{47}^{\text{final}}\right)^2 = \sum_i \omega_i^2\, \sigma\left(\overline{\Delta_{47}}\right)_i^2 = 1 \Big/ \sum_i \sigma\left(\overline{\Delta_{47}}\right)_i^{-2} \tag{16}$$

3. Discussion

3.1. Properties of Standardization Errors

The standardization model of Section 2.4 is mathematically equivalent to the least squares fitting of a two-dimensional plane described by Equation 3 in a three-dimensional space (δ_{47}, Δ_{47}, Δ_{47}^{raw}). Most properties described below arise naturally from this geometry.

Standardization uncertainties depend greatly on the bulk (δ_{47}) and clumped-isotope (Δ_{47}) composition of unknown samples relative to the anchor samples (Daëron & Blamart, 2016; Kocken et al., 2019). It is thus useful to describe this uncertainty in terms of an "error field" which can be mapped in (δ_{47}, Δ_{47}) space, as shown in Figure 1. The minimum standardization error coincides, in (δ_{47}, Δ_{47}) space, with the barycenter of the anchor analyses, and its value is equal to $\sigma_{47} / \sqrt{N}$, with N being the total number of anchor analyses.

Outside of a polygon defined by the anchor samples, standardization errors increase steeply. As illustrated in Figure 1, this increase is comparatively slower if analyses are evenly distributed between anchor samples, which tightens constraints on parameters a and b.

Figure 1 also illustrates the benefits of using anchors with extreme isotopic compositions, which increase the area of the anchor polygon. One potential drawback of relying on isotopically extreme anchors, however, is that our "planar" model approximation might then break down. For instance, a small quadratic component to the compositional nonlinearity (term $b\ \delta_{47}$ in Equation 3), whose effect would be negligible over a δ_{47} range of 30‰, might introduce a significant bias over a range of 60 or 100‰ (e.g., Figure 7 from He et al., (2012)).

The properties outlined above are fully consistent with the Monte Carlo simulations of Kocken et al. (2019). In particular, they explain all of the main patterns displayed in their Figure 5. The primary difference between our approach and that of Kocken et al., beyond the difference in mathematical methods, is that their simulations focus on the empirical transfer function (Equation 2), which corresponds to parameters a and c. Here we show that the uncertainty from compositional nonlinearities (Equation 1) behaves in a similar way, and that all of these corrections can be propagated explicitly in a unified manner, side-stepping the need for Monte Carlo simulations.

3.2. Impact of Standardization Errors

Because the relative contributions of the autogenic and allogenic error components defined above (Equations 11–13) are sensitive to the distribution of analyses among anchor and unknown samples and on the isotopic composition of unknowns relative to the anchor polygon, they are expected to vary greatly between laboratories and/or sessions. The InterCarb data set, comprising over five thousand analyses from 22 different laboratories (Bernasconi, Daëron, et al., 2021), offers an excellent opportunity to quantify these two components in a wide range of realistic settings.

A compilation of σ_s versus σ_u for the average Δ_{47} value of unknown samples obtained in each of the 77 InterCarb sessions is shown in Figure 2. As expected, standardization errors for IAEA-C1, a marble sample

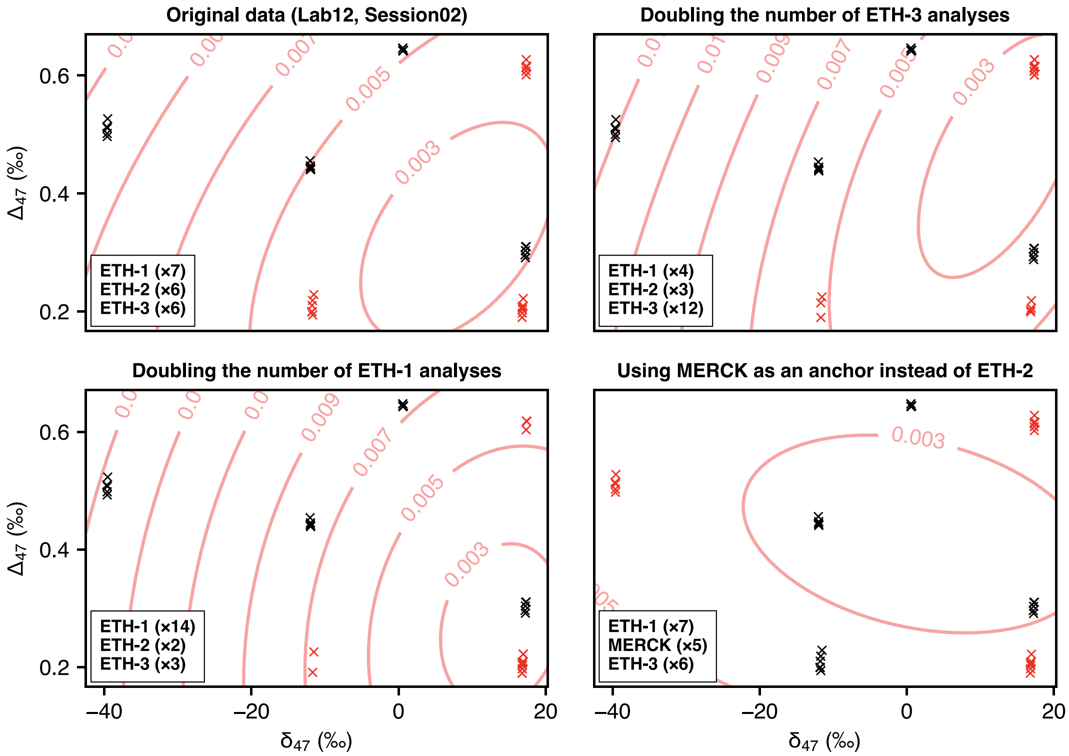
Original data (Lab12, Session02)
Doubling the number of ETH-3 analyses
Doubling the number of ETH-1 analyses
Using MERCK as an anchor instead of ETH-2

Figure 1. Properties of standardization errors. Upper left panel shows the unknown and anchor analyses (black and red crosses, respectively) and contours of the standardization error field (red lines) for Session #2 of Lab #12 in the InterCarb data set. Upper right and lower left panels modify the original data by changing the distribution of anchor analyses between ETH-1, ETH-2, and ETH-3, keeping the total number of anchor analyses constant, illustrating that the error minimum coincides in (δ_{47}, Δ_{47}) space with the barycenter of anchor analyses. The lower right panel corresponds to the original data but treats ETH-2 as an unknown and MERCK as an anchor, illustrating the benefits of using isotopically extreme anchors.

which plots within the anchor polygon defined in (δ_{47}, Δ_{47}) space by ETH-1/2/3, are generally slightly smaller than autogenic errors, resulting in a modest increase of the total Δ_{47} error (σ_{47}) relative to the autogenic error. Samples IAEA-C2 (natural travertine) and ETH-4 (synthetic calcite), both of them located outside of the anchor polygon, display larger standardization errors, thereby increasing σ_{47} by an average factor of 1.5 and up to a factor of 2. Finally, in the case of the MERCK sample, a synthetic carbonate with extremely depleted δ^{13}C and δ^{18}O values, standardization errors generally dominate. As a result, propagating them into the total Δ_{47} error increases σ_{47} by an average factor of 2.5 and up to a factor of 4.

3.3. Correlations Between Samples

Standardization errors contribute a sizable portion of analytical uncertainties, but it is notable that they do so in a way that is strongly correlated between samples, as illustrated by the joint 95% confidence ellipses for the average Δ_{47} values of unknown samples shown in Figure 3. As a result, Δ_{47} measurements of samples analyzed in one or more common sessions are not independent measurements. In many cases, this precludes using simple statistics such as the widely used formula for calculating the standard error of the average of replicate measurements, which assumes independent measurement errors.

