LIMNOLOGY
and
OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 14, 2016, 268–277

doi: 10.1002/lom3.10087

Locally interpolated alkalinity regression for global alkalinity estimation

B. R. Carter,*[1,2] N. L. Williams,[3] A. R. Gray,[4] R. A. Feely[2]

[1]Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, Seattle, Washington
[2]Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric Administration, Seattle, Washington
[3]College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon
[4]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, New Jersey

Abstract

We introduce methods and software for estimating total seawater alkalinity from salinity and any combination of up to four other parameters (potential temperature, apparent oxygen utilization, total dissolved nitrate, and total silicate). The methods return estimates anywhere in the global ocean with comparable accuracy to other published alkalinity estimation techniques. The software interpolates between a predetermined grid of coefficients for linear regressions onto arbitrary latitude, longitude, and depth coordinates, and thereby avoids the estimate discontinuities many similar methods return when transitioning from one regression constant set to another. The software can also return uncertainty estimates scaled by user-provided input parameter uncertainties. The methods have been optimized for the open ocean, for which we estimate globally averaged errors of 5.8–10.4 μmol kg^{-1} depending on which combination of regression parameters is used. We expect these methods to be especially useful for better constraining the carbonate system from measurement platforms—such as biogeochemical Argo floats—that are only capable of measuring one carbonate system parameter (e.g., pH). It may also provide a useful way of simulating alkalinity for Earth system models that do not resolve the tracer prognostically.

An emerging strategy for monitoring the ocean carbon cycle involves using sensors on profiling floats. The primary advantages of this strategy are significant cost savings relative to shipboard measurements and the possibility of extending data coverage to regions and seasons that ships cannot routinely access with current resources. However, while float-capable sensors can now measure several biogeochemical properties including oxygen (O_2) and nitrate (N) (e.g., Johnson et al. 2010), float sensors can only currently measure one of two carbonate system parameters required to constrain the carbonate system in seawater. Ion-Sensitive Field Effect Transistor (ISFET)-based sensors now allow pH measurements on moorings, autonomous underwater vehicles (AUVs), and profiling floats (e.g., Johnson et al. 2012, unpubl.; Bresnahan et al. 2014; Talley et al. 2014; Schuller et al. 2015). However, while options exist for moorings (Sutton et al. 2014; Fassbender et al. 2015), inexpensive, small volume, low-power draw, fast response time, reagent-free, and pressure-tolerant sensors are not yet available for profiling float measurements of total seawater titration alkalinity (A_T), total dissolved inorganic carbon or (C_T), or partial pressure of CO_2 (pCO_2).

The need for additional carbonate system constraints led many (e.g., Millero et al. 1998; Lee et al. 2006; McNeil et al. 2007; Alin et al. 2012; Sasse et al. 2012; Bostock et al. 2013; Velo et al. 2013) to regress A_T data measured on hydrographic cruises against measurements of other seawater properties. The regression constants obtained allow A_T to be later estimated from other property measurements where A_T measurements are not also available. A_T is an ideal carbonate system parameter to estimate in this manner and use with pH for several reasons: it is nearly orthogonal to pH as a constraint for the carbonate system; it mixes linearly and is unaffected by temperature, gas exchange, or the continuing ocean uptake of anthropogenic carbon; and it varies predictably and linearly with other seawater properties. A_T has similar measurement uncertainties to C_T and pH (Bockmon and Dickson 2015).

A_T regression estimates have advanced since Millero et al. (1998) showed that A_T could be estimated across large

Additional Supporting Information may be found in the online version of this article.

*Correspondence: Brendan.carter@noaa.gov

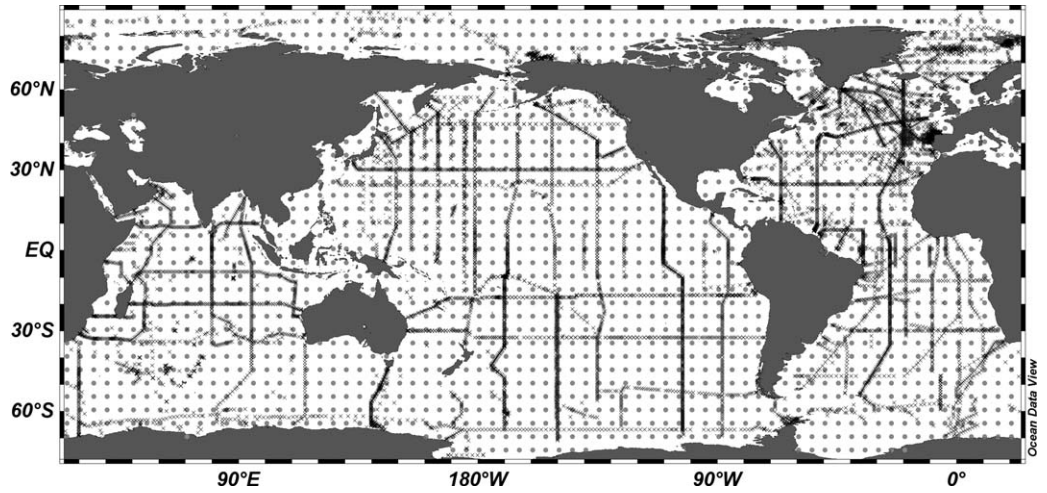

Fig. 1. A map of the locations of measurements included for our analysis (black x's) and of locations where we have estimated regression constants (grey dots).

regions of the surface ocean from simple regressions with high accuracy, and similar regressions have been used to simulate A_T distributions in Earth system models (e.g., Galbraith et al. 2015). Scientists have worked to develop regressions for new regions and to incorporate new A_T measurements into regression estimates. Sasse et al. (2012) and Velo et al. (2013) recently showed that superior fits can be obtained using neural networks or self-organizing maps that divide measurement sets into optimized "neurons" instead of regions, where a neuron can be thought of as a collection of measurements that share similarities in their meta, physical, or chemical data. This approach has the advantage of eliminating the need for arbitrarily prescribed regional boundaries, but the disadvantage of needing to optimize the arbitrary number of allowed neurons.

