

Our response to the editors comments (29.08.2014):

Hi Andy,

thank your very much for your comments.

We provide a new version of the ms in which we respond on your suggestions as follows:

(1) We added a paragraph on the ALK-DIC metric, see section 2, p 5 (in bold).

(2) We provide a paragraph on the interpretation of Fig. 11 (model evaluation) at the end of section 6, p 18/19 (in bold). We also state more clearly in the paper its focus: method assessment rather than model evaluation (e.g. at the end of the introduction).

(3) We include some of the discussion on PALK-PALK0 and acknowledge the respective suggestion of reviewer 2 in section 2, p 8/9 (in bold).

(4) We have carefully checked the manuscript and improved it at various places (some marked in bold, others not).

We hope that this version is ready for publication.

Best regards, Wolfgang Koeve (on behalf of all co-authors),

Editors comments:

Comments to the Author:

Hi Wolfgang,

Thank you for submitting your revised paper. I really enjoyed the original submission which provided a helpful analysis of ALK evaluation metrics and made good use of your new preformed ALK tracer. The reviewers added some thoughtful suggestions for improvements and hence a clearer paper with greater potential uptake by the community.

I must therefore admit to being a little disappointed, despite a thorough reply to the reviewers, that you made so few and so minor changes to the manuscript. Unfortunately, I don't think you have quite yet adequately address these in the manuscript itself. Please go back to the reviews and your replies, and see where your manuscript could be further clarified and usefully incorporate some of the discussion arising (e.g. as in your replies).

For instance:

(1) While I agree with your reply that the ALK-DIC metric was not intended as an ALK evaluation metric and hence contrary to John Dunne's suggestion, given the apparent common use of ALK-DIC and potential confusion in the context of your manuscript, you do need to include e.g. a short paragraph of discussion regarding it (even if you do not calculate it and carry out any in-depth intercomparison) and cite the relevant literature.

(2) You need to bring out a little more from the CMIP5 model tests – I agree with points 3 and 4 of John Dunne. Method evaluation is not entirely separable from model evaluation, particularly here as you include multi model results in 2 different places. In diagnosing why models differ in a particular method, you additionally elucidate and test the method itself. I hence do not see this as 'diluting' your focus or intention.

(3) Regarding the main 'Comment' of Referee 2: could you not include some of this discussion and calculations in improving the manuscript? You call it an 'interesting suggestion' yet nothing appears in the revision. If both you and the Referee find it interesting, could you not share this with all readers?

(4) Please see where else clarifications and/or additions, perhaps based directly on some of your replies, might be incorporated in order to create an improved manuscript with maximum clarity.

This should not take you too long nor overly hold final publication up, as you have already clearly thought about the reviewers comments and provided a reasonably compre-

hensive response to many of them. I look forward to seeing final product.
andy

Original reviews and responses:

General response: First, we like to thank the referees for their time, comments, constructive criticism and helpful suggestions. (Reviewer comments in italics, our answers in normal font.) (Please note that we provide a slightly updated version of our final response to the discussion here. In particular we have added hints to where in the manuscript text (page numbers in pdf version) we have made the respective changes. In the manuscript itself we have highlighted changes (new text) in bold font and deleted text as struck through. In addition to these marked changes the text has seen minor (editorial) changes of the wording where appropriate.)

20. June 2014, W. Koeve on behalf of all authors.

J. Dunne (Referee john.dunne@noaa.gov Received and published: 22 January 2014

Comment:

General Comments

The manuscript "Evaluating CaCO₃-cycle modules in coupled global biogeochemical ocean models" by Koeve et al. assesses the utility of a variety of derived water column metrics of alkalinity for calcium carbonate cycling, describing the difficulty of deconvolving biases in physical and biogeochemical pathways, and concluding that the Alk method is considerably more robust than others in the ability to parse out the preformed (Alk₀), remineralized organic (Alkr), and dissolved CaCO₃ (Alk*) contributions through an uncertainty analysis in an offline model. The authors then apply the method to three coupled physical-biogeochemical models to assess their representation of Alkr and Alk* and find that while the models all represent Alkr fairly well, they have large differences in their representation of Alk* both compared to the observational estimate and each other. The major strengths of the manuscript are in the detailed fashion the authors explore the complications associated with interpreting alkalinity variations in observations as they confound model verification and how they arrive at a concrete recommendation.*