Appendix B provides full computational details for the covariance between the session-averaged Δ_{47} values of two unknowns samples (B.1); the uncertainties characterizing Δ_{47} differences between samples (B.2); and weighted mean Δ_{47} values averaged over several samples (B.3). The key point to keep in mind is that full analytical errors are not independent between samples of the same session, with the following consequences: (1) when averaging many Δ_{47} measurements within a single session, analytical errors will not tend to zero but to the standardization error for this sample; (2) the error on Δ_{47} differences between samples of similar

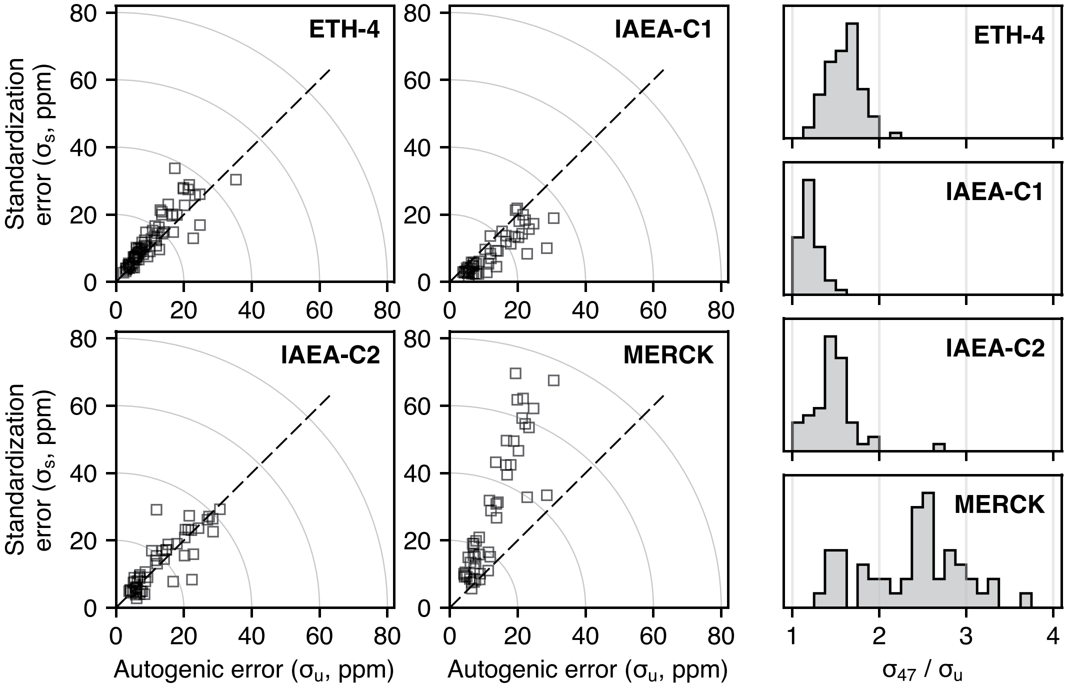

Figure 2. Autogenic vs. standardization errors. Each square marker corresponds to the error components for the average Δ_{47} value of an unknown sample in each of the InterCarb sessions. Histograms characterize the ratios of total analytical error (σ_{47}) to autogenic error (σ_u) for each of the unknown samples.

compositions (as is often the case in paleoclimate records) is largely unaffected by standardization errors, but only if they are analyzed within the same session.

3.4. Gaussian Approximation of Standardization Errors

The error propagation formula of Equation 8 is a first-order Taylor approximation. Because Δ_{47} is not a linear combination of (Δ_{47}^{raw}, δ_{47}, a, b, c), propagated errors in Δ_{47} are not strictly Gaussian. However, after quantifying the non-Gaussian effects of these approximations using Monte Carlo simulations of the full InterCarb data set (Appendix C), we find that these deviations from normality remain entirely negligible in practice, with Gaussian estimates of mean Δ_{47} values and their corresponding standard errors being typically off, respectively, by only 0.02σ and 0.01σ (with σ denoting Gaussian estimates), considering all sessions and all unknown samples in the InterCarb data set.

As an extreme example, Figure 4 shows the Monte Carlo distribution of full analytical errors for the average Δ_{47} value of IAEA-C2 in one of the InterCarb sessions, chosen because it is the "least Gaussian" distribution of the whole data set, i.e., the least likely to be Gaussian based on a Kolmogorov-Smirnov test ($p = 0.0003$). Even in this worst-case example, differences between the Monte Carlo cumulative distribution function (CDF) and the Gaussian CDF computed from Equation 8 remain minuscule: the Monte Carlo average of Δ_{47} for this sample in this session is 0.6734 ‰ (vs. 0.6713‰ for the Gaussian estimate), and the corresponding Monte Carlo standard error is 0.0357‰ (vs. 0.0348‰ for the Gaussian approximation, noted σ), off by -0.06σ and -0.03σ respectively.

3.5. Statistical Weighting Options

For the sake of simplicity, the error model described above rests on simple assumptions, for example by assigning equal statistical weights to all analyses. In the following sections, we briefly discuss various ways in which this error model could be modified to better reflect real-life analytical conditions.

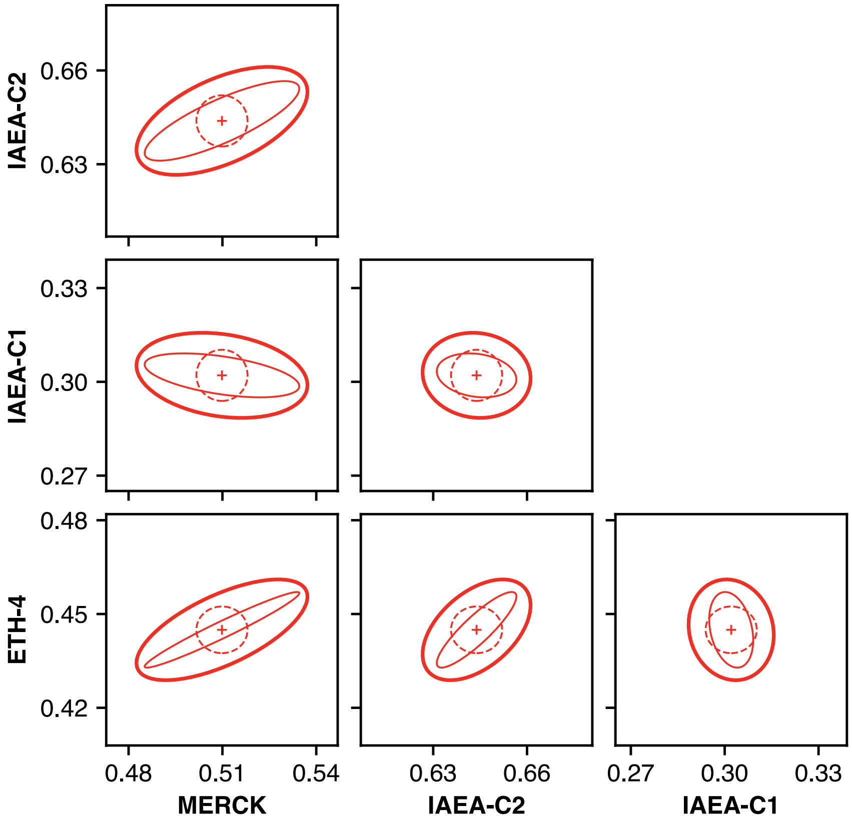

Figure 3. Covariance of errors in session-averaged Δ_{47} values for unknowns samples. Thick red lines correspond to joint 95% confidence ellipses for the average Δ_{47} values of each unknown sample in Session02 of Lab12 (cf. upper left panel of Figure 1). Thin red lines and dashed red lines correspond to joint 95% confidence ellipses only taking into account standardization and autogenic errors, respectively.

3.5.1. Equal Session Weights

In the general case where all sessions are considered equal, we recommend that each session should first be standardized using Equations 21 and 23. The *overall* $\Delta_{47}^{\mathrm{raw}}$ repeatability should then be computed using a slightly modified version of Equation 6, with N_A being the number of anchor samples, N_U the number of unknown samples, and N_a the total number of analyses:

$$\left(\sigma_{47}^{\mathrm{raw}}\right)^2 = \frac{1}{\mathrm{N_a} - \mathrm{N_A} - \mathrm{N_U}} \Sigma\left(\Delta_{47}^{\mathrm{raw}} - a\,\overline{\Delta_{47}} - b\,\delta_{47} - c\right)^2 \quad (17)$$

This overall repeatability should then be used to scale the covariance matrix of each session according to Equation 7.

