Review of OS manuscript egusphere-2025-3588

Title: Metrological concepts applied to Total Alkalinity measurements in seawater: reference materials, inter-laboratory comparison and uncertainty budget

General comments

The authors have characterized a reference material (RM) prepared from artificial seawater by means of gravimetric measurements and quantified its measurement uncertainty using a bottom-up approach in accordance with the GUM, taking into account the homogeneity and stability of the RM. To validate the results, a comparison measurement was carried out. In addition, another RM based on natural seawater was investigated. In this case, a top-down approach was applied to quantify the measurement uncertainty based on the results of a second comparison. Here as well, homogeneity and stability were assessed. This RM has been characterization using a common TA measurement procedure. The RMs and the corresponding measurement results and their interpretation provide a first metrological evaluation of the measurement uncertainty of TA measurements, which is currently lacking. The manuscript is therefore addressing a question of scientific interest and is relevant for reliable measurements of seawater TA. Therefore, publication of the paper is recommended/accepted. Nevertheless, a revision is necessary.

Before publication, the paper must address the following issues:

1. The traceability of the TA values and their uncertainties - which must be the core element of any RM, particularly if the authors, as they claim, worked in accordance with ISO 17034 – are inconsistent to some extent and are not sufficiently discussed. As the authors have noted themselves, TA measurements using the conventional method according to Dickson's Guide have no proper traceability. A detailed analysis of its traceability is currently lacking. Nevertheless, the authors use this method to characterize TA of the natural seawater TA. Moreover, even though the gravimetric approach to characterize the artificial RM is traceable to the SI, the quantification of impurities in NaCl is conducted with the common method, which introduces an inconsistency into the characterization. Additionally - as the authors themselves note - the artificial seawater differs chemically from natural seawater. This also affects the regression method used to determine TA and thus its associated measurement uncertainty. Consequently, measurements of seawater TA that are referred to the artificial RM may still contain significant biases, even though the comparison measurements show reasonably good agreement.
   Beyond these more fundamental issues, the overall concept of traceability that the authors seem to have in mind is not presented clearly. Traceability can fundamentally be established only to one metrological reference—either to the artificial or to the natural seawater RM. However, two very different RMs are introduced: the artificial RM, characterized gravimetrically, and the natural seawater RM, measured using the standard method according to Dickson's SOP 3a/b. The first is proposed to establish SI

traceability while the latter is proposed to serve as an additional reference for quality control. However, what if this kind of quality control provides a deviating result, which is the RM to believe? In turn, if both RM provide compatible results, why is there a need for second kind of RM?
The manuscript is somewhat vague with respect to those traceability issues. The authors are not expected to solve these difficult, general problems of traceability in this paper, which would likely be beyond the scope of the study that mainly aims to evaluate the RMs, which has sufficient value in its own. Nevertheless, they must discuss and contextualize both RMs in light of these traceability challenges and clearly define their respective limitations.

2. The structure of the paper should be reconsidered. The typical format—theory, results, discussion—makes this paper rather confusing, since each of the three sections successively addresses several different topics. There are two RMs, two characterizations, homogeneity and stability studies, and two comparison measurements. For each, the theoretical background is first explained, then all results are presented together, and finally everything is revisited for discussion. As a result, the reader easily loses track of which parts belong together and is constantly forced to flip back and forth between sections.
This is not a mandatory requirement to be addressed for publication. However, to the reviewer's opinion readability would be improved if the authors restructure the paper by addressing the artificial seawater RM first—covering its preparation, characterization, stability, and homogeneity, including the relevant calculation principles and results, and concluding with the discussion. The same should then be done for the natural seawater RM. Finally, the comparison measurements should be presented, allowing both RMs to be contrasted, and a proposal for traceability should be discussed in more detail, also from a practical perspective.
Additionally, the paper is rather long, considering that it essentially evaluates RMs using well-established procedures. The authors should consider shortening it to some extent.
3. The reviewer recommends using an LLM-based AI to improve the language. In parts, the paper is difficult to read due to linguistic weaknesses, which the reviewer did not further correct.

Specific technical comments

line	**comment**
45	How do uncertainty limits illustrate climatic variations in TA? Please rephrase for more clarity.
54	"Not fully traceable" is a strong statement that requires discussion. If characterized HCl was used for the titration, why is the measurement not traceable to the SI? This should be discussed — if not in the introduction, then elsewhere in the manuscript.