Comment: *The major weaknesses of the paper is that (1) the authors fail to incorporate into their analysis comparative assessment of a common approach (Alk-DIC, or "excess alkalinity") used to verify CMIP5 class models (e.g. Seferian et al., 2013; Dunne et al., 2013),*

Answer: The focus of our paper is the evaluation of methods (metrics or properties) which can be used for the data-based evaluation of CaCO₃-cycle ocean models. Is the property ALK-DIC a potential candidate for this purpose? Firstly, neither Seferian nor Dunne suggest to use it for that purpose. Secondly, considering biogeochemical modifications in the interior of the ocean, ALK is mainly modified by CaCO₃ dissolution, while DIC is influenced by both the organic tissue pump and CaCO₃ dissolution. The property ALK-DIC will hence show a combined response to both processes which may even vary with depth or ocean basin (North Pacific vs. North Atlantic). Unfortunately, we are not able to fully test this question in this paper and within the rationale of our approach, since we have no preformed DIC tracer available from our model runs. We instead decided to present here some of our unpublished work (Koeve et al. in prep) with CMIP5 model output. Comparing the global patterns of ALK-DIC and of AOU on the 2000m depth

level, we find large similarities (see suppl. Figs 1 and 2 of this rebuttal). It is obvious from this comparison that the organic tissue pump dominates the patterns of ALK-DIC. In the IPSL model, for example, AOU explains 88% of the variations of ALK-DIC while TA* (i.e. CaCO₃ dissolution) explains only 4%. From this preliminary analysis we conclude that ALK-DIC is not a property particularly suitable for the evaluation of the CaCO₃ cycle in models. We plan to present this material in more detail in a follow-up paper on comparing different CMIP5 models with observations (see below).

Comment: (2) choose a CaCO₃ dissolution scheme that puts most dissolution in the upper water column rather than below the saturation horizon,

Answer: We have stated this limitation more explicitly in the revised methods section (page 11) and the conclusions section (page 19) of our paper. However, keeping in mind the clear focus of our paper (EVALUATION of METHODS that can be used for data-based evaluation of CaCO₃ cycle models), it may be sufficient to use a model that to first order reproduces the vertical patterns of CaCO₃ dissolution (see Fig. 7a).

Comment: (3) fail to describe the factors leading to the differing Alk* distributions in the three OCMIP5 models, and (4) similarly fail to come to any conclusions as to the relative strengths and weaknesses of these model formulations.

Answer: We like to emphasize again that the major objective of our paper is the evaluation of methods (metrics or properties) which can be used for the data-based evaluation of CaCO₃-cycle ocean models. For that purpose we compare in particular certain computed properties with respective idealised tracers. At the end of the paper we apply the TA* method also to three models for which output is freely available and present global mean profiles of TA and TA* from these models and the GLODAP database. These models serve only as examples and it is not the objective of our current paper to discuss individual strengths and weaknesses of the respective model formulations. We understand the interest in the actual MODEL EVALUATION, but we think that including this aspect in our paper would dilute the clear focus on METHOD EVALUATION in particular since this necessary first step has not been done before. **To better emphasize our focus we have slightly changed the title of the paper from 'Evaluating CaCO₃-cycle modules in coupled global biogeochemical ocean models' to 'Methods to evaluate CaCO₃-cycle modules in coupled global biogeochemical ocean models'.** We are currently preparing a follow up paper in which we apply, among others, the TA* method to a range of CMIP5 models and discuss the patterns of TA* along side the patterns of CaCO₃ production (export) and saturation/undersaturation characteristics.

Comment: *These weaknesses aside, I think the present manuscript provides a very thoughtful and helpful contribution to advance the verification and future development of ocean CaCO₃ models and would hope that the authors are planning on a more targeted application of this method to assess OCMIP5 and CMIP5 models in a following manuscript.*

Answer: As stated above (and also demonstrated by suppl. Figs. 1 and 2), we are working on this.