3.5.2. Different Session Weights

It may be justified in some cases to assign different statistical weights to analyses from different sessions. We would not generally recommend doing so based only on observed differences in σ_{47} (which will inevitably vary slightly between sessions), unless these differences are statistically significant with a high level of confidence. On the other hand, data produced under different analytical conditions may in some cases reasonably be expected to be more or less precise: for example, measurements obtained using greater ion currents should be more precise due to counting statistics alone.

In such cases, we may first divide sessions into groups expected to share similar analytical precision levels. Pooled $\Delta_{47}^{\mathrm{raw}}$ repeatabilities for each group may then be computed according to Equation 17, and subsequently applied to covariance matrix scaling according to Equation 7.

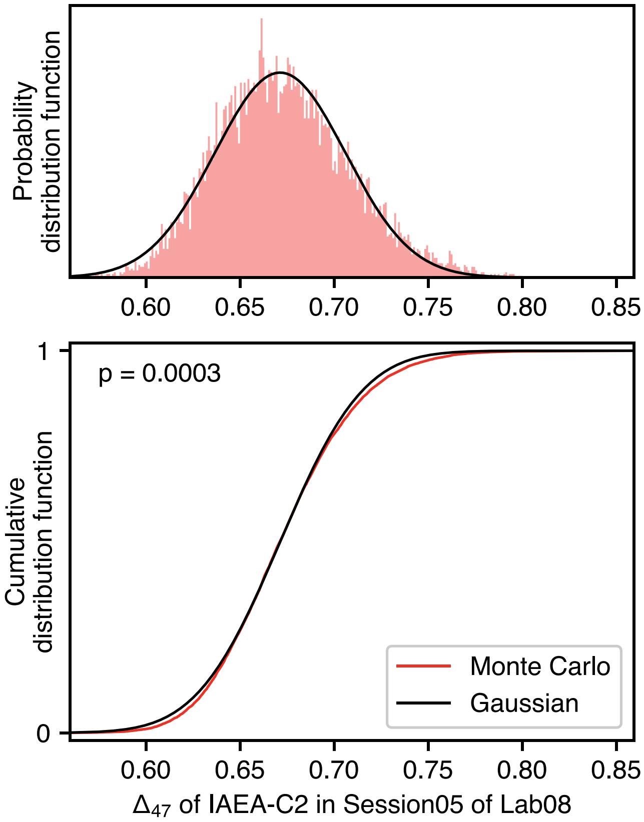

Figure 4. Error distribution for the "least Gaussian" average Δ_{47} value in the InterCarb data set. Upper panel: red area corresponds to the Monte Carlo histogram of the average Δ_{47} value of IAEA-C2 in Session #5 of Lab #8, black line is the Gaussian probability distribution computed from Equations 11–13. Lower panel: Monte Carlo (red line) and Gaussian (black line) cumulative distributions functions for this average Δ_{47} value. In the upper left corner, *p* is the Kolmogorov-Smirnov *p*-value for the null hypothesis that these two distributions are identical.

3.6. Pooled Standardization Model Taking Unknown Samples into Account

3.6.1. Principle

By only considering anchor samples to constrain the standardization parameters (a, b, c) of each session, the models described so far neglect some useful information. As a matter of fact, even without prior knowledge of the Δ_{47} values of unknown samples, we expect the relative mapping of anchors and unknowns in (δ_{47}, Δ_{47}) space to be preserved between sessions. This approach is only useful when some of the sessions have unknown samples in common, but in that case it is likely to substantially increase the number of observations constraining the standardization model, making it more robust (less sensitive to outliers in the anchor analyses) and slightly more precise (by virtue of increasing the model's degrees of freedom).

In practice, instead of treating each session as a separate least squares problem, we now aim to minimize a "pooled" version of the χ^2 statistic defined in Equation 4, this time summed over all analyses (including both anchors and unknowns) in all sessions considered:

$$\chi^2 = \sum\left(\Delta_{47}^{\mathrm{raw}} - a_i\,\Delta_{47} - b_i\,\delta_{47} - c_i\right)^2 \tag{18}$$

where $\Delta_{47}^{\mathrm{raw}}$ and δ_{47} are the observations from each analysis, (a_i, b_i, c_i) are the standardization parameters for session (i), and Δ_{47} is either a nominal value assigned *a priori* (for anchor analyses) or an additional, free model parameter corresponding to the Δ_{47} value of the relevant unknown sample. The pooled regression model now rests on a number of observations equal to the total number of analyses, with a number of model parameters equal to the number of unknown samples plus three times the number of sessions.

Because some of the χ^2 terms include the product of two model parameters, this is not a linear least squares problem and the direct solution of Appendix A no longer applies. One may, however, call upon well-established numerical approaches designed to optimize nonlinear problems. In our experience, the classical Levenberg-Marquardt method (Levenberg, 1944; Marquardt, 1963), as implemented by the LMFIT Python package (Newville et al., 2014), is well suited to this task. Even for large data sets of several thousand analyses, it is able to quickly and reliably output a vector of best fit values for all model parameters (including Δ_{47} values for all unknown samples) along with the corresponding covariance matrix, thus directly providing standard errors and covariances between unknown sample Δ_{47} values.

3.6.2. Benefits

The benefits of a pooled standardization model may not be immediately obvious, but this approach should yield systematic improvements in the robustness and accuracy of the standardization procedure. For instance, considering the samples shown in Figure 1, it may be clear that forcing the Δ_{47} value of MERCK to remain consistent between sessions should greatly contribute to constrain variations in the compositional slope (b) between sessions, even without knowing MERCK's true composition. The same argument could be made if one were to analyze heated and equilibrated gases along with carbonate standards, treating them as entirely unknown samples: even without any knowledge of CO_2 equilibrium values nor of acid fractionation effects, the large spread of Δ_{47} between heated and equilibrated gases would strongly constrain variations of the scrambling factor (a) between sessions, thereby reducing standardization errors for all samples.

Figure 5 illustrates this reduction in standardization errors by showing (δ_{47}, Δ_{47}) plots for the four sessions from Lab #12 in the InterCarb data set, comparing the error fields resulting either from the pooled standard-

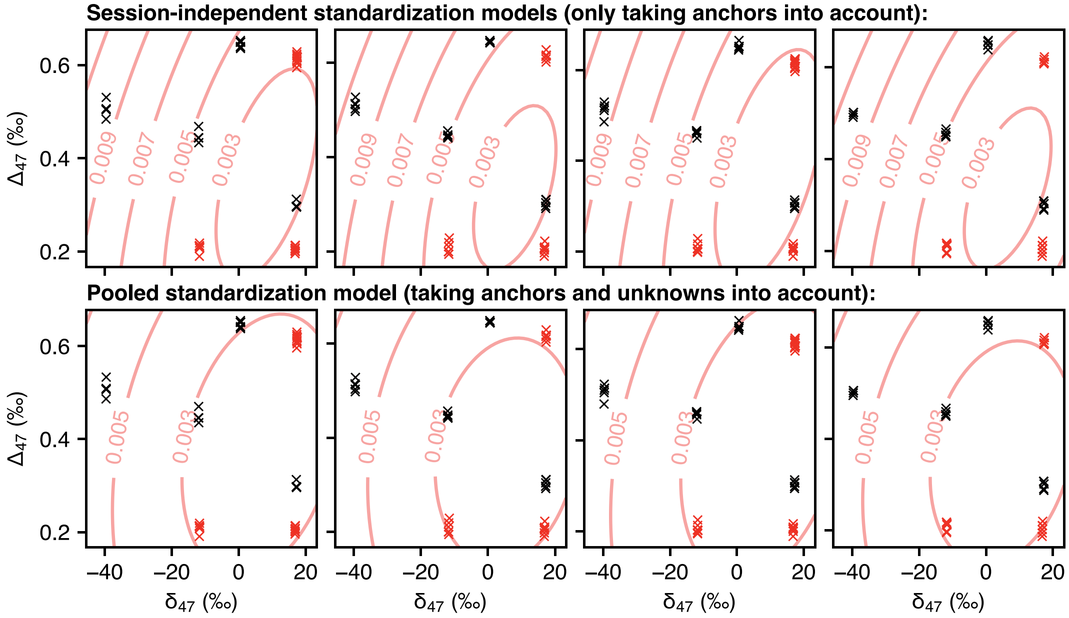

Figure 5. Benefits of a pooled standardization model. Unknown and anchor analyses (black and red crosses, respectively) and contours of the standardization error fields (red lines) for the four sessions of Lab #12 in the InterCarb data set. Upper row: using four independent models only taking anchor analyses into account, with 20, 16, 24, and 16 degrees of freedom, respectively. Lower row: using a pooled standardization model with 153 degrees of freedom taking anchors and unknowns into account as described in Section 3.6.