We argue boundaries between regions and between neurons limit the usefulness of the A_T estimates obtained from some of these methods and are unnecessary. One can imagine a float drifting from one region to another, or transitioning into a new neuron when measuring across a thermocline. In these cases, A_T estimates will show a discontinuity where the transition between one set of regression coefficients to another occurs. A simple example is the boundary between the Pacific and other sectors of the Southern Ocean where the estimates returned by equations from Millero et al. (1998) change by 9 μmol kg^{-1}. A neuron transition can also happen over time at a fixed location provided there is warming or freshening. These discontinuities could show up as abrupt and spurious changes in the C_T calculated from, for instance, measured pH and estimated A_T. We therefore argue that A_T estimate consistency is at least as important as A_T estimate accuracy. Lee et al. (2006) recognized this limitation and forced second order polynomials for sea surface temperature and salinity, applied to regimes of both physical and property space, to return identical estimates at regime boundaries. This approach has the drawback of biasing regression fits away from the values that return the smallest residuals.

We present methods for estimating A_T and associated uncertainty globally at all ocean depths. These methods are similar to the 3DwMLR method advocated by Velo et al. (2013), which circumvents the need to carve datasets into regions by considering a window around each estimate to be its own region. We take the 3DwMLR approach a step further by recognizing that linear coefficients can be linearly interpolated, or triangulated in three dimensions, onto any given location of interest. This ensures that transitions between regression coefficient sets happen smoothly from location to location. This aspect of our approach is very similar to the recently published methods for estimating carbonate mineral saturation in the North Pacific by Kim et al. (2015). We call our approach LIAR, for "locally interpolated A_T regression." The traditional meaning for our acronym serves as a reminder that the A_T values generated are only estimates, not measurements.

In the "Procedures" section, we detail the methods we use to generate regression coefficients. Then we detail how one can estimate A_T from LIAR coefficients. In the "Assessment" section, we estimate the uncertainty of LIAR A_T estimates and how estimate uncertainty varies with changing inputs and input uncertainties. In the "Discussion" section, we enumerate the advantages of LIAR strategy with respect to scope, convenience, consistency, and accuracy over similar estimation strategies.

Procedures

The merged data product

We merged the PACIFICA (Suzuki et al. 2013), GLODAPv1.1 (Key et al. 2004), and CARINA (Velo et al. 2009)

datasets, and then eliminated duplicates of bottle measurements that appear in more than one of these products. We eliminated any data flagged with a quality control code corresponding to "bad" or "questionable" for all of our regression parameters. We removed data collected before 1990—or the approximate advent of seawater reference materials which were later certified for A_T (Dickson et al. 2003)—for the data product we use to derive regression constants, although we retain this data in a separate data product and use that set for regression error estimate calculations in "Assessment" section. We omitted data from the PACIFICA dataset falling along the Ocean Station Papa line because some of the station profiles disagree with neighboring A_T profiles at depth (> 3000 m) by as much as 100 μmol kg^{-1}. We also removed measurements from GLODAPv1.1 stations 24048 through 24065 because the station profiles appeared noisy. Finally, we removed data from CARINA stations 11000 through 11013 because these Southern Ocean station profiles disagreed with neighboring profiles in GLODAPv1.1 (there were no neighboring profiles in CARINA). We are left with a merged data product with 204,110 sets of A_T measurements and other regression parameters. The locations of stations at which we have data are mapped as x's in Fig. 1.

Estimating regression coefficients

We estimate regression constants for each location on a three-dimensional (3D) grid, which is a subset of the World Ocean Atlas grid subsampled at 5° resolution. Specifically, we use all coordinates that have a longitude of [0.5°: 5°: 355.5°], a latitude of [−84.5°: 5°: 85.5°], and a depth z of [0, 10, 20, 30, 50, 75, 100, 125, 150, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1750, 2000, 2500, 3000, 3500, 4000, 4500, 5000, or 5500 m]. There are 44,957 combinations of these coordinates in total in our subset. The locations of coordinates are mapped as circles in Fig. 1.

We select a subset of the data product to use for each regression at each coordinate set. To prevent using data from seawater measured on opposite sides of Central America or the Bering Strait in a single regression, we exclude data in the Arctic and Atlantic Oceans when estimating regression constants outside of these oceans and exclude data from outside of these oceans when estimating regression constants within them. Exceptions are that data in the Southern Atlantic, south of 0° N, is never excluded for this reason and no data is excluded for this reason for regression constant estimates in the Southern Atlantic. The latitude-longitude polygons we use for these basins are provided as Supporting Information. We also exclude all data with a salinity less than 30 to ensure our open ocean coefficients are not overly biased by river water. Finally, we exclude all data not within latitude, longitude, depth, and potential density σ_θ windows of our measurement values. Window sizes W are given by criteria (1–4):

$$W_{\mathrm{Lat}} = 5^\circ \times i \tag{1}$$

$$W_{\mathrm{Lon}} = \frac{10^\circ \times i}{\cos(\mathrm{Latitude})} \tag{2}$$

$$W_{\mathrm{Depth}} = 50\ \mathrm{m} + \frac{z}{10} \times i \tag{3}$$

$$W_{\sigma\theta} = 0.05\ \frac{\mathrm{kg}}{\mathrm{m}^3} \times i \tag{4}$$

Data are included if they satisfy criteria (1), (2), and either (3) or (4). We iteratively increment the integer i whenever exclusion criteria result in fewer than 100 viable measurements. We divide by the cosine of the latitude in (2) to maintain an approximately constant window width at all latitudes.