56,57	The metrological terminology is somewhat imprecise. Comparability of results is achieved through traceability to the same metrological reference, not through uncertainty. Measurement uncertainty defines the limits within which differences between measurement results — or their equivalence — become meaningful (see also VIM: compatibility). Only deviations exceeding the measurement uncertainty can be regarded as significant. Similarly, the term "uncertainty of a measurement method" is incorrect — a method itself has no uncertainty; only a measured value has one. The authors should also verify whether the word "trueness" in line 56 expresses what they intend to say. According to the VIM, *trueness* refers to "the closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value." I am not sure, if this meant.
59	ISO Guide 35 has been replaced by ISO 33405. A paper related to metrological science should not refer to outdated standards.
76	Section 2.1: The purpose of this brief summary of Dickson's SOP 3b is not clear. Usually, reference to Dickson's Guide would be suffice, all the more, the paper is already quite long. If there is a reason for the repetition, it should mentioned. I assume the formulas are mentioned because it is relevant for the uncertainty calculations in subsequent sections?
164	The measurement result at zero NaCl mol/kg sol is shown …
164	Figure 2 is mentioned in the main text before Figure 1. Figures should be cited in the order of their appearance.
165	Replace "theory" with "reasonable assumption."
164	Even with goodwill, Fig. 2 does not support the assumption of a linear relationship passing through the origin. It rather shows a square root like behavior, which is difficult to explain. Alternatively, the ΔTA values at 1, 2 and 3 mol/kg NaCl solution content indeed suggest that there is a linear relationship - which one can expect in dependence of NaCl content – but with an offset at zero NaCl content. Which raises the question, why the measured ΔTA value is zero at zero NaCl content? I suspect, that the reason for this discrepancy can be found in the different metrological references involved in the gravimetric and measured TA values. See also comments related to lines 236 and 522. Anyhow, the linear extrapolation might be used as a rough estimate for the background alkalinity. However, the authors must comment on the difficulties I have mentioned.
168, 170	The purpose of measuring practical salinity and dissolved nutrients should be stated.
205	A reference to ISO 33405 would be more appropriate here.

206	The authors claim to evaluate the proposed RMs in accordance with ISO 17034. If so, they must fulfill the experimental requirements for short-term stability testing. Using a single, undefined transport of the RM does not meet these requirements. Since this uncertainty contribution can presumably not be readily quantified, I recommend refraining from claiming that this value has been determined. Instead, it should be stated that the value represents a first estimate, while a proper evaluation according to ISO 17034 is still pending.
214	"ISO Guide 35" — see comment on line 59.
236 et seq	It is unclear whether the calculation of the bias introduced by NaCl impurities is used solely to correct the reference value or to quantify its contribution to the uncertainty. This must be explicitly stated to avoid confusion. In any case, the approach appears to lead to a circular argument regarding traceability. The authors aim to correct the bias and/or assign a corresponding uncertainty to the RM. To do so, they measure TA and subtract this value from the one obtained via gravimetric measurements. However, in order to measure the TA with proper traceability, they would need a characterized RM traceable to the same metrological reference as the artificial RM — which is not available except for the proposed one. If, as assumed, the authors used Dickson's SOP to measure TA, then the traceability of the bias/uncertainty is subject to the same limitations inherent to that SOP (as mentioned in the introduction). Thus, traceability of the assigned TA value of the artificial RM, or its uncertainty, respectively, is questionable. Fundamentally, the bias/uncertainty must be quantified independently of the RM it is intended to characterize and with respect to the same metrological reference.
245	"Systematic uncertainty sources, such as those arising from the device, the operator, or the procedure, are cancelled here." This statement is not self-evident. Which uncertainties cancel out, and how are they correlated? The authors should explain this important aspect in more detail.
271	"It was chosen to neglect the within-bottle homogeneity." Again, the authors claim compliance with ISO Guide 35 (ISO 33405, respectively); however, their homogeneity analysis appears superficial to some extent. A one-way ANOVA must be applied to account for both within-unit and between-unit homogeneity. One might decide to disregard within-unit uncertainty for the reasons mentioned by the authors. In that case, only between-bottle homogeneity should be calculated according to the (corrected) Eq. 11. Otherwise, a proper one-way ANOVA analysis is expected for the homogeneity values given in Table 4. The authors must also evaluate the repeatability standard deviation of the homogeneity with respect to the target uncertainty (see Section 7.5.1 of ISO 33405).
274	Equation 11 is incorrect. In the simplest approach, assuming within-unit homogeneity,