ization approach (one model with 153 degrees of freedom) or from the earlier approach ignoring unknown samples (four independent models with 20, 16, 24, and 16 degrees of freedom, respectively). These statistical improvements are not a result of over-fitting, despite the increase from 12 to 16 model parameters, because the number of observations used to compute the χ^2 statistic increases even more, from 88 to 169. Although the locations and values of the error field minima remain largely unaffected by the choice of standardization method, in this case the pooled model strongly reduces standardization errors for analyses plotting outside of the anchor polygon (from 10–11 ppm down to 6 ppm for MERCK), despite the fact that no assumption was made regarding the true Δ_{47} values of unknown samples. It should be noted, however, that uncertainties on final, average Δ_{47} values tend not to be as greatly reduced as those on (a, b, c), reflecting the fact that the pooled regression approach is primarily designed to improve accuracy rather than precision.

3.6.3. Caveat

The pooled approach depends critically on our earlier assumption that samples are homogeneous, which we acknowledge to be generally but not universally true. It is however simple enough, in the presence of samples suspected to be heterogeneous (i.e., whose Δ_{47} repeatability is demonstrably worse than for carbonate standards with a statistically high level of confidence), to treat each of the corresponding analyses as belonging to separate samples.

4. ClumpyCrunch and D47crunch

The calculations discussed above may be tedious to implement from scratch. The simplest way to take advantage of these error models is to use the latest version of the open-source ClumpyCrunch web application (https://clumpycrunch.pythonanywhere.com), which implements both the independent-sessions method of Section 2.6 and the pooled standardization approach of Section 3.6. Those wishing to experiment at a deeper lever may install the underlying, open-source D47crunch library for Python (https://doi.org/10.5281/zenodo.4314550), which also supports computing different repeatabilities for different groups of sessions (Section 3.5.2); explicitly treating some samples as potentially inhomogeneous (Section 3.6); modeling temporal drifts in parameters a, b, c (Appendix A.2); computing standard errors for Δ_{47} differences and/or means accounting for analytical covariance between samples (Appendices B.2–B.3); and assessing whether the Δ_{47}

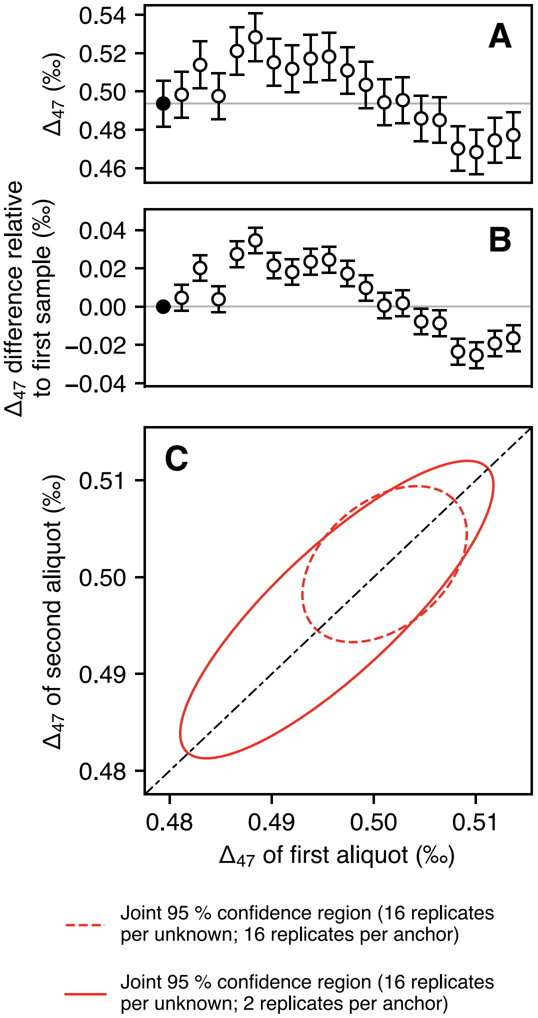

Figure 6. (a–b): Simulated series of 20 samples with similar bulk compositions, analyzed within a single session (four replicates for each anchor and unknown sample, $\sigma_{47} = 0.01‰$). Because of shared standardization errors, the uncertainties on the absolute Δ_{47} values of each sample (panel a, 95% confidence limits) are much larger than the uncertainties on the Δ_{47} differences (panel b, 95% confidence limits) between each sample and the first one (black marker). Note that panels a and b have identical vertical scales. (c): Simulated comparison of Δ_{47} values measured for two samples with identical compositions in (δ_{47}, Δ_{47}) space. Precisely comparing two unknowns samples with similar compositions only requires a few anchor analyses: increasing the number of replicate analyses per anchor from 2 to 16 reduces the uncertainties on the absolute Δ_{47} values of each unknown sample but does not improve constraints on the Δ_{47} difference between them. Both simulations were produced using the D47crunch library (cf. Section 4).

repeatabilities of two samples differ significantly. Both D47crunch and ClumpyCrunch also output robust 95% confidence limits for final Δ_{47} values based on the number of degrees of freedom in the standardization models. Links to the source code and documentation for D47crunch and ClumpyCrunch are provided below (see "Data and Code" section).

5. Recommendations

Based on the findings above we may offer the following recommendations, several of which are reiterations or reformulations of oft-repeated best practices.

Allocate anchors wisely. As illustrated by Figure 1, the standardization error field in (δ_{47}, Δ_{47}) space is primarily controlled by the Δ_{47} repeatability (σ_{47}), by the compositional distribution of anchor samples, and by the number of analyses performed for each anchor. The predicted properties of this error field are entirely consistent with the Monte Carlo simulations of Kocken et al. (2019), who called attention to the importance of optimizing the distribution of anchor replicates. When unknown samples of interest are close, in (δ_{47}, Δ_{47}) space, to one of the anchors, we again recommend analyzing many replicates of that anchor and just enough replicates of other anchors to constrain the standardization parameters. "Just enough replicates" is not entirely subjective, because we are now able to model quantitatively, as in Figure 1, how the standardization error field responds to different allocations of replicates among the anchor samples. In other cases, where unknown samples plot outside of the anchor polygon in (δ_{47}, Δ_{47}) space, the optimal choice of anchor analyses is less obvious, making this simulation approach even more useful (see below for a practical method to perform such simulations). Finally, analyses of specific types of natural samples with exotic isotopic compositions (e.g., methane seep carbonates) should greatly benefit from defining new, bespoke carbonate standards expressly chosen for this purpose.

*When in doubt, simulate standardization uncertainties***.** As mentioned above, it may be useful to predict the error fields resulting from arbitrary combinations of anchor/unknown analyses. D47crunch implements such simulations using the D47data.simulate() function, for any combination of user-defined samples, number of replicate analyses, and Δ_{47} repeatability (σ_{47}).