Specific comments

6120,5 - *Where does the 45% come from, a specific depth horizon? For a maximum, I get more like >95% remineralized PO₄ in the N Pacific.* **Answer:** We are referring to the fraction of remineralised PO₄ to total PO₄. In the interior ocean observed PO₄ can be thought of being composed of two fractions, preformed PO₄ and remineralised PO₄ stemming from the degradation of organic matter. Waters inhabiting the deep North Pacific are dominantly ventilated in the Southern Ocean known for its high winter-time (and hence preformed) PO₄ concentrations (see e.g. Broecker et al 1985, JGR; Duteil

et al., 2012, BG). The highest fractions of remineralized PO₄ are to be expected under conditions where all oxygen has been used up for the degradation of organic matter, and also in shallow waters which outcrop under conditions of very low PO₄. Considering the latter case (very low PO₄ at outcrop) one can of course observe fractions of up to 100%. Hence the signal strength which we discussed on page 6120 can be even larger than quoted by us. However, such waters are not important volume-wise in the global ocean. We have been more specific in the revised version of the paper (**page 3/4**).

6120,15 - suggest replacing "are not from the" with "reflect not only" and removing "alone". **Answer:** Has been changed in the revised version (**page 4**).

6121,13 - This is where I would be interested in seeing a comparison with the Alk- DIC metric used in Figure 9 of (Seferian R., Bopp L., Gehlen M., Orr J. C., Ethe C., Cadule P., Aumont Olivier, Melia D. S. Y., Voldoire A., Madec G. Skill assessment of three earth system models with common marine biogeochemistry. In : Presentation and analysis of the IPSL and CNRM climate models used in CMIP5. *Climate Dynamics*, 2013, 40 (9-10), p. 2549-2573.) and Figure 5 of (Dunne, J. P., and Coauthors, 2013: GFDLs ESM2 Global Coupled Climate—Carbon Earth System Models. Part II: Carbon System Formulation and Baseline Simulation Characteristics*. *J. Climate*, 26, 2247–2267. doi: <http://dx.doi.org/10.1175/JCLI-D-12-00150.1>). Admittedly these assessments were more generally configured for evaluation of anthropogenic carbon uptake rather than specifically the CaCO₃ models within them, but it would provide a helpful critical assessment for feedback to these groups. **Answer:** See our response above and the suppl. Figs. 1 + 2. We postpone a more detailed discussion of the metric ALK-DIC to our work in progress on CaCO₃ cycle in CMIP5 models.

6122, 2 and 11 - Can you provide the r^2 for these to help the reader quantitatively assess the pattern similarity. **Answer:** We agree and provide statistics in the **caption of Figs. 3 and 4**, similar to what we did for Fig. 2.

6123,5,9 - I am having trouble interpreting what the authors mean by "spurious" and why "the salinity-normalized TA0-anomaly should be constant everywhere" Do they simply mean that the distribution is not helpful for CaCO₃ cycling assessment as it reflects surface compositional differences determined other than through Salinity differences? It is clear that 6a and 6c look similar and 6b and 6d look similar and that some is preformed, but I am not sure how any of the distributions are "spurious". **Answer:** Considering also the comments of Reviewer 2 concerning our discussion of Fig. 6, i.e. the concepts of potential alkalinity and Howard's metric, we have reworded this section (**pages 6-7**). For details see our respective response to Reviewer 2 below. Also, our use of the word 'spurious' was probably not to the point and has been removed.

6125,21 and 6130,1 - In acknowledging that undersaturation during water mass formation leads to AOU overestimating oxygen utilization by 20-25%, more detail is warranted here, particularly with respect to why it was ignored and the consequences to the underlying uncertainties in Alk*... i.e., that it leads to an overestimation of Alk* of about 3 mmol/m-3 or 5%. **Answer:** In the paper of Duteil et al. 2013 (BG) we analysed a number of models in which we implemented idealized oxygen utilization tracers (TOU) and found that, on average, AOU computed in the classical way of assuming oxygen saturation at outcrop, actually overestimates the time integrated imprint of oxygen utilization. In Duteil et al. 2013 we further developed a new pragmatic metric (EOU) which performed much better in our models. We applied this metric also to observations and found that globally AOU overestimates true oxygen utilization by about 20-25%. From this finding one can deduce a respective effect on TA* estimates of the real ocean.

