Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-269-AC1, 2020

Biogeosciences
Discussions

Interactive comment on "Technical note: Single-shell δ^{11}B analysis of *Cibicidoides wuellerstorfi* using femtosecond laser ablation MC-ICPMS and secondary ion mass spectrometry" *by* Markus Raitzsch et al.

Markus Raitzsch et al.

mraitzsch@marum.de

Received and published: 9 September 2020

AC: We thank Dennis Mayk for his constructive and thorough review of our manuscript and will address all his remarks and suggestions, which are listed below:

Raitzsch et al. present an interesting and timely manuscript about a comparative study of B isotopes in the benthic foraminifera Cibicidoides wuellerstorfi analysed using LA-MC-ICP-MS and SIMS. Despite the relevance of δ11B as a paleo-pH-proxy, very few studies have been published showing intra and inter foraminfera test (shell) δ11B vari-

ability as these analysis have proven to be challenging due to low B concentration and fragility of foraminifera tests. This study provides an interesting comparison between different heterogeneity levels within and between individual foraminfera which will be of widespread interest and should be published after revision of the issues listed below:

Main comments: 1. Data processing: The manuscript lacks a general explanation of how the data were treated after collection. Fig. 2 shows a typical time-resolved laser ablation profile for a clean and a contaminated (clay filled) foraminifer. In the caption, it is mentioned that some points have been removed from the ablation trend by a 2-sigma outlier test, however in the methods there is no explanation of the data processing involved. It would be important to mention the general data reduction routine that was employed.

AC: Right, the explanation of the data reduction routine is obviously too sparse (l. 119-120), and will be replaced by a more detailed version.

Furthermore, the ablation intensity profiles appear very bulgy and do not present apparent plateaus. Please report how the shell signal was extracted from the rest?

AC: The reason for the bulgy shape of the signal intensities in Fig. 2 is probably related to changing ablation efficiency for some samples, i.e. a more efficient ablation of material from a surface progressively getting rougher after a couple of "helix turns". As we always attempted to match the signal, i.e. to gain the same signal intensities between sample and standard, we often had to increase the ablation frequency to enhance ablation at the beginning of a measurement. Conversely, after some time, we often had to decrease the frequency, when ablation was liable to become too strong, resulting in a bulgy profile as shown in Fig. 2. For integration of the shell signal, we chose an interval, where the signal ratio clearly showed a smooth plateau (see Fig. 2), which will be better explained in the revised manuscript.

2. Sample size estimation: The estimation of the required sample size to resolve 0.1 pH unit is a very important part of the manuscript but the R function "combn()" used for

that purpose lacks a detailed explanation in the manuscript – in addition it is unclear if the presumptions made in the manuscript are correct or lead to an underestimation of the required sample size. In detail: On line 237 it is reported that the sample size simulation is based on the assumption that the entire population (P) consists of the 18 shells analysed. Although this holds true for this particular study it is not a representation of the actual (true) population size which is what future studies would be interested in to estimate required sample sizes. In other words, the presumption of P =18 holds only true within this study but has no real world application. Instead it should be discussed what population sizes are realistic within similar pH-environments and simulations should be based on these. Furthermore, it is not clear how the simulation are carried out using the 18 shells as they have not been measured in the same way according to the Supplemental Material. The “large crater” was analysed on 16 shells and the “umbilical knob” was also analysed on 16 shells suggesting that for the simulation using 18 shells two different measurement “types” were merged which further complicates its validity. It would be more informative to separate the two and report required sample sizes based on measurement type i.e., for measurements on the “umbilical knob” and for “large crater” measurements.

AC: This is a good point raised by Dennis, as only the 14 and 16 individuals were analyzed using the "umbilical knob" and "large crater", respectively. Hence, for the analysis of "sample size requirement" we will exclude the 2 individuals analyzed for inter-chamber variability, and only examine the separate variabilities based on "knob" and "large crater" analyses, as well as on the variability where both analysis types are averaged. Accordingly, we will update Fig. 7 showing the results.

Considering the sample size of 16 or 18 it appears to be useful to consider the use of a conventional sample size estimation approach in comparison to a resampling approach as drawing from the same small population may result in errors. In the figure below, the estimated sample sizes required for e = 0.5 (2SD) and 1-α = 0.05 in relation to the population size is given as estimated by the R function “sample.size.mean()”

(https://CRAN.R-project.org/package=samplingbook) for both measurement “types”. Given an overall population size of e.g., 500 specimens in the same pH-environment, it would require n = 87 specimens if the “umbilical knob” was measured and n = 40 if the “large crater” was measured (based on the variability observed in this study) to achieve the desired significance level. Even if the population size consisted of only 16 individuals, the estimated sample size would be n = 16 and n = 14, respectively and thus twice as large as reported in the manuscript.