Analyze related samples together. As discussed above, Δ_{47} measurements of samples analyzed in one or more common sessions are not independent measurements (Section 3.3). As a result, Δ_{47} differences between unknown samples which were analyzed together are often more precisely constrained than their absolute Δ_{47} values. Figures 6a and 6b provide such an example, in which a simulated series of samples with identical bulk compositions but different Δ_{47} values are analyzed together. Similarly, when testing whether two samples with similar compositions in (δ_{47}, Δ_{47}) space have different Δ_{47} values (e.g., when testing different carbonate aliquots for homogeneity), standardization errors largely cancel out and autogenic errors dominate. In such cases, we recommend the unorthodox approach of short sessions with many unknown analyses and few anchor analyses (Figure 6c).

Report full uncertainties. Accurate comparisons of clumped-isotope data produced by different laboratories have long remained a challenge (Petersen et al., 2019, and references therein). A striking result of the InterCarb comparison exercise (Bernasconi, Daëron, et al., 2021) is that despite data sets from different labs

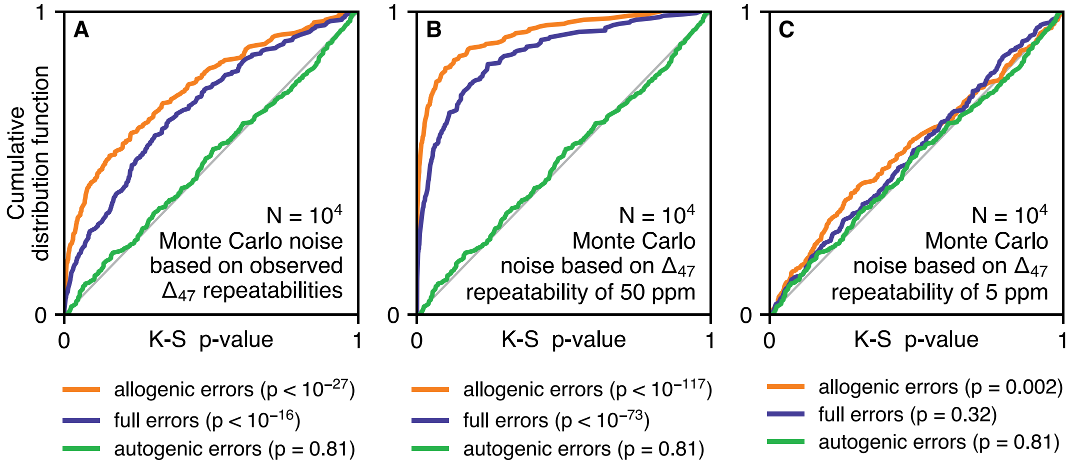

Figure 7. Monte Carlo simulation results All three panels display the cumulative distribution function of the *p*-values obtained from 10^4 Monte Carlo simulations of the full InterCarb data set (see Appendix C for computational details). When the random offsets used by the simulation are scaled according to the original data (a), autogenic errors behave in a Gaussian manner, but the standardization ("allogenic") errors do not, due to the limits of the first-order Taylor approximations used here for error propagation. As expected, greatly increasing (b) or decreasing (c) the random offsets used by the simulation results modulates the non-Gaussianity of standardization errors, while autogenic errors, despite being respectively increased or decreased (not shown here), remain Gaussian.

having extremely diverse analytical errors, the overall scatter between all laboratories is accurately predicted (i.e., neither too large nor too small) by the error propagation models described here, implying that carbonate-standardized Δ_{47} measurements are free of unrecognized systematic inter-laboratory discrepancies. It is thus reasonable to expect that we are now capable of quantitative comparisons between results from different laboratories, but this requires that future studies report full analytical uncertainties. At present, two options for estimating these uncertainties are available. One is to use the software described in Section 4 (ClumpyCrunch or D47crunch); the other is to implement Monte Carlo simulations similar to those described by Kocken et al. (2019). We recommend that existing, widely used software such as Easotope (John & Bowen, 2016) should eventually report full analytical error estimates by default.

Experiment with session length. There has been little discussion so far in the literature regarding the choice of analytical session length. Shorter sessions may obviously suffer from less robust statistics due to fewer observations. Conversely, longer sessions risk overestimating Δ_{47} repeatabilities in case of slow, non-motononic instrumental drifts on the same order as σ_{47}. Although this increase in apparent σ_{47} is counter-acted by a larger number of observations (N_a in Equation 5), apparently keeping modeled standardization errors small, the overall accuracy of the error model may suffer because slow drifts are by definition not random and do not necessarily cancel out over time. We recommend checking for such slow drifts by testing whether σ_{47} at short time scales (e.g., a few tens of analyses) is substantially smaller than at longer time scales (a few hundred analyses). This can easily be performed in post-processing by redefining session bounds (or session names in ClumpyCrunch).

Use pooled regression by default. Although the pooled approach described in Section 3.6 is not without limitations, it has been tested on over a year's worth of real-world data from several laboratories, and so far appears to offer greater statistical robustness at very little cost. Beyond rare pathological cases where some unknown samples are believed to have changed in composition over time, we recommend using this approach by default, or at least testing whether its output differs significantly from that of other methods.

6. Conclusion

The framework presented here provides a quantitative/predictive description of Δ_{47} error propagation, fully taking into account standardization errors and their properties. It corroborates and extends earlier investigations based on Monte Carlo simulations (Kocken et al., 2019). This mathematical formulation is found not to introduce large deviations from normality: in other words, if $\Delta_{47}^{\mathrm{raw}}$ errors are Gaussian, the fully propagated Δ_{47} errors may also be treated as Gaussian for all practical purposes. What is more, as reported by Bernasconi, Daëron et al. (2021), using this framework yields a very reasonable ($p = 0.19$) prediction for the distribution of inter-laboratory scatter in Δ_{47} values within the InterCarb data set.

Based on this framework, we describe a new, "pooled" standardization method designed to make full use of the constraints available from both anchor and unknown analyses. This approach is expected to yield substantially improved standardization models, in terms of both robustness and accuracy. We also provide new online resources and a Python library aiming to make the use of such error models as simple as possible. This library being open-source and fully documented, implementing the methods described here in existing software such as Easotope (John & Bowen, 2016) should be straightforward.

Most published clumped-isotope studies so far have lacked a rigorous propagation of standardization errors. This, of course, is not a problem in itself, but the InterCarb results unambiguously demonstrate that these standardization uncertainties are both necessary and sufficient to explain the inter-laboratory scatter observed in this large data set ($N = 5{,}329$). Going one step further, it could be argued that the ongoing persistence of inter-laboratory discrepancies in Δ_{47} calibrations (Petersen et al., 2019) is due, at least in part, to largely ignored standardization errors (Anderson et al., 2021). Whatever the case, it seems likely that future comparisons between results obtained in different laboratories would greatly benefit from more accurate error estimates.

Finally, although all statistical models are interpretative approximations, their ultimate value depends less on their exactness than on their practical usefulness. At the very least, the framework described here should help improve the manner in which we report analytical data and/or compare them across laboratories, and may inform our choice of standardization protocols (e.g., anchor/unknown ratios, compositional distribution of anchors, new reference materials).

Appendix A: Least Squares Regression

A.1. General Linear Case

Consider a linear model f defined as:

$$y = f(x, a_1, a_2 \ldots a_p) = \sum_{i=1}^{p} a_i f_i(x) \tag{19}$$

where x is a scalar or vectorial explanatory variable; y the response variable; ($f_1 \ldots f_p$) a series of functions of x; and ($a_1 \ldots a_p$) a series of scalar factors which are the model parameters to be estimated.