The difference between AOU and TOU was variable between models. In fact, the model

we used in the study under review here, showed a very small difference between TOU and AOU. Considering the correct determination of TA0 to be the major challenge in the TA* approach we decided to isolate this uncertainty and to use TOU instead of AOU when quantifying TAR in our paper.

We agree with the reviewer that in the actual application of the TA* approach for a data-based model assessment for example of CMIP5 models uncertainties of the TAR term will have to be considered. We will pick up on this in our follow-up paper and we mentioned this challenge in the conclusions of the current paper (**page 17/18**).

6127,8 - Representation of dissolution as an exponential scaling is extremely crude in ignoring the role of saturation state and is a potentially important limitation of the error analysis in its applicability to the real world. This structural limitation should be acknowledged. **Answer:** This structural limitation has been acknowledged in the revised version (**page 19**). However, we note that on the global scale our simple dissolution formulation gives a reasonable global mean profile of TA* compared with that deduced from observations (Fig. 7 green line). **See also new text in the methods section (page 11).**

6130,9 - Discussion of Figure 7 - What is TA0ub? Why are there no observational lines in the regional plots? **Answer:** TA0ub stands for unbiased sampling, i.e. sampling during both the winter and summer seasons. This was mentioned in the original caption but without using the abbreviation. **We have updated the caption of Fig. 7 and the text on page 13.**

The absence of observational lines in the regional plots is intentional. In this paper the ground truth is the solid black line (the TA* tracer). The algorithm to compute TA* is successful if the computed TA* matches the model's true tracer TA* as well as possible. Showing the TA* estimates for the observations also for all ocean basins may mislead some readers and shift the focus from evaluating the METHOD to evaluating our model against the observational TA* estimate.

6132, 10 - Some description of the CaCO3 dissolution formulations in these models is warranted. **Answer:** Again we hesitate to discuss the three example models in more detail. We think that this would turn a METHOD EVALUATION paper more and more into a MODEL COMPARISON paper. Given the remarks of the reviewer concerning the ad hoc dissolution scheme used in our model, however,

we have included a brief statement that UVIC applies a similar ad hoc dissolution parameterization while still showing the best comparison with TA* derived from observations Fig. 11 and in Tab. 4. Advantages and disadvantages of certain dissolutions schemes in models will be discussed in more detail in our paper on CaCO3 cycle in CMIP5 models (Koeve et al. in prep).

6132, 21 - Table 4 versus Figure 11 - Table 4 has the UVIC model as slightly higher than GLODAP, but in Figure 11, it looks like the UVIC line is always at or below the GLODAP line, how does that work out? **Answer:** Table 4 shows global inventories while Fig. 11 shows global mean (concentration) profiles of TA*. As we noted in the caption of Tab. 4, the ocean volumes covered by GLODAP and the three models are different. In particular GLODAP is not defined north of 65N and hence has a smaller volume compared to the models. Correcting for this effect would give a higher inventory for GLODAP consistent with Fig. 11.

6133, 21 - What is causing these differences? **Answer:** As stated above, it is not the objective of this paper to discuss in detail why the three models show certain characteristics, for example with respect to TA* and certain Omega values.

6133, 24 - suggest replacing "compared with the representation of the organic tissue pump" with "than their organic tissue pump modules" **Answer:** Ok. (**p 17**)

Comment:

General Comments

This study aims to diagnose CaCO₃ cycling from modeled or observed alkalinity distributions. It evaluates different methods that have been proposed and used before for this purpose. It concludes that the TA method is superior and that other methods, such as the potential alkalinity (PALK) method are "disqualified" by the analysis. I think the evidence presented to support this assertion is insufficient and unconvincing. The authors arrive at their conclusion using incorrect assessments. E.g. on page 6123 lines 5-6 they state that "the salinity normalised TA0-anomaly should be constant everywhere." In other words, the author's expectation is that it should be constant. But I think it cannot reasonably be expected that the TA0 anomaly should be constant. Since TA0 is affected by CaCO₃ and organic matter (OM) production its surface distribution will be not equal to salinity and therefore one cannot expect its interior distribution to be equal to salinity either. This statement is followed by assessing the patterns of salinity normalized TA0 as "spurious" (line 9). The assessment of PALK is similarly flawed (lines 19-21). Again, since surface distributions of PALK0 will be different from salinity distributions due to production of CaCO₃ its interior distributions will also be different. So, it CANNOT be expected that PALK0 displays a uniform distribution.*