	$u_{\text{hom}} = M_{\text{between}}/n_0$, where $M_{\text{between}}$is the mean square of the TA results of the units and $n_0$is the number of measurements per unit (assuming they are equal for each unit). See, for example, Section 7.7.3 and Annex B1 of ISO 33405.
281, 282	The equations are mutually inconsistent. u_{stab}cannot comply with both Eq. 12 and Eq. 13. I would recommend referring to Eq. 11 in Section 8.7.3 and Annex B3 of ISO 33405 instead.
296	Improve clarity: Do the authors mean that the artificial RM and the stabilized natural RM were prepared at approximately the same time, or one after the other?
301	Replace “calibrated” with “characterized.”
390	“… and is included in the reference values given above.” The meaning is unclear. The authors should be more precise: do the TA values in Table 3 really include the TA contribution from NaCl impurities, meaning the bias has not been corrected, or do they mean it has been considered, meaning the values in Table 3 have been corrected for this bias? See also the comment on line 236.
404	Table 4: The authors should add the units more precisely, as not all quantities are given in µmol/kg.
404/406	Table 4: It was mentioned that within-unit homogeneity was neglected; nevertheless, corresponding values are shown. Moreover, the results suggest that within-unit variability is even larger than between-bottle homogeneity, which seems unlikely. This supports the recommendation that measurement repeatability should be assessed in relation to the evaluation of homogeneity, stability and target uncertainty (also see comment on line 271).
409	“Its stability has been studied …” I recommend adding a figure to illustrate the stability results.
458	Results should not be excluded solely for statistical reasons, especially when only a small number of participants is involved. Have potential causes of deviation related to the measurement itself been investigated?
475	Table 6: Using Eq. 16, the values for $s_L$and $s_r$given in Table 6, $n = 3$, $p = 4$, and $u(\mu) = 1.08$, I calculate $u(\Delta) = 1.40$ μmol/kg. The authors should verify the values, or clarify which numbers were used in their calculation.
483	Tables 6 and 7 are nearly identical. I recommend combining them.
510	Since natural seawater consists of around 90 % NaCl, it is unlikely that the difference in composition between natural and artificial seawater could account for a tenfold discrepancy between the expected and observed differences in practical and absolute salinity. In any case, it is unclear why this matters. The authors should clarify the relationship between TA and salinity to make the relevance of this discrepancy for the study apparent.

522	I doubt the validity of this method for the reasons already mentioned in the comment on line 236 et seq. It would be more appropriate to quantify the impurities affecting TA using independent measurement methods. One cannot use the same instrument or procedure intended to be calibrated with the RM to determine the bias of that RM — this constitutes circular reasoning. It becomes impossible to distinguish whether the bias originates from the RM itself (e.g., NaCl impurities) or from the instrument or measurement procedure. For instance, if the bias depends on the ionic strength of the RM, demonstrating that the ΔTA at zero NaCl is zero does not resolve the issue.
551	Why only "linear": There could be other types of dependencies.
553	"However, it does not allow for the accuracy verification of TA values obtained using the nonlinear least-squares regression method …" This statement is not incorrect, but it is misleading, as it implies that the evaluation method itself is the cause of the problem. In fact, if an RM with an assigned value is available, the evaluation method is not critical — any method-related bias can be compensated by the known value of the RM. The real issue lies in the different chemical composition of natural versus artificial seawater. This difference necessitates using an evaluation method (NLLS) that differs from the one used to assign the value to the artificial seawater RM (Gran's method).
555	If a natural seawater RM is needed anyway to measure natural seawater, what is the benefit of using the artificial seawater RM?
558	"Having a natural seawater reference material that is easy to collect during open-ocean oceanographic cruises …" I find it difficult to see how this proposal could be implemented in practice, or what its benefit would be. Which institution would characterize such an *in situ* RM prepared by the operator during a cruise? And if that were feasible, why would the user rely on any other RM? If the operator is capable of characterizing an RM, they could directly apply the same method to measure their samples.
563	"… method's limited precision": Where has this been discussed? As mentioned above, this evaluation should indeed be addressed (following the guidance in ISO 33405.
575	The purpose of the DIC measurements is not stated. I assume they were intended to demonstrate that the carbon content did not change over time. Consequently, any instability of the RM must result from sources other than carbon, such as silicates. The authors should not leave it to the reader to infer the reasons for including specific results in the investigation.
586	"… lack stability": A good observation that appropriately addresses the scope of the paper.
587	"… indicating potential secondary processes influencing alkalinity." Such as? It is indeed a peculiar finding that the TA results do not reflect the increase in

	silicate. Identifying secondary processes in natural seawater may be difficult because of its complex composition. However, the composition of the artificial seawater is known—except perhaps for the NaCl impurities—so, an evaluation of the discrepancy should at least be feasible for the artificial RM.
599	The potential failure should also be mentioned in the results section (see comment on line 458).