Given n observations $\left((x_1, y_1) \ldots (x_n, y_n)\right)$ to fit, we construct the following matrices:

$$A = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_p(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_p(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_n) & f_2(x_n) & \cdots & f_p(x_n) \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{20}$$

The best fit parameters ($a_1 \ldots a_n$) and their unscaled variance-covariance matrix V_0 are then:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \mathrm{V}_0 \times \mathrm{A}^{\mathrm{T}} \times \mathrm{Y} \qquad \text{with} \qquad \mathrm{V}_0 = \left(\mathrm{A}^{\mathrm{T}} \times \mathrm{A}\right)^{-1} \tag{21}$$

A.2. Application to the Standardization Model

The standardization model of Equation 3 is equivalent to the above formulation if:

$$
\begin{aligned}
y &= \Delta_{47}^{\mathrm{raw}} & f_1(x) &= \Delta_{47} \\
x &= (\delta_{47}, \Delta_{47}) & f_2(x) &= \delta_{47} \\
(a_1, a_2, a_3) &= (a, b, c) & f_3(x) &= 1
\end{aligned}
\tag{22}
$$

In this case:

$$
\mathrm{A} = \begin{bmatrix} \Delta_{47} & \delta_{47} & 1 \\ \Delta_{47} & \delta_{47} & 1 \\ \vdots & \vdots & \vdots \\ \Delta_{47} & \delta_{47} & 1 \end{bmatrix}
\begin{array}{c} \leftarrow \text{analysis \#1} \rightarrow \\ \leftarrow \text{analysis \#2} \rightarrow \\ \vdots \\ \leftarrow \text{analysis \#}n \rightarrow \end{array}
\begin{bmatrix} \Delta_{47}^{\mathrm{raw}} \\ \Delta_{47}^{\mathrm{raw}} \\ \vdots \\ \Delta_{47}^{\mathrm{raw}} \end{bmatrix} = \mathrm{Y}
\tag{23}
$$

To take into account an uncertainty, noted σ, assigned to the observations, A and Y should both be divided by σ, which will leave the best fit parameters unchanged and scale the variance-covariance matrix V_0 by a factor of σ^2 (as in Equation 7).

Alternatively, to assign individual uncertainties, noted $(\sigma_1 ... \sigma_n)$ to the n analyses, each line of A and each element of Y should be divided by the corresponding σ value:

$$
\mathrm{A} = \begin{bmatrix} \Delta_{47} / \sigma_1 & \delta_{47} / \sigma_1 & 1 / \sigma_1 \\ \Delta_{47} / \sigma_2 & \delta_{47} / \sigma_2 & 1 / \sigma_2 \\ \vdots & \vdots & \vdots \\ \Delta_{47} / \sigma_n & \delta_{47} / \sigma_n & 1 / \sigma_n \end{bmatrix}
\begin{array}{c} \leftarrow \text{analysis \#1} \rightarrow \\ \leftarrow \text{analysis \#2} \rightarrow \\ \vdots \\ \leftarrow \text{analysis \#}n \rightarrow \end{array}
\begin{bmatrix} \Delta_{47}^{\mathrm{raw}} / \sigma_1 \\ \Delta_{47}^{\mathrm{raw}} / \sigma_2 \\ \vdots \\ \Delta_{47}^{\mathrm{raw}} / \sigma_n \end{bmatrix} = \mathrm{Y}
\tag{24}
$$

Extending this model with additional parameters should be rather straightforward. For instance, in order to account for a temporal drift in the compositional nonlinearity, one could reformulate the model, with t denoting time and an additional standardization parameter d, as:

$$
\Delta_{47}^{\mathrm{raw}} = a\, \Delta_{47} + (b + td)\, \delta_{47} + c
\tag{25}
$$

Which would correspond to:

$$
\mathrm{A} = \begin{bmatrix} \Delta_{47} & \delta_{47} & 1 & t\delta_{47} \\ \Delta_{47} & \delta_{47} & 1 & t\delta_{47} \\ \vdots & \vdots & \vdots & \vdots \\ \Delta_{47} & \delta_{47} & 1 & t\delta_{47} \end{bmatrix}
\begin{array}{c} \leftarrow \text{analysis \#1} \rightarrow \\ \leftarrow \text{analysis \#2} \rightarrow \\ \vdots \\ \leftarrow \text{analysis \#}n \rightarrow \end{array}
\begin{bmatrix} \Delta_{47}^{\mathrm{raw}} \\ \Delta_{47}^{\mathrm{raw}} \\ \vdots \\ \Delta_{47}^{\mathrm{raw}} \end{bmatrix} = \mathrm{Y}
\tag{26}
$$

Appendix B: Δ_{47} Covariance

B.1. Covariance Between Unknown Samples

Consider two unknown samples A and B, whose session-averaged compositions ($\overline{\delta_{47}}$, $\overline{\Delta_{47}^{\mathrm{raw}}}$, and $\overline{\Delta_{47}}$) are respectively noted δ_{A}, δ_{B}, $\Delta_{\mathrm{A}}^{\mathrm{raw}}$, $\Delta_{\mathrm{B}}^{\mathrm{raw}}$, Δ_{A}, and Δ_{B}. Defining X as the column vector $[\Delta_{\mathrm{A}}, \Delta_{\mathrm{B}}]$, we can express its Jacobian J_{X} relative to the system of variables ($\Delta_{\mathrm{A}}^{\mathrm{raw}}$, $\Delta_{\mathrm{B}}^{\mathrm{raw}}$, a, b, c) and the covariance C of this quintuplet as:

$$
\mathrm{J}_{\mathrm{X}} = \frac{1}{a} \begin{bmatrix} 1 & 0 & -\Delta_{\mathrm{A}} & -\delta_{\mathrm{A}} & -1 \\ 0 & 1 & -\Delta_{\mathrm{B}} & -\delta_{\mathrm{B}} & -1 \end{bmatrix}
\qquad
\mathrm{C} = \begin{bmatrix} \left(\sigma_{47}^{\mathrm{raw}}\right)^2 / \mathrm{N}_{\mathrm{A}} & 0 & 0 & 0 & 0 \\ 0 & \left(\sigma_{47}^{\mathrm{raw}}\right)^2 / \mathrm{N}_{\mathrm{B}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_a^2 & \mathrm{c}_{ab} & \mathrm{c}_{ac} \\ 0 & 0 & \mathrm{c}_{ab} & \sigma_b^2 & \mathrm{c}_{bc} \\ 0 & 0 & \mathrm{c}_{ac} & \mathrm{c}_{bc} & \sigma_c^2 \end{bmatrix}
\tag{27}
$$

The covariance matrix of X is then:

$$\mathrm{C_X} = \mathrm{J_X} \times \mathrm{C} \times \mathrm{J_X^T} \tag{28}$$

Because of the structure of J and C, the non-zero terms of $\mathrm{C_X}$ are equal to:

$$\mathrm{cov}(\Delta_\mathrm{A},\Delta_\mathrm{B}) = \frac{1}{a^2}\begin{bmatrix}\Delta_\mathrm{A} & \delta_\mathrm{A} & 1\end{bmatrix} \times \begin{bmatrix}\sigma_a^2 & \mathrm{c}_{ab} & \mathrm{c}_{ac} \\ \mathrm{c}_{ab} & \sigma_b^2 & \mathrm{c}_{bc} \\ \mathrm{c}_{ac} & \mathrm{c}_{bc} & \sigma_c^2\end{bmatrix} \times \begin{bmatrix}\Delta_\mathrm{B} \\ \delta_\mathrm{B} \\ 1\end{bmatrix} \tag{29}$$

$$\begin{aligned}\mathrm{cov}(\Delta_\mathrm{A},\Delta_\mathrm{B}) = \; & \frac{1}{a^2}\Delta_\mathrm{A}\Delta_\mathrm{B}\sigma_a^2 + \delta_\mathrm{A}\delta_\mathrm{B}\sigma_b^2 + \sigma_c^2 + (\Delta_\mathrm{A}\delta_\mathrm{B} + \delta_\mathrm{A}\Delta_\mathrm{B})\mathrm{c}_{ab} \\ & + (\Delta_\mathrm{A} + \Delta_\mathrm{B})\mathrm{c}_{ac} + (\delta_\mathrm{A} + \delta_\mathrm{B})\mathrm{c}_{bc}\Big)\end{aligned} \tag{30}$$