Once we have selected a subset of our merged data product, we perform regressions with 16 combinations of regression parameters. Like Velo et al. (2013), we use robust linear regression, which iteratively re-estimates regression coefficients, following an initial traditional least squares estimate, by assigning smaller weights to measurements with larger residuals. The iterative outlier un-weighting step addresses inaccuracies in the assumption that measurement errors are adequately described by a normal distribution. We use a bisquare outlier test with the turning constant of 4.685 for this step, meaning data with residuals in excess of 4.685 the standard residual are given no weight. The first of the 16 regressions has all regression parameters we consider:

$$A_T = \alpha_0 + \alpha_S S + \alpha_\theta \theta + \alpha_{\mathrm{AOU}} \mathrm{AOU} + \alpha_N N + \alpha_{\mathrm{Si}} \mathrm{Si} \tag{5}$$

Here α terms are regression coefficients we estimate for the subscripted properties, S is salinity, θ is potential temperature in °C, N is nitrate concentration in μmol kg^{-1}, AOU is apparent oxygen utilization in μmol kg^{-1}, and Si is total dissolved silicate concentration in μmol kg^{-1} We use apparent oxygen utilization in place of O_2 concentration because preliminary testing found this to be a slightly more powerful predictor and one that is less correlated with temperature. We do not use phosphate since these are highly correlated with N—for which sensor measurements are more common—and including both measurements would therefore risk overfitting A_T. We henceforth call Eq. 5 "Regression 1." Predictors used in Regressions 1 through 16 are indicated in Table 1.

Regressions 9 through 16 do not include potential temperature because McNeil et al. (2007) found temperature terms can create large spurious surface seasonal A_T estimate swings in estimates calibrated using data with incomplete seasonal coverage. α_{Si}, α_{AOU}, and α_N are omitted from some regressions because the related measurements are frequently unavailable.

Estimating A_T

The LIAR method requires two steps to estimate A_T. The first step is interpolating the regression coefficients to the location of interest. Linear interpolation can be done easily once appropriate points are chosen to interpolate between, and choosing points that bound the location of interest is

Table 1. Constant terms used for each of the 16 regressions.

Reg. #	α_0	α_S	α_θ	α_{AOU}	α_N	α_{Si}
1	✓	✓	✓	✓	✓	✓
2	✓	✓	✓		✓	✓
3	✓	✓	✓	✓		✓
4	✓	✓	✓			✓
5	✓	✓	✓	✓	✓	
6	✓	✓	✓		✓	
7	✓	✓	✓	✓		
8	✓	✓	✓			
9	✓	✓		✓	✓	✓
10	✓	✓			✓	✓
11	✓	✓		✓		✓
12	✓	✓				✓
13	✓	✓		✓	✓	
14	✓	✓			✓	
15	✓	✓		✓		
16	✓	✓				

Table 2. R^2 fits for the 16 regressions against measured A_T for Variant 2 of the LIAR method. The "Ranking" ranks the 16 regressions in order of how well they reproduce measured A_T with Variant 2.

Reg. #	Parameters used	R^2 Variant 2	Ranking
1	*S*, θ, *N*, AOU, Si	0.973	1
2	*S*, θ, *N*, Si	0.972	3
3	*S*, θ, AOU, Si	0.971	5
4	*S*, θ, Si	0.970	7
5	*S*, θ, *N*, AOU	0.966	9
6	*S*, θ, *N*	0.964	11
7	*S*, θ, AOU	0.966	10
8	*S*, θ	0.955	15
9	*S*, *N*, AOU, Si	0.972	2
10	*S*, *N*, Si	0.971	4
11	*S*, AOU, Si	0.970	6
12	*S*, Si	0.969	8
13	*S*, *N*, AOU	0.961	12
14	*S*, *N*	0.959	13
15	*S*, AOU	0.957	14
16	*S*	0.940	16

made simple by our use of a regular grid. However, there are edge cases to consider when interpolating near holes in our grid (e.g., Greenland). For this reason, we use MATLAB Delaunay Triangulation 3D linear interpolation routines (see Lee and Schachter 1980) after dividing depth differences by a factor of 25 to equate 100 m depth with ~4° latitude in the triangulation distance calculation. Delaunay triangulation selects nearby points that bound the location of interest while avoiding sets of points that make "skinny" polygons with small minimum interior angles. Triangulation is faster and performs less smoothing than objective mapping. We note that additional smoothing is unnecessary because of the smoothing inherent in our regression constant estimation process. The changing window sizes with latitude and depth (see "Estimating regression coefficients" section) ensure our estimates are not strongly sensitive to the choice of 25 for the depth-to-latitude conversion or to our assuming latitude differences equate to longitude differences regardless of coordinate latitude in this step. The latter of these simplifications allows us to use a single interpolant for all estimate coordinates for each regression coefficient, and greatly reduces the LIAR estimation routine computational burden. Extrapolated values outside the domain over which we have regression constant estimates (e.g., near sediments and some coasts) are set equal to the regression constants interpolated at the nearest location inside the domain. As when selecting data for a regression, we interpolate the Atlantic and Arctic Oceans separately to avoid interpolating across Central America or the Bering Strait. In the second step, the interpolated regression coefficients are used to estimate A_T directly from the intended equation (e.g., Eq. 5) in the second step.

Estimate requirements

The LIAR code is written for MATLAB R2014b, with backward compatibility tested through version 2012. Computation time varies with application and machine, although our desktop 64 bit PC operating with a 2.0 GHz processor returns an average of 35,000 A_T estimates per second.