AC: Also this is a very important point, which made us to rethink about this issue. Firstly, it is right that our approach is not the most appropriate resampling method, as the entire data population only consists of the 18 measured individuals. The R function "combn()" searches for all possible combinations within this population and hence does not apply replacement of a sample, i.e. it does not resample one foram multiple times for generating one subsample. This ultimately results in an underestimation of the uncertainty as a function of measured individuals, which was also correctly pointed out by Dennis. However, we think that the resampling approach proposed by Dennis (R function “sample.size.mean()”) is not the most appropriate neither. That is because this function assumes that the value of a measurand from the entire population may be approached by measuring a random subsample, the size of which is dependent on the population size and the target uncertainty. In other words, this function allows for determining the required subsample size in order to gain the "true" average value of the entire population to within a quoted uncertainty, but it does not reflect whether the average value accurately records the influencing variable, in our case pH that influences δ11B. The output plot provided by Dennis implies that a few individuals are sufficient to gain an accurate value, if the population is small, but this is not true as each specimen has a large uncertainty in terms of the closeness of the agreement between the measured δ11B and the influencing pH. Consequently, the relationship between the "accuracy" and number of analyzed individuals must be independent of the population size.

Based on these thoughts, we will apply a different method, but which partly goes the same direction as the resampling approach suggested by Dennis. In the revised manuscript, we will use a Monte Carlo simulation, where a large artificial population (n=10,000) is created by randomly generating δ11B values around the "true" δ11B value within the determined individual uncertainty of $\pm$0.84 and $\pm$1.38 ‰ (SD) for "large craters" and "umbilical knobs", respectively. From this population, we will randomly resample and average N values to determine the 2SD uncertainty (= the potential error of δ11B) as a function of N analyzed individuals. We think that this approach is the most appropriate one, as it is independent of the population size and does not underestimate the uncertainty, as does our initial approach.

Minor comments: This is a non-comprehensive list of minor issues Line 30: Consider removing the last clause of the abstract. “Vital effect” is a loaded term and since it is not further discussed (Line 185) of little value for this manuscript.

AC: Correct, will be removed.

Line 35: Space missing between 27.2 and $\pm$0.6 ‰

AC: OK.

Line 57: Comma missing after “Also”

AC: OK.

Line 69: Considering that this study looked at a total of 23 specimens the term “tens of specimens” seems excessive, better report the actual number of individuals.

AC: Right, will be changed.

Line 179: Why was a non-parametric test used? Please specify what data the test was used on? Please report the Wilcoxon-Mann-Whitney test summary i.e., (W = XXX, p <0.001)

AC: We used a non-parametric test because it does not imply defined probability distri-

butions a priori, but is open to the model structure. The WMW uses randomly selected values X and Y from two populations, and tests the null hypothesis whether the probability that X>Y is equal to the probability that Y>X. However, I am very happy that Dennis made this a subject of discussion, since I walked right into a trap when testing the null hypotheses. Because of the few datapoints for each chamber, I applied the statistical test on Monte-Carlo simulated d11B values (n=10) around quoted uncertainties, yielding p values smaller than 0.05, meaning that the differences in d11B between chambers f-1 and f-5 are statistically significant at a 95 % SL. This artificially increased population size, however, led to a biased uncertainty estimation, which was also subject to papers in mathematical journals (e.g. Lin et al. (2013), Too Big to Fail: Large Samples and the p-Value Problem, http://dx.doi.org/10.1287/isre.2013.0480). If just the original data are taken into account, both the WMW and Welch t-test yield p values $\sim$0.07, and hence the d11B differences between the chambers are not statistically different at a 95 % SL, based on the small datasets of this study. I have to apologize for this incautious and naive application of statistical tests on our data. This will be clarified in the revised manuscript.

Line 184: Space missing between “large-scale” and “suggesting”

AC: OK

Line 186: “Somewhat” not useful, report how much δ11B was elevated in the umbilical knob

AC: OK, will be changed.

Line 197: a total of 18 shells “were” used

AC: OK, thanks.

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2020-269, 2020.