The covariance between mean Δ_{47} values of two samples averaged over several sessions is zero if the samples were never analyzed in the same session. Otherwise, with $\omega_{\mathrm{A}i}$ and $\omega_{\mathrm{A}i}$ weights defined as in 15:

$$\Delta_\mathrm{A}^\mathrm{final} = \sum_i \omega_{\mathrm{A}i}\,\Delta_{\mathrm{A}i} \qquad \text{with } i \text{ denoting all sessions including A} \tag{31}$$

$$\Delta_\mathrm{B}^\mathrm{final} = \sum_j \omega_{\mathrm{B}j}\,\Delta_{\mathrm{B}j} \qquad \text{with } j \text{ denoting all sessions including B} \tag{32}$$

$$\mathrm{cov}\left(\Delta_\mathrm{A}^\mathrm{final},\Delta_\mathrm{B}^\mathrm{final}\right) = \sum_k \omega_{\mathrm{A}k}\,\omega_{\mathrm{B}k}\,\mathrm{cov}\left(\Delta_{\mathrm{A}k},\Delta_{\mathrm{B}k}\right) \qquad \text{with } k \text{ denoting all sessions including both A and B} \tag{33}$$

B.2. Standard Errors on Δ_{47} Differences Between Samples

Consider two unknown samples A and B, whose session-averages compositions ($\overline{\delta_{47}}$, $\overline{\Delta_{47}^\mathrm{raw}}$, and $\overline{\Delta_{47}}$) are respectively noted δ_A, δ_B, $\Delta_\mathrm{A}^\mathrm{raw}$, $\Delta_\mathrm{B}^\mathrm{raw}$, Δ_A, and Δ_B. Defining x as the difference ($\Delta_\mathrm{A} - \Delta_\mathrm{B}$), we can express its Jacobian J_x relative to the system of variables ($\Delta_\mathrm{A}^\mathrm{raw}$, $\Delta_\mathrm{B}^\mathrm{raw}$, a, b, c) as:

$$\mathrm{J}_x = \frac{1}{a}\left[1, -1, \Delta_\mathrm{B} - \Delta_\mathrm{A}, \delta_\mathrm{B} - \delta_\mathrm{A}, 0\right] \tag{34}$$

and compute the variance of x using the same covariance matrix C as above:

$$\sigma_x^2 = \mathrm{J}_x \times \mathrm{C} \times \mathrm{J}_x^\mathrm{T} \tag{35}$$

$$\sigma_x^2 = \sigma_{47}^2\left(\frac{1}{\mathrm{N_A}} + \frac{1}{\mathrm{N_B}}\right) + \frac{(\Delta_\mathrm{B} - \Delta_\mathrm{A})^2\sigma_a^2 + (\delta_\mathrm{B} - \delta_\mathrm{A})^2\sigma_b^2 + 2(\Delta_\mathrm{B} - \Delta_\mathrm{A})(\delta_\mathrm{B} - \delta_\mathrm{A})\mathrm{c}_{ab}}{a^2} \tag{36}$$

B.3. Standard Errors on Mean Δ_{47} Values Averaged Over Several Samples

As an example, we treat here the problem of a weighted average of three samples. Consider three unknown samples A, B, and C, whose session-averages compositions ($\overline{\delta_{47}}$, $\overline{\Delta_{47}^\mathrm{raw}}$, and $\overline{\Delta_{47}}$) are respectively noted δ_A, δ_B, δ_C, $\Delta_\mathrm{A}^\mathrm{raw}$, $\Delta_\mathrm{B}^\mathrm{raw}$, $\Delta_\mathrm{C}^\mathrm{raw}$, Δ_A, Δ_B, and Δ_C. Defining W as the weighted average ($x_\mathrm{A}\Delta_\mathrm{A} + x_\mathrm{B}\Delta_\mathrm{B} + x_\mathrm{C}\Delta_\mathrm{C}$) and w as the weighted average ($x_\mathrm{A}\delta_\mathrm{A} + x_\mathrm{B}\delta_\mathrm{B} + x_\mathrm{C}\delta_\mathrm{C}$), we can express the Jacobian of W relative to the system of variables ($\Delta_\mathrm{A}^\mathrm{raw}$, $\Delta_\mathrm{B}^\mathrm{raw}$, $\Delta_\mathrm{C}^\mathrm{raw}$, a, b, c) as:

$$J_W = \frac{1}{a}\begin{bmatrix}x_A & x_B & x_C & -W & -w & -1\end{bmatrix} \tag{37}$$

and compute the variance of W using the same method as above:

$$\sigma_W^2 = \mathrm{J}_W \times \begin{bmatrix} \left(\sigma_{47}^{\mathrm{raw}}\right)^2 / \mathrm{N_A} & 0 & 0 & 0 & 0 & 0 \\ 0 & \left(\sigma_{47}^{\mathrm{raw}}\right)^2 / \mathrm{N_B} & 0 & 0 & 0 & 0 \\ 0 & 0 & \left(\sigma_{47}^{\mathrm{raw}}\right)^2 / \mathrm{N_C} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_a^2 & \mathrm{c}_{ab} & \mathrm{c}_{ac} \\ 0 & 0 & 0 & \mathrm{c}_{ab} & \sigma_b^2 & \mathrm{c}_{bc} \\ 0 & 0 & 0 & \mathrm{c}_{ac} & \mathrm{c}_{bc} & \sigma_c^2 \end{bmatrix} \times \mathrm{J}_W^{\mathrm{T}} \tag{38}$$

$$\sigma_W^2 = \sigma_{47}^2 \left(\frac{x_{\mathrm{A}}^2}{\mathrm{N_A}} + \frac{x_{\mathrm{B}}^2}{\mathrm{N_B}} + \frac{x_{\mathrm{C}}^2}{\mathrm{N_C}} \right) + \frac{W^2\sigma_a^2 + w^2\sigma_b^2 + \sigma_c^2 + 2\left(wW\mathrm{c}_{ab} + W\mathrm{c}_{ac} + w\mathrm{c}_{bc}\right)}{a^2} \tag{39}$$

Note that the second term above is equal to the value of the standardization error field at the weighted barycenter of the samples in (δ_{47}, Δ_{47}) space.

Appendix C: Monte Carlo Assessment of the Normality of Δ_{47} Errors

Because Δ_{47} is not a linear function of ($\Delta_{47}^{\mathrm{raw}}$, δ_{47}, a, b, c), the propagation of standardization errors described in Section 2.6 is an approximation. Here we used a Monte Carlo simulation based on the full InterCarb data set to investigate how much autogenic and allogenic errors deviate from a Gaussian approximation. In each step of the simulation, we offset the original $\Delta_{47}^{\mathrm{raw}}$ values observed in each of the 5,329 analyses by random, independent, zero-centered Gaussian errors with a standard deviation equal to the session's $\sigma_{47}^{\mathrm{raw}}$ value. We then standardize all sessions of the modified data set and record the final, session-averaged Δ_{47} values of each unknown sample ($N = 226$) for a total of 10^4 iterations. Each of these session averages is submitted to a Kolmogorov-Smirnov (KS) test of normality (Massey, 1951), comparing the distribution of these 10^4 values to a normal distribution centered on the original session-averaged value and whose width depends of the original propagated errors. Each of the 226 KS tests yields a p-value corresponding to the null hypothesis that the two distributions are identical. By design, if the Gaussian approximation of the propagated errors holds true, these p-values should be evenly distributed in the [0–1] interval. We may quantify how well they do so by performing a final KS test comparing the distribution of p-values to the uniform distribution, yielding a new, final p-value for the hypothesis that the errors in the InterCarb data set follow Gaussian distributions.

We run this simulation in three different configurations, considering only autogenic errors, only standardization errors, or both. Initially, the random errors introduced in each iteration are scaled according to the Δ_{47} repeatability of each session (Figure 7a). We then repeat the simulations twice, by scaling the random errors according to a constant Δ_{47} repeatability of 50 and 5 ppm, respectively (Figures 7b and 7c).

Predictably, based on Equations 12 and 13, we find that autogenic errors behave in a Gaussian manner ($p = 0.81$), but this is clearly not the case for standardization errors ($p < 10^{-27}$). Because the error propagation formula of Equation 8 is equivalent to a first-order Taylor expansion, the non-normality of standardization errors is expected to worsen as Δ_{47} repeatability increases, as is the case in Figure 7b, and to become negligible when Δ_{47} repeatability is small enough (Figure 7c).