LIAR estimates require any combination of the following measurements, with *S* being the only mandatory input: *S*, Si, *N*, O_2 or AOU, θ or in situ temperature. Concentrations can be provided in molar or molal units. Temperature and O_2 are converted to θ and AOU, respectively. These conversions are made using the CSIRO MATLAB seawater package version 3.1 (Morgan and Pender 2006). The (now deprecated) seawater package is used in place of updated Thermodynamic Equation of Seawater 2010 (TEOS-10) functions to ensure converted measurements are consistent with the GLODAPv1.1, CARINA, and PACIFICA data products used to estimate regression coefficients. Input uncertainties are an optional input; default values, corresponding to the uncertainties in Table 3 scaled to typical deep water properties, are assumed if none are provided.

Table 3. Assumed input measurement uncertainties

Parameter	*U*
S	0.005
θ	0.005°C
N	2% meas.
AOU	1% meas. O_2
Si	2% meas.

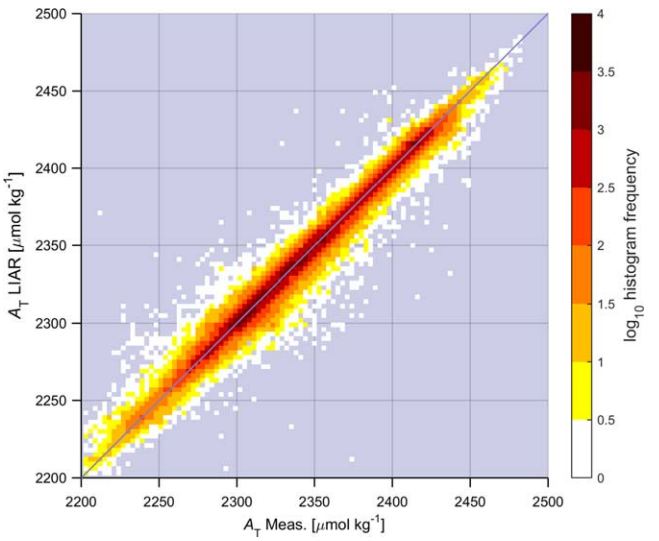

Fig. 2. 2D histogram with the number of measurements falling within small square bins of LIAR-estimated A_T (y axis) and measured A_T (x axis) shown as color on a log scale for Regression 1 (see Supporting Information for histograms for other regressions). More than 90% of measurements fall within the darker bins corresponding to log histogram frequencies > 2. Thin blue 1 : 1 foreground lines and a background grid are provided for reference. No measurements fall within bins where the blue-grey background is visible.

Assessment

We use two variants on LIAR to assess aspects of the estimation strategy. For both variants, we estimate regression coefficients at the locations of each of our 204,110 A_T measurements for which we have measurements of all regression parameters instead of at the 44,957 WOA coordinates, allowing us to bypass the interpolation step when estimating A_T at the measured locations. Variant 2 also does not use measurements from a cruise as training data for regression coefficients estimated along that cruise track. For assessments with the "unmodified LIAR method," we interpolate regression coefficients from the 44,957 WOA coordinates to the 204,110 measurement coordinates.

Figure 2 has two-dimensional (2D) histograms of unmodified LIAR-estimated A_T against measured A_T for measurements in our merged data product for regression 1 (see Supporting Information for histograms for all 16 regressions). The strong linear relationships demonstrate that the LIAR method estimates the global A_T field well. R^2 statistics are given for the 16 regressions for LIAR Variant 2 estimates in Table 2 for all data in our merged data product. Variant 2 estimates, like estimates that will be made when the LIAR method is applied, do not benefit from using the measured alkalinity at the location of interest to estimate the regression constants used for that location. Variant 2 estimates are therefore a more appropriate test for the strength of the fit than unmodified LIAR estimates. There is also no need to adjust Variant 2 R^2 values in light of the differing degrees of freedom since a larger number of predictors does not guarantee a better fit to data withheld from the regression constant estimation procedure. We caution that estimate uncertainty can increase with an increasing number of predictors when collinear regression parameters are used, but note that our uncertainty estimation procedure (detailed later) accounts for this by propagating uncertainty with the regression coefficients. Variant 2 R^2 values suggest the relative importance of non-salinity regression parameters is $\theta < \mathrm{AOU} < N < \mathrm{Si}$. All four parameters improve the fit.

We rank the 16 regressions by their Variant 2 R^2 values in Table 2, but note that the best regression to use depends both on which regression parameters are available and the uncertainties of the input parameter measurements. We therefore develop an approach to estimating LIAR estimate uncertainty from the bottom up. This approach can be applied to measurements of arbitrary quality, so we are able to return estimate uncertainties from user-provided input uncertainties as part of the LIAR software.

We consider four sources of error E for our bottom-up error uncertainty estimate:

1. E_{Alk} from errors in the A_T data used to fit the regression constants,
2. E_{Input} from input parameter measurement uncertainties,
3. E_{MLR} from the inadequacies inherent to the use of multiple linear regression to reproduce the global A_T distribution,
4. and errors associated with interpolating regression coefficients E_{Interp}.

We combine these errors as the square root of the sum of squares to produce the overall uncertainty, E_{LIAR}:

$$E_{\mathrm{LIAR}} = \sqrt{E_{\mathrm{Alk}}{}^2 + E_{\mathrm{MLR}}{}^2 + E_{\mathrm{Interp}}{}^2 + E_{\mathrm{Input}}{}^2} \quad (6)$$

We start with an E_{Alk} estimate of 3.3 μmol kg^{-1} for the uncertainty of the A_T measurements in our merged data product (Velo et al. 2009). This is the minimum possible uncertainty for LIAR estimates. For reasons discussed shortly, over or underestimation of this uncertainty is not of great concern for this assessment.