Answer: We agree that this part of our arguments needs improvement. Certainly CaCO₃ (and organic matter) production will affect TA0 and PALK0 and hence a simple relationship to salinity is not to be expected. Perhaps even more important is the upwelling of waters that have seen strong modifications of alkalinity in the deep ocean. This is particularly obvious when waters returning from the North Pacific upwell in the Southern Ocean. These waters carry the imprint of CaCO₃ dissolution and nutrient release on alkalinity. In the revised version we have completely reworded the text of pages 6122 and 6124 (**see p 6/7 of new ms**) and provided a more consistent line of arguments.

Comment: *In fact I wonder if not the PALK part from dissolution and TA* are very similar. I suggest a simple analysis by calculating the PALK component resulting from dissolution only as $PALK_{dis} = PALK - PALK^0$. In Fig. 6 it would be the difference between panels c) and d. Plot this against TA*, e.g. as global average vertical profiles, or basin wide profiles or sections. This is my mayor comment.*

Answer: This is an interesting suggestion. Ignoring the effect of denitrification and assuming N/P and O₂/P ratios of 16 and 170 the property PALK_{dis} proposed by the reviewer can be rewritten as

$$PALK_{dis} = (TA - TA^0 + NO_3 - NO_3^0) / S * 35 \text{ or}$$

$$PALK_{dis} = (TA - TA^0 + NO_3remin) / S * 35$$

This is similar, though not identical to the correct relationship (re-arranging of our equation 1):

$$TA^* = TA - TA^0 + NO_3remin * R_{TA:NO_3}.$$

The term $NO_3remin * R_{TA:NO_3}$ is very similar to our term TA^r which we estimate from AOU. The major difference is that we, in agreement with the literature, use a $R_{TA:NO_3}$ of 1.26 (instead of 1) and that we do not perform a salinity normalization of TA*.

Hence, we agree with the reviewer that the property PALK_{dis} would have patterns very similar to TA*. Calling for a property like PALK_{dis} hence supports our choice of TA* for the evaluation of the CaCO₃ cycle very much. Both for TA* or PALK_{dis}, the essential step (and challenge) is the correct determination of TA⁰. (We stick to using TA* since that term has been introduced to the literature which PALK_{dis} has not.

Comment: *Other than that I think the manuscript is well written and presents interesting and new material, which advances the understanding of ocean biogeochemistry by introducing and testing different analysis methods.* **Answer:** We thank the reviewer for his stimulating words.

Comment: *Minor comments are embedded directly into the manuscript.*

Answer: We received the supplementary material (manuscript with detailed comments) on request from the editorial office. Most of the remarks refer to points raised by the reviewer already above. We do not repeat these here. In the following we have extracted, and respond to, the other comments. In some cases, text was highlighted but no comments were provided. We ignore these.

6129, 14: Hemispheric sampling: how is this done in detail? **Answer:** We have provided the requested details in the final version of the ms (**page 12/13**).

6132, 23-26: MPI-ESM. I don't agree with this assessment. The vertical gradient in the observations is > 80 , whereas in MPI-ESM is less than 40 mmol/m^3 . **Answer:** We refer here to the structural elements of the observations discussed earlier on that page (shallow maximum, minimum at about 500m, broad maximum in the deep ocean). We are now provide a more detailed discussion on the differences (overall gradient) of MPI-ESM relative to observations in the final ms version (**page 16**).

6133, 20-21: reference or explanation missing **Answer:** We now provide a more elaborate explanation about the expectations from dissolution chemistry, i.e. that TA^* should be largest in the waters undersaturated with respect to both calcite and aragonite and provide respective references in the final paper version (**page 17**).

We close this final response by again thanking both reviewers for their time and helpful comments.