Notations

a	scrambling factor, one of the standardization parameters, quantifying the amount of molecular recombination during the analytical procedure; its value should lie between 0 and 1.
b	compositional slope, one of the standardization parameters, quantifying small systematic errors in the electrical background of the ion collectors; it may be positive or negative and its absolute value should ideally remain small (10^{-2} or less).

c	working gas offset, one of the standardization parameters, accounting for the fact that the working gas is not necessarily stochastic; in settings where the working gas is equilibrated at room temperature, $c \approx -a$
N_a	Number of analyses.
N_f	Degrees of freedom in a regression model.
N_S	Number of samples.
N_A	Number of anchor samples.
N_U	Number of unknown samples.
Δ_{47}	delta notation (in ‰) for the clumped-isotope anomaly associated with mass-47 CO_2; either denotes the "true" value for a given sample, or the "absolute" value computed from one or more IRMS measurements after standardization.
Δ_{47}^raw	"raw" Δ_{47} value from an IRMS measurement, before standardization.
δ_{47}	delta notation (in ‰) for the mass-47 to mass-44 abundance ratio of an analyte CO_2, generally defined relative to a working reference gas.
σ_{47}	analytical error/uncertainty assigned to individual measurements of Δ_{47} (Equation 5).
σ_{47}^raw	analytical error/uncertainty assigned to individual measurements of Δ_{47}^raw (Equation 6).
σ_s	allogenic error, i.e., the analytical error/uncertainty on a Δ_{47} measurement arising from the standardization function (Equation 13).
σ_u	autogenic error, i.e., the analytical error/uncertainty on a Δ_{47} measurement arising from the analyses of the unknown sample itself (Equation 12).

Data Availability Statement

The complete raw data and all associated code used in this work are available under a Modified BSD License at https://doi.org/10.5281/zenodo.4314593. The preferred way to comment on the code or to suggest improvements is to raise an issue at https://github.com/mdaeron/D47_error_propagation. D47crunch is easily installed through the Python Package Index ("pip install D47crunch"). To download the latest versions of the code source, contribute improvements, report bugs, or suggest new features, see https://github.com/mdaeron/D47crunch. Full documentation is available at https://mdaeron.github.io/D47crunch. The ClumpyCrunch source code is also available at https://github.com/mdaeron/clumpycrunch.

Acknowledgments
I wish to thank all InterCarb participants, whose involvement in generating a unique clumped-isotope data set provided a much-needed "ground truth" of what real Δ_{47} data looks like in the wild. Thanks to D. Blamart, M. Peral, and W. Defliese, who took the time to comment on the manuscript and the ClumpyCrunch website. I am also grateful to G. Olack, I. Kocken, and a third, anonymous reviewer, whose detailed comments and suggestions helped substantially improve this work. Finally, many thanks to D. Blamart for his thoughtful advice and guidance over the years. This is LSCE contribution 7537.

References

Anderson, N. T., Kelson, J. R., Kele, S., Daëron, M., Bonifacie, M., Horita, J., et al. (2021). A unified clumped isotope thermometer calibration (0.5°C –1,100°C) using carbonate-based standardization. *Geophysical Research Letters*, *48*. https://doi.org/10.1029/2020GL092069

Bernasconi, S. M., Daëron, M., Bergmann, K. D., Bonifacie, M., Meckler, A. N., et al. (2021). InterCarb: A community effort to improve inter-laboratory standardization of the carbonate clumped isotope thermometer using carbonate standards. *Geochemistry, Geophysics, Geosystems*. https://doi.org/10.1029/2020GC009588

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*(5), 603–612. https://doi.org/10.1002/rcm.6490

Bernasconi, S. M., Müller, I. A., Bergmann, K. D., Breitenbach, S. F. M., Fernandez, A., Hodell, D. A., et al. (2018). Reducing uncertainties in carbonate clumped isotope analysis through consistent carbonate-based standardization. *Geochemistry, Geophysics, Geosystems*, *19*, 2895–2914. https://doi.org/10.1029/2017GC007385

Bonifacie, M., Calmels, D., Eiler, J. M., Horita, J., Chaduteau, C., Vasconcelos, C., et al. (2017). Calibration of the dolomite clumped isotope thermometer from 25°C 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

Daëron, M. & Blamart, D. (2016). *Uncertainties and standardization in the absolute reference frame. Abstract, Fifth International Clumped Isotope Workshop*. url. 02492075 https://hal.archives-ouvertes.fr/hal-

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

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

Eiler, J. M. (2013). The isotopic anatomies of molecules and minerals. *Annual Review of Earth and Planetary Sciences*, *41*(1), 411–441. https://doi.org/10.1146/annurev-earth-042711-105348

Eiler, J. M., & Schauble, E. (2004). $^{18}O^{13}C^{16}O$ in Earth's atmosphere. *Geochimica et Cosmochimica Acta*, *68*(23), 4767–4777. https://doi.org/10.1016/j.gca.2004.05.035

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*(12), 4375–4386. https://doi.org/10.1002/2017gc007106

Ghosh, P., Adkins, J., Affek, H., Balta, B., Guo, W., Schauble, E. A., et al. (2006). ^{13}C–^{18}O bonds in carbonate minerals: A new kind of paleothermometer. *Geochimica et Cosmochimica Acta*, *70*, 1439–1456. https://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

Huntington, K. W., Eiler, J. M., Affek, H. P., Guo, W., Bonifacie, M., Yeung, L. Y., et al. (2009). Methods and limitations of “clumped” CO_2 isotope (Δ_{47}) analysis by gas-source isotope ratio mass spectrometry. *Journal of Mass Spectrometry*, *44*(9), 1318–1329. https://doi.org/10.1002/jms.1614

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

Kocken, I. J., Müller, I. A., & Ziegler, M. (2019). Optimizing the use of carbonate standards to minimize uncertainties in clumped isotope data. *Geochemistry, Geophysics, Geosystems*, *20*(11), 5565–5577. https://doi.org/10.1029/2019gc00854510.1029/2019gc008545

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. *Quarterly of Applied Mathematics*, *2*, 164–168. https://doi.org/10.1090/qam/10666

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*, 431–441. https://doi.org/10.1137/0111030, http://www.jstor.org/stable/2098941

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.1080/01621459.1951.10500769

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

Newville, M., Stensitzki, T., Allen, D. B., & Ingargiola, A. (2014) LMFIT: Nonlinear least-square minimization and curve- fitting for python. https://doi.org/10.5281/zenodo.11813

Olack, G., & Colman, A. S. (2019). Modeling the measurement: Δ_{47}, corrections, and absolute ratios for reference materials. *Geochemistry, Geophysics, Geosystems*, *20*, 3569–3587. https://doi.org/10.1029/2018gc008166

Petersen, S. V., Defliese, W. F., Saenger, C., Daëron, M., Huntington, K. W., John, C. M., et al. (2019). Effects of improved ^{17}O correction on interlaboratory agreement in clumped isotope calibrations, estimates of mineral-specific offsets, and temperature dependence of acid digestion fractionation. *Geochemistry, Geophysics, Geosystems*., *20*, 3495–3519. https://doi.org/10.1029/2018gc008127

Schauble, E. A., Ghosh, P., & Eiler, J. M. (2006). Preferential formation of ^{13}C–^{18}O bonds in carbonate minerals, estimated using first-principles lattice dynamics. *Geochimica et Cosmochimica Acta*, *70*, 2510–2529. https://doi.org/10.1016/j.gca.2006.02.011

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*(24), 2607–2616. https://doi.org/10.1002/rcm.7743

Schmid, T. W., & Bernasconi, S. M. (2010). An automated method for 'clumped-isotope' measurements on small carbonate samples. *Rapid Communications in Mass Spectrometry*, *24*(14), 1955–1963. https://doi.org/10.1002/rcm.4598

Tellinghuisen, J. (2001). Statistical error propagation. *Journal of Physical Chemistry A*, *105*, 3917–3921. https://doi.org/10.1021/jp003484u