Next we estimate E_{Input}. We assume measurement uncertainties (other than A_T uncertainties) have negligible influence on the regression constant α values due to the large number of measurements used to estimate each constant. E_{Input} is due rather to uncertainties in the singular sets of parameter measurements used to estimate A_T from the 16 regressions. We estimate E_{Input} as the input uncertainty propagated through the regression equations (e.g., Eq. 5). For a regression with n predictors, E_{Input} is:

Table 4. Error estimates for the subset of our data product found within the open-ocean salinity range of 33–38. E_{MLR} is error arising from the use of a MLR approach, E_{Input} is error arising from uncertainties in our input data, and E_{LIAR} is the overall estimate uncertainty. Errors are expressed as errors in μmol A_T kg^{-1}.

Reg. #	Parameters used	E_{MLR}	E_{Input}	E_{LIAR}
1	*S*, θ, *N*, AOU, Si	2.8	2.1	5.8
2	*S*, θ, *N*, Si	3.0	2.1	5.9
3	*S*, θ, AOU, Si	3.3	1.7	6.0
4	*S*, θ, Si	3.5	1.7	6.1
5	*S*, θ, *N*, AOU	3.6	2.3	6.3
6	*S*, θ, *N*	3.5	3.2	6.7
7	*S*, θ, AOU	4.2	1.6	6.5
8	*S*, θ	6.5	1.9	8.2
9	*S*, *N*, AOU, Si	3.1	2.1	5.9
10	*S*, *N*, Si	3.2	2.2	6.1
11	*S*, AOU, Si	3.6	1.6	6.1
12	*S*, Si	4.0	1.8	6.4
13	*S*, *N*, AOU	4.9	3.0	7.4
14	*S*, *N*	5.2	3.2	7.7
15	*S*, AOU	6.3	1.7	8.0
16	*S*	9.1	1.8	10.4

$$E_{\mathrm{Input}} = \sqrt{\sum_{j=1}^{n} (U_j \alpha_j)^2} \tag{7}$$

Here, U_j is assumed uncertainty for the *j*th parameter used. Here we assume our measurement uncertainties are independent despite potential small correlations between errors in, for instance, temperature and AOU calculated from temperature. Our input parameter uncertainty estimates for our merged data product are given in Table 3. These estimates are inferred from Suzuki et al. (2013)'s minimum adjustments for the PACIFICA data product.

We estimate E_{Interp} by comparing the root mean squared error (henceforth: error) for Variant 1 estimates (or E_1) to the error for estimates from the unmodified LIAR method (or E_0). Variant 1 has no interpolation step, so E_{Interp} is 0. The methods are the same otherwise. This allows us to write:

$$E_{\mathrm{Interp}} = \sqrt{(E_0^2 - E_1^2)} \tag{8}$$

E_{Interp} is statistically indistinguishable from 0 over the open ocean salinity range of 33–38, and small relative to E_{MLR}(<10%) outside this range. We henceforth assume E_{Interp} is 0.

The overall LIAR uncertainty estimate is the error for Variant 2 A_T estimates, or E_{LIAR}. We use error estimates from Variant 2 in place of similar estimates from the unmodified LIAR method or Variant 1 because the Variant 2 estimates are not derived from regression coefficients determined using the target A_T values, as will be the case for future LIAR estimates.

We rearrange Eq. 6 and neglect E_{Interp} to solve for E_{MLR}:

$$E_{\mathrm{MLR}} = \sqrt{E_{\mathrm{LIAR}}{}^2 - E_{\mathrm{Alk}}{}^2 - E_{\mathrm{Input}}{}^2} \tag{9}$$

This E_{MLR} estimate is highly sensitive to the assumed E_{Alk} estimate, although this is unimportant for our final uncertainty estimates because E_{MLR} and E_{Alk} will always be combined in Eq. 6; an underestimation of one is compensated for by an overestimation of the other.

In Table 4, we report mean E_{MLR}, E_{Input}, and E_{LIAR} for each of the 16 regressions for the portion (>95%) of our data product found within the open-ocean salinity range of 33–38.

Critically, LIAR estimates have similar or superior accuracy to alternative estimates. Velo et al. (2013) suggested their neural network and 3DwMLR methods have average absolute residuals (note: not standard deviations) of < 5 μmol kg^{-1}. Regression 1 is our most comparable regression to Velo et al. (2013)'s method—they use O_2 in place of AOU and include pressure and phosphate as additional regression parameters—for which we have an average absolute residual of 4.1 μmol kg^{-1}. McNeil et al. (2007) reported an error of 8.1 μmol kg^{-1} for the depth range, regression, and region over which we estimate an error of ~5.8 μmol kg^{-1}. Bostock et al. (2013) obtain an error of 9.8 μmol kg^{-1} for the region south of 20° S while the LIAR regression with the same parameters (regression 7) returns an error of 6.4 μmol kg^{-1} for this region. Lee et al. (2006) used six constant terms in five 2nd degree regional regressions to estimate surface (<30 m) alkalinity with a combined error of 8.1 μmol kg^{-1}, while we achieve smaller errors for this depth range with all regressions except 8 and 16 (i.e., with only θ and *S* and with just *S*, respectively). The self-organizing map approach of Sasse et al. (2013) achieved an error of 9.2 μmol kg^{-1} for a similar region to that considered by Lee et al. (2006). Alin et al. (2012) used a four-term function of temperature and salinity to estimate A_T with an error of 6.4 μmol kg^{-1} above 500 m depth in the CALCOFI region, and the LIAR error in this region is 3.9–6.2 μmol kg^{-1} (depending on regression). We apply Millero et al. (1998)'s estimate for the Pacific Gyres to data in our data product shallower than 50 m depth, between 20°S to 30°N, and between 150° and 240°E and estimate an error of 7.5 μmol kg^{-1}. LIAR error for this subset of our data product using the equivalent Regression 16 is 6.4 μmol kg^{-1}. We do not exclude data beyond the open ocean salinity range or measured before 1990 for these error comparisons.

As with other estimation strategies (e.g., Velo et al. 2013) LIAR estimation performs substantially worse in the ~4% of our dataset that does not fall within the open ocean salinity range. For example, the RMSE values in Table 4 increase by an average of 3% when we extend the minimum salinity measurement used for the calculation to 32, and increase by

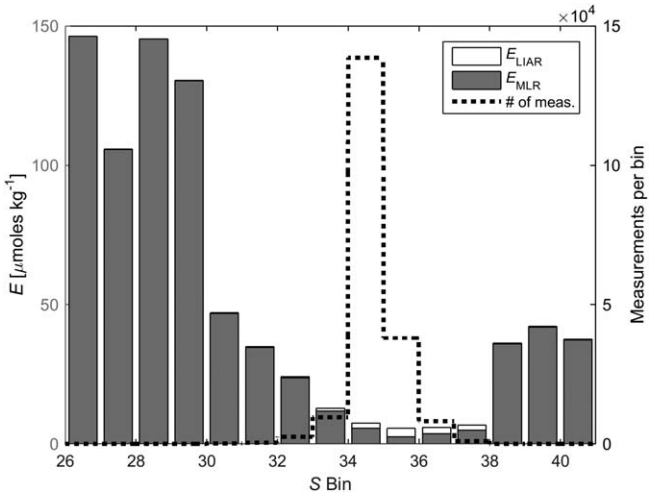

Fig. 3. E_{MLR} (grey bars) and E_{LIAR} (larger or similarly sized white bars behind grey bars) estimates for Regression 1 (left *y*-axis), and the number of measurements (dashed line, right *y*-axis) for each 1 unit salinity bin (*x*-axis).

an average of 36% when all measurements are included. Salinities between 32 and 33 are common in the surface North Pacific, and LIAR only performs slightly worse for these estimates (32% higher RMSE values for this surface subset). Very large errors are typically found in regions with unusual A_T to S relationships resulting from river water (e.g., the Arctic or Bay of Bengal) or in evaporative marginal seas where there are not enough measurements to estimate LIAR coefficients locally (e.g., the Red and Mediterranean Seas) (Carter et al. 2014). While the LIAR method could, at higher resolution, be adapted to reflect the distinct A_T to S relationships characteristic of these regions, we decided to instead optimize our method for the open ocean with a coarse resolution grid, the requirement of >100 measurements per regression, and the omission of data with $S<30$ from the data used to estimate regression constants. Nevertheless, we develop methods to estimate the greater uncertainties for LIAR estimates for unusually fresh or saline seawater. We do this by calculating E_{MLR} separately for all data product measurements falling within each 1 unit S bin that our dataset spans. For bins for which we have no measurements, we linearly interpolate between estimates for neighboring bins. Our final error estimate then uses the E_{MLR} estimate appropriate to the salinity bin the measurement is found within. Figure 3 shows E_{MLR} for Regression 1 as well as the number of data product measurements within each bin. Other regressions have similar distributions.

Our software therefore returns A_T estimate uncertainty E_{Est}:

$$E_{Est}=\sqrt{E_{Alk}{}^2+E_{MLR}{}^2+\sum_{j=1}^{n}(U_j\alpha_j)^2} \quad (10)$$

where E_{Alk} is the constant 3.3 μmol kg^{-1}, E_{MLR} is determined from the histogram in Fig. 3 (or an equivalent histogram appropriate to the regression used), and U values are provided by the user (or assumed to equal the values in Table 3 if not provided).

LIAR estimate bias is indistinguishable from 0 for all 16 regressions. However, LIAR errors are not normally distributed about 0 due to large errors for a small number of measurements. For all 16 regressions, more than 94% of LIAR errors are less than the standard error estimates in Table 4, whereas for normally distributed errors we would expect ~68% of deviations to be less than or equal to the standard error. However, ~87% of errors are less than E_{Est}. This suggests Eq. 10 does scale estimate uncertainty with estimate error for our data product to a degree, but not sufficiently to ensure that the ratio of the deviations to E_{Est} is normally distributed. We do not anticipate this will be a problem for applications in the open ocean, but note that LIAR uncertainties are likely underestimated for river plumes, marginal seas, and areas without many historical A_T measurements.

We tested the LIAR method on the recent occupation of the P16 repeat hydrographic line. Data from this occupation were collected over two cruises with three total legs from early 2014 through mid-2015 (Talley 2014; Cross 2015; Macdonald 2015). These data were not used for estimating the LIAR regression constants, so they provide a preliminary demonstration of how well the method could perform in regions where there are ample measurements available in the training dataset. Figure 4 maps (preliminary) measured A_T, LIAR-estimated A_T (from Regression 1), and differences between these values. It can be seen that LIAR does an excellent job of capturing the broad-scale patterns observed on this cruise, including several localized features of the fronts in the Antarctic Circumpolar Current and the North Pacific. The measurement-estimate disagreement averaged -0.1 ± 3.2 μmol kg^{-1} A_T for data from this cruise.

Discussion

LIAR A_T estimates improve on the previously available suite of estimation strategies in several important ways without compromising the high estimate accuracy characteristic of recent A_T estimation efforts. First, LIAR estimates are applicable globally at all ocean depths. Second, lacking regional or neural boundaries, LIAR provides estimate precision when transitioning between regions of physical or property space. Third, LIAR can be used with many combinations of parameter measurements, including combinations that are measureable by float-capable biogeochemical sensors. Fourth, LIAR provides uncertainty estimates that scale with input uncertainties and seawater S. Fifth, LIAR regression coefficients are determined using more data and more recent data than some alternatives. Finally, we provide MATLAB code and documentation that make LIAR estimates accessible. It is our hope LIAR estimates will be used to supplement incomplete constraints

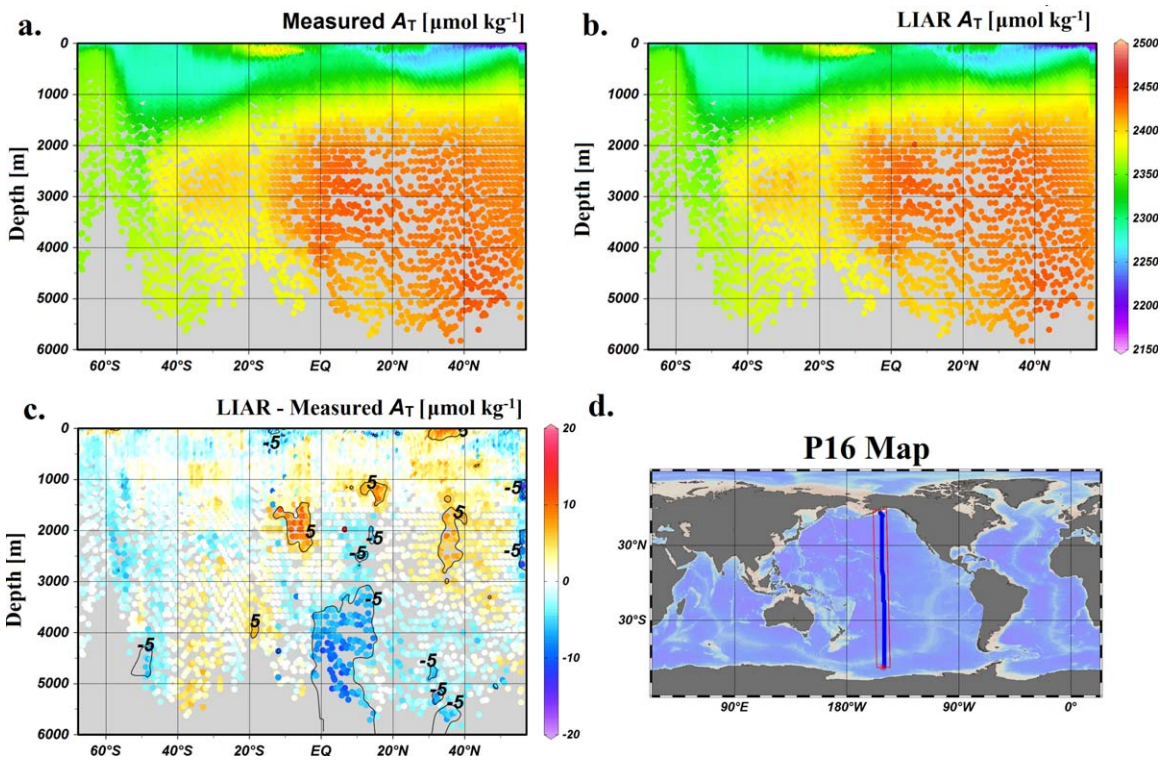

Fig. 4. (**a**) Measured and (**b**) LIAR-estimated (Regression 1) A_T mapped against latitude and depth using the same colorscale, and (**c**) differences between these values. Data are from the 2014 to 2015 occupation of the P16 hydrographic section (mapped in **d**). Contours in (**c**) demark regions where the average offset between measured and estimated A_T exceeds ± 5 μmol kg^{-1} for a version of the same plot smoothed with weighted average gridding (with 8 and 9 permille length scales in the X and Y directions, respectively).

of the seawater carbonate system from, for example, sensor measurements and Earth system models that do not include prognostic A_T due to computational or data storage constraints (e.g., Galbraith et al. 2015).

The high, and in some cases improved, accuracy of LIAR estimates relative to other A_T estimates is likely due to the larger quantities of data used to produce the regression coefficients, the large fraction of data collected in the years following the introduction of reference materials for alkalinity A_T in our data product, and to the circumvention of the limitation of alternate approaches (except 3DwMLR) that each measurement in the training data set be used to constrain only one set of regional regression constants. Also, LIAR implicitly relies on sample position information through the regression constant interpolation step. This is the reason LIAR achieves comparatively small errors even for regressions with few parameters (e.g., 10.4 μmol kg^{-1} globally using only S as a predictor) and is able to automatically adopt regression coefficients appropriate for both dynamic frontal regions and stable subtropical gyres. The local-interpolation step is added for the convenience of deriving regression coefficients appropriate to arbitrary locations in the ocean and has little impact on the estimate accuracy.

An important question for our estimation strategy is how well the methods reproduce temporal A_T changes from natural variability and long term changes. LIAR does capture natural variability to an extent. For example, the standard deviation of surface (<25 m) A_T is 13 μmol kg^{-1} and 11 μmol kg^{-1} at ocean stations ALOHA and BATS, respectively (Joyce and Robbins 1996; Karl and Lukas 1996), while the standard deviation between measured and LIAR-estimated A_T is only 6 μmol kg^{-1} and 6 μmol kg^{-1}, respectively. Lacking a temporal component, LIAR cannot capture the long term changes expected with biogeochemical feedbacks with ocean acidification. However, these impacts have been estimated to only become detectable after ~2040 (Ilyina et al. 2009), so they are not a large concern for the immediate future. Furthermore, LIAR estimates may be of use as a baseline for detecting such changes. For instance, regression of surface A_T at BATS normalized to a salinity of 35 against time reveals a statistically significant (at 95% conf.) increase of ~0.24 μmol kg^{-1} per year over the record, while no increase is found in the difference between measurements and LIAR estimates. These observations suggest this observed surface increase can be attributed to captured natural variability rather than to long term changes.

Comments and recommendations

The next step for development of the LIAR method is to update regression coefficients with the planned version 2 of the GLODAP data product. This data product will have quality controlled data from over 700 cruises including data from more recent cruises than those included in GLODAPv1.1, CARINA, and PACIFICA. This data product will also likely have cruises from several under-represented regions in our merged data product, such as the Gulf of Mexico, the Mediterranean, and the South China Sea.

It would be desirable to extend LIAR to other programming platforms commonly used by oceanographers, especially freely-distributed platforms such as Python, Fortran, Ocean Data View, and R. Implementations in Fortran, a common language for Earth system models and ocean circulation models, would allow the LIAR method to more easily be used to simulate A_T distributions in models that do not resolve A_T prognostically.

It may be useful to estimate and interpolate E_{MLR} regionally instead of against S. This would allow uncertainty estimates to increase where residuals are larger due to enhanced measurement uncertainty or variability that is not well captured by our regression approach. We expect such a strategy would further reduce the non-normality of our uncertainty-estimate-normalized error distribution.

Methodological adaptations near river mouths and in marginal seas may also allow the LIAR method to return better estimates for these regions. For instance, deriving regression constant sets specific to riverine or estuarine outflows and placing these regression constant sets in the LIAR regression constant grid at the locations of river mouths may allow for better estimates to be returned in these areas. Currently, LIAR uncertainties in these regions are quite large, and possibly underestimated.

References

Alin, S. R., R. A. Feely, A. G. Dickson, J. M. Hernández-Ayón, L. W. Juranek, M. D. Ohman, and R. Goericke. 2012. Robust empirical relationships for estimating the carbonate system in the southern California Current System and application to CalCOFI hydrographic cruise data (2005–2011). J. Geophys. Res. **117**: 2156–2202. doi:10.1029/2011JC007511

Bockmon, E. E., and A. G. Dickson. 2015. An inter-laboratory comparison assessing the quality of seawater carbon dioxide measurements. Mar. Chem. **171**: 36–43. doi:10.1016/j.marchem.2015.02.002

Bostock, H. C., S. E. Mikaloff Fletcher, and M. J. M. Williams. 2013. Estimating carbonate parameters from hydrographic data for the intermediate and deep waters of the Southern Hemisphere oceans. Biogeosciences **10**: 6199–6213. doi:10.5194/bg-10-6199-2013

Bresnahan, P. J., T. R. Martz, Y. Takeshita, K. S. Johnson, and M. LaShomb. 2014. Best practices for autonomous measurement of seawater pH with the Honeywell Durafet. Methods Oceanogr. **9**: 44–60. doi:10.1016/j.mio.2014.08.003

Carter, B. R., J. R. Toggweiler, R. M. Key, and J. L. Sarmiento. 2014. Processes determining the marine alkalinity and calcium carbonate saturation state distributions. Biogeosciences **11**: 7349. doi:10.5194/bg-11-7349-2014

Cross, J. 2015. Preliminary Cruise Report P16N Leg 1. Available from http://cchdo.ucsd.edu/cruise/33RO20150410

Dickson, A. G., J. D. Afghan, and G. C. Anderson. 2003. Reference materials for oceanic CO_2 analysis: A method for the certification of total alkalinity. Mar. Chem. **80**: 185–197. doi:10.1016/S0304-4203(02)00133-0

Fassbender, A. J., C. L. Sabine, N. Lawrence-Slavas, E. H. De Carlo, C. Meinig, and S. Maenner Jones. 2015. Robust sensor for extended autonomous measurements of surface ocean dissolved inorganic carbon. Environ. Sci. Technol. **49**: 3628–3635. doi:10.1021/es5047183

Galbraith, E. D., and others. 2015. Complex functionality with minimal computation: Promise and pitfalls of reduced-tracer ocean biogeochemistry models. J. Adv. Model. Earth Syst. **7**: 2012–2028. doi:10.1002/2015MS000463

Ilyina, T., R. E. Zeebe, E. Maier-Reimer, and C. Heinze. 2009. Early detection of ocean acidification effects on marine calcification. Global Biogeochem. Cycles **23**: GB1008. doi:10.1029/2008GB003278

Johnson, K. S., S. C. Riser, and D. M. Karl. 2010. Nitrate supply from deep to near-surface waters of the North Pacific subtropical gyre. Nature **465**: 1062–1065. doi:10.1038/nature09170

Johnson, K. S., Y. Gu, T. R. Martz, and S. C. Riser. 2012. Development of an integrated ISFET pH sensor for high pressure applications in the deep-sea. Monterey Bay Aquarium Research Institute, Moss Landing, CA.

Joyce, T. M., and P. Robbins. 1996. The long-term hydrographic record at Bermuda. J. Clim. **9**: 3121–3131. doi:10.1175/1520-0442(1996)009<3121:TLTHRA>2.0.CO;2

Karl, D. M., and R. Lukas. 1996. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. Deep-Sea Res. II **43**: 129–156. doi:10.1016/0967-0645(96)00005-7

Key, R. M., and others. 2004. A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). Global Biogeochem. Cycles **18**: GB4031. doi:10.1029/2004GB002247

Kim, T.-W., G.-H. Park, D. Kim, K. Lee, R. A. Feely, and F. J. Millero. 2015. Seasonal variations in the aragonite saturation state in the upper open-ocean waters of the North Pacific Ocean. Geophys. Res. Lett. **42**: 4498–4506. doi:10.1002/2015GL063602

Lee, D. T., and B. J. Schachter. 1980. Two algorithms for constructing a Delaunay triangulation. Int. J. Comput. Inform. