Review by Manuel F. G. Weinkauf

Manuel.Weinkauf@unige.ch

I have been reviewing the manuscript by Mallo et al. entitled 'Low planktic foraminiferal diversity and abundance observed in a spring 2013 West–East Mediterranean Sea plankton tow transect', and submitted to the journal Biogeosciences, in its second revised version.

This paper studies planktonic Foraminifera, sampled with plankton nets in 200 m water depth during spring/summer 2013 across the entire Mediterranean Sea. It reports abundance patterns of several species across a Mediterranean transect which is characterized by large differences in physical ocean properties (e.g. temperature, salinity). It further tries to infer the influence of those environmental parameters on the abundance and shell calcification intensity of selected (abundant) species. The study finds that the species composition changes across the Mediterranean, with *Globigerina bulloides* and *Trilobatus sacculifer* dominating in the western part, *Globorotalia inflata* in the central part, and *Globigerinoides ruber* (white)/*Globigerinoides elongatus* in the east. The species investigated for their abundance and calcification intensity show distribution and calcification patterns that differ between regions in the Mediterranean Sea, and can partly be correlated with environmental factors.

I appreciate this study for its large potential in filling in gaps in our current knowledge about species distribution in the Mediterranean and their changes both seasonally and across longer timespans by comparison of their results with earlier studies. It can also be a significant contribution to the still relatively scarce set of literature about shell calcification in planktonic Foraminifera. The sections are logically ordered, and the abstract gives a sufficient and well structured overview over the manuscript. Otherwise the manuscript has an appropriate length (although the discussion is rather long I do not think it is excessive). The figures and tables are suitable.

The manuscript has again been improved since the first revised version. Doing so, however, revealed a new problem of which I was not aware. The major problem is that the data are even less comparable to earlier studies than I believed so far, because they do not represent the depth range that has so far been indicated by the authors (compare General comments). As a result, part of the comparison with older works must be further toned down. Additionally, there are still some small issues with the analyses, but those can now be dealt with reasonably quickly.

I must admit that the analytical quality in the manuscript is not to the highest standards (I still believe that principal component analysis (PCA) is a subpar analysis for what the authors want to do). But I do not see how this can be helped, and although I would have wished for a more thorough and robust analysis, PCA as applied by the authors (i.e. for ordinating on the environmental parameters) is not wrong. I therefore believe that once

the authors dealt with the remaining problems, we finally reached the state where the manuscript can be published in Biogeosciences.

General comments

I have only one major problem left with the new iteration of the manuscript, that is a change in the description of the data. So far I, was led to believe, that the assemblages investigated by the authors have a bias towards coming from 200 m depth, but that they are representative of the entire water column from 0–200 m. This was indicated by the phrasing used by the authors in the first revised manuscript version, lines 165–167 'Twenty samples were collected with BONGO nets (mesh size 150 µm and 40 cm diameter, for further details see Posgay, 1980) primarily 200 m depth, but also including tow time integrating the upper water column from 200 m to the surface (Table 1).'. This statement has now been revised to 'Those nets sampled primarily 200 m depth, but also caught foraminifera during the net descent and ascent to the ocean surface, which both involve **negligible** towing and capturing time compared to the sampling at 200 m depth (Table 1).' (lines 167–169, emphasis by me).

This statement is in stark contrast to the statement in the first revised version of the manuscript. It now means that the data available to the authors are even less suitable for any comparison with earlier studies, which always used the data from the surface to a particular depth. The authors now not only have a snap-shot in time, but also in space, which is very critical given the distinct depth distribution of planktonic foraminiferal species (Rebotim et al., in press) when one wants to compare abundance data with other studies. It in fact means, that the authors should have missed representative populations of all species identified as shallow and most species from the intermediate depth range *sensu* Rebotim et al. (in press). As it stands, I do not see how this can be corrected in any way, and actually I doubt that the assemblages mainly represent the standing stock at 200 m depth, because of the large abundances of shallow species like *G. ruber* (white) and *T. sacculifer* (both occurring mainly above 70 m, compare Rebotim et al. (in press)) in some samples. Alas, the authors seem not to know what exactly they were catching themselves either, which is why I have to conclude that any attempt to compare the assemblages presented here with any of the older studies on a more than very basic quantitative level are futile and should be taken out of the manuscript. It also raises the question of how well and with what attention the experimental design was thought through, because if the goal was a comparison with earlier studies, it would have been wise to stick as close as possible to sample schemes of such earlier studies, instead of doing something basically incomparable.

My request is therefore to further tone down the comparison with earlier studies, specifically removing the text passages between lines 397 and 416.

Detailed comments

Line 40, 'predatory presence': I have never heard that. In fact, we do not have the faintest idea if there are selective predators targeting Foraminifera.

Lines 41f, '(i.e., Schiebel and Hemleben, 2005; Hemleben et al., 1989': Replace 'i.e.' with 'e.g.', consider citing Rebotim et al. (in press).

Lines 50f, 'Pujol and Vergnaud-Grazzini, 1995': I still believe it should be 'Pujol and Vergnaud Grazzini, 1995' (without hyphen) in the entire manuscript, since this is how the author wants to be referred to according to the original publication.

Lines 105f, 'Gulf of Naples (de Castro Coppa et al., 1980), the Alboran Sea': Should be 'Gulf of Naples (de Castro Coppa et al., 1980) or the Alboran Sea'.

Lines 167–169, 'Those nets sampled primarily 200 m depth, but also caught foraminifera during the net descent and ascent to the ocean surface, which both involve negligible towing and capturing time compared to the sampling at 200 m depth (Table 1).': This new information makes the data collected by the authors now fully unsuitable for any in-depth quantitative comparison with earlier studies. Compare General comments.

Lines 176f, ' These three parameters of the carbonate system were then integrated for the upper 200 m water depth.': Which does not make any sense when your specimens are basically exclusively from 200 m water depth.

Line 188, 'When necessary, samples were split into aliquots of 1/4 and 1/6': How do you split to $\frac{1}{6}$? All splitters I am aware of split a sample in two. So you get $\frac{1}{2}$ from the first split, $\frac{1}{4}$ by splitting one of the $\frac{1}{2}$ samples, and then $\frac{1}{8}$ when splitting one of the $\frac{1}{4}$ samples. There is no reasonable way to end up with $\frac{1}{6}$ using this technique.

Line 202, '*G. ruber*': Should probably be '*G. ruber* (white)'.

Line 207, '(±1 µg of nominal precision)': As I already requested in my first review, the measurement error must be analysed in more detail. A precision of 1 µg is rather low for such a study (normally it should be at 0.2 µg or below). Given that many measurement values in the appendix are only 2–4 times larger than the precision of the balance, this leads to theoretical errors of 25–50 %. The authors must show, that the measurement in triplicate is enough to yield meaningful results across all weights.

Line 217, 'Varimax rotation': Varimax is an orthogonal rotation method, which is not suitable if factors are correlated, which is likely the case here. The authors should either use an oblique rotation method (e.g. oblimin or promax), or at the very least show that

no correlation problem exists according to Tabachnik and Fidell (1996) (i.e. correlations between all relevant factors (PCs 1 and 2) after oblique rotation must be < 0.32).

Line 221, 'salinity (and to a lesser degree on $\left[{CO_3}^{2-}\right]$;': I would say salinity and carbonate saturation leaks into both axes.

Lines 249–250, 'no differences are observed between samples collected during day and night.': Consider showing this graphically.

Lines 280f, 'stations influenced by the incoming waters from the Atlantic and lowest $\left[{CO_3}^{-2}\right]$ values score highest.': But this makes no sense with the PCA. Highest scores on the second PC are correlated with high $\left[{CO_3}^{2-}\right]$ values, which also makes more sense with the description of the authors that stations close to the Atlantic indeed plot at the lower end of the second axis. Additionally, $\left[{CO_3}^{-2}\right]$ should be $\left[{CO_3}^{2-}\right]$ in the entire manuscript.

Line 288, 'and at stations where pH is higher': Where? The highest abundances of *G. bulloides* are all on the negative side of the second PC (stations 3, 5, 2), correlating with low pH values.

Lines 311-313, 'Similar growth patterns can be seen in *G. ruber* (white), *G. bulloides*, and *O. universa* with that correlation, graphically represented by the shape of a power function (Fig. S2).': This is a misconception by the authors, that I already noted again in their response to both Reviewer #1 and myself. You cannot call those curves growth curves, as they do not represent ontogenetic growth! You simply have a lot of individuals at different sizes, which can be at very different or very similar ontogenetic stages depending on their individual ontogeny. The size of adult Foraminifera of the same species can be very different, and all shells above 150 µm diameter can reasonably be considered adult and in a reproductive state (Peeters et al., 1999). What you have there is a diameter–cross-section scaling, and you have no idea how much of it is ontogenetic growth and how much is intra-specific variation.

Lines 313–315, 'Planktic foraminifera grow faster when they are younger and smaller (steepest in the lower left part of the regression line) and slower when they are older and bigger (less steep in the upper right part of the regression line; Fig. S2).': This is nonsense (see comment to lines 311–313), and could only be studied if you followed single individuals during their ontogeny by repeated measurements designs, or at least made a chamber-by-chamber analysis as in Caromel et al. (2015). What you see from the curves is rather the very trivial and predictable fact that a 1D and a 2D size measurement are linked via some kind of power function.

Lines 315–317: Those p-values must again be corrected for multiple testing, using for example corrections for the family-wise error rate (e.g. Bonferroni correction) or the false discovery rate (e.g. Benjamini and Hochberg, 1995). Since those analyses are basically the

post-hoc tests for an ANOVA, you may alternatively want to use the classical correction in that case, which would be Tukeys honest significance difference, as implemented in most packages. Certainly, all those pairwise comparisons should be performed for all possible pairs of groups defined in fig. 6, not just subjectively selected ones, as this artificially inflates the significance of the results. Compare my review on the original version of the manuscript for details.

Lines 317f, ' In the other two species *G. bulloides* and *O. universa*, a similar trend is observed regarding the two basins,': Then where are p-values to prove that?

Line 322, 'The long axis-to-weight relation': Those size–weight curves (most notably those shown in fig S4) are still not forced to go through origin, and the authors completely misrepresent the comments of both Reviewer #1 and myself in this regard in their response. Reviewer #1 wrote '... but disagree with assertion that the curves ... should not go through origin. That is physically impossible.' I wrote 'Especially when the authors argue that a zero-intercept model would not make sense because it would imply the existence of individuals with zero mass and size, is it not logical to them that non-zero-intercept model which allows a foraminifer to have mass at size zero or have a certain size without mass is even more problematic!'. In both cases, the authors effectively seem to believe (at least they argue as such in their response) that we would both agree that curves should **not** go through the origin. **This is demonstrably false!** Reviewer #1 seems adamant, that they have to go through origin. I would personally relax that necessity, depending on the purpose, as I already argued in my first review. If those curves are only supposed to be local approximations of a relationship, that should not be extrapolated, I am fine with them using simply the best fitting curve. However, it has been the authors who were constantly talking about those things as a growth curve, and it has been the authors who were arguing that a foraminifer of size and weight zero would not make sense, because the proloculus already has a certain size and weight. This implies that the authors believe the curves to be a biological model, that can be extrapolated across the reasonable size range of Foraminifera. As soon as this is the case, **the curves must be forced to go through the origin**, because otherwise they would allow a foraminifer with zero size and positive or negative mass, or zero mass and positive or negative size to exist. The authors must therefore either stop trying to sell these curves as some biological growth model, or at last apply proper regression to ensure a zero intercept of the fitted curves. Under no circumstances is it acceptable that they try to blatantly misinterpret what either of the reviewers criticised, just to fit their needs.

Lines 324f, '*O. universa* was finally discarded for comparisons between ρ_A at different locations due to a low area–weight correlation and no remarkable trend observable between locations (Fig. S4c; Fig. 3i);': I absolutely do not understand this argumentation. The correlation (i.e. R^2 value) is only marginally lower for *O. universa* (0.64) than for *G. bulloides* (0.69). The correlation between cross-sectional area and weight is probably significant for both species, if it is not then *G. bulloides*

is the problematic species, not *O. universa*. And the PCA of area densities shows no clearer signal in *G. bulloides* than in *O. universa*. Yet *G. bulloides* is used in the ensuing analyses, while *O. universa* is not. This makes absolutely no sense!

Lines 326–328, 'The eastern Mediterranean specimens are the lightest in both species (*G. ruber* (white), *G. bulloides*), with more extreme W–E differences in *G. ruber* (white) than in *G. bulloides* (Fig. S4d–e).': Any prove of this statement, while probably true judged by the graphic, is still missing though.

Lines 329f, 'The data of all the locations show a similar CV value.': Well, there is certainly some variation. Calculating the 95 % confidence interval for the coefficient of variation (e.g. using the method by Vangel (1996) or a bootstrap approach) could help to interpret those values more meaningful.

Lines 332f, 'highest median value ($1.55 \cdot 10^{-4}$ µg µm^{-2}) and IQR.': I believed we now finally established, that variation is always correlated with mean value due to stochastic reasons. This is why you should, and finally did, calculate the coefficient of variation. So why do you still compare IQRs here?

Line 337, 'seven locations were compared (Fig. 7).': Why are the locations not the same for *G. ruber* and *G. bulloides*?

Lines 338f, 'Specimens from the Atlantic have the lowest median ρ_A ($8.75 \cdot 10^{-5}$ µg µm^{-2}) and the smallest IQR': I believed we now finally established, that variation is always correlated with mean value due to stochastic reasons. This is why you should, and finally did, calculate the coefficient of variation. So why do you still compare IQRs here?

Lines 343f, 'Results show a less clear overall trend for *G. bulloides* than for *G. ruber* (white), with higher ρ_A associated with slightly higher pH in the eastern Mediterranean (Fig. 3h).': I actually fail to see any trend in *G. bulloides* at all.

Lines 397–416: Given the new evidence of your true sampling depth (i.e. only 200 m, instead of the upper 200 m integrated with only a bias towards the 200 m level) I consider all attempts of in-depth quantitative comparisons to be futile. However, you can very well keep the qualitative comparisons in the paragraphs before here and in section 5.2.

Line 434, 'Basinit': Should be 'Basin it'.

Line 449, 'abundanceat': Should be 'abundance at'.

Line 453, 'tempreatures': Should be 'temperatures'.

Line 487, 'Atlanticclose': Should be 'Atlantic close'.

Line 506, 'itdecreases': Should be 'it decreases'.

Line 550, '*G. bulloides*': Neither from fig. 3 nor fig. 7, I see any such trend for *G. bulloides*, and I am certain that a quantitative comparison of area density between the basins shows no significant differences either.

Line 569, 'The reason why the $\rho_{\mathbf{A}}$ of *O. universa* is particularly low and highly variable': Where do you take this information from. Fig. 3i shows that the area density of *O. universa* is the highest of all investigated species, and its spatial variation does not seem to be exceedingly high.

Line 574, 'The $\rho_{\mathbf{A}}$ of *G. ruber* (white) is only partly controlled by carbonate chemistry': Just by eye from fig. 3g, I would say that the highest correlation of *G. ruber* area density is with $\left[{CO_3}^{2-}\right]$ (i.e. diagonal trend from bottom right to top left).

Line 575, 'similar to *O. universa*': So now all of a sudden there is a trend in *O. universa*!?

Lines 575f, 'In contrast to *O. universa*, the $\rho_{\mathbf{A}}$ data of *G. ruber* and *G. bulloides* follow systematic correlations.': You have indeed more significant correlations in *O. universa* than in any of the other species, according to your table 2, so I do not know what you are talking about. Needless to say, that all correlations presented in table 2 must be corrected for multiple testing anyways, and I do not know how many will remain significant after this is done. For a further explanation of this I refer you to my review of the first iteration of this manuscript, where I already explained to you in great detail, why this is necessary and how it is done.

Lines 584f, 'The $\rho_{\mathbf{A}}$ of *G. bulloides* tests increases from the Atlantic toward the eastern Mediterranean.': I do see such a trend in neither fig. 3h nor fig. 7. This is also apparent from fig. S4, where the relationship between size and weight indicates that the area density is nearly constant: $3\,\mu g/30\,000\,\mu m^2 = 6\,\mu g/60\,000\,\mu m^2$.

Line 585: Delete 'In both species larger IQRs are found toward higher absolute $\rho_{\mathbf{A}}$ (Fig. 7).'

Line 596, 'weight-to-long axis relations': In the following paragraph you are comparing raw weights in a particular size fraction, not weight to size relations.

Line 603, 'comparison of the weight-to-long axis relation': In the following paragraph you are comparing raw weights in a particular size fraction, not weight to size relations.

Line 621, 'upper 200 m of the water column': According to your own words, you do not have data from the upper 200 m but only from more or less exactly 200 m depth.

Caption table 2, 'r-values in bold are significant at $p < 0.05$, *$p < 0.1$.': This is, as the analyses in the first version of the manuscript, again a case of hefty multiple testing. As such, all p-values of correlations per species would have to be corrected for this fact, either using family-wise error rate (e.g. Bonferroni) or false discovery rate (e.g. Benjamini and Hochberg, 1995) corrections! Compare my review to the original manuscript for an in-depth explanation, why this is necessary, and where to find further information. A graphical depiction of all significant correlations could further be very helpful. Please also note again, that r, p and all such variables need to be typeset in italics.

Table 2: 'Density area' should be 'Area density'.

Figure 5: For station 12, a sample size of 1 is given. This can hardly be right. How can one individual belong to different size classes (as depicted in the barplot)?

Figure 7: Why are the regions different for *G. ruber* and *G. bulloides*?

References

Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. B Met., 57, 289–300, doi:10.2307/2346101, 1995.

Caromel, A. G. M., Schmidt, D. N., Fletcher, I., and Rayfield, E. J.: Morphological change during the ontogeny of the planktic Foraminifera, J. Micropalaeontol., 35, 2–19, doi:10.1144/jmpaleo2014-017, 2015.

Peeters, F., Ivanova, E., Conan, S., Brummer, G.-J., Ganssen, G., Troelstra, S., and van Hinte, J.: A size analysis of planktic Foraminifera from the Arabian Sea, Mar. Micropaleontol., 36, 31–63, doi:10.1016/S0377-8398(98)00026-7, 1999.

Pujol, C. and Vergnaud Grazzini, C.: Distribution patterns of live planktic foraminifers as related to regional hydrography and productive systems of the Mediterranean Sea, Mar. Micropaleontol., 25, 187–217, doi:10.1016/0377-8398(95)00002-I, 1995.

Rebotim, A., Voelker, A. H. L., Jonkers, L., Waniek, J. J., Meggers, H., Schiebel, R., Fraile, I., Schulz, M., and Kučera, M.: Factors controlling the depth habitat of planktonic Foraminifera in the subtropical eastern North Atlantic, Biogeosciences, doi:10.5194/bg-2016-348, in press.

Tabachnik, B. G. and Fidell, L. S.: Using Multivariate Statistics, Harper Collins, New York, 3rd edn., 1996.

Vangel, M. G.: Confidence intervals for a normal coefficient of variation, Am. Stat., 50, 21–26, URL http://www.jstor.org/stable/2685039, 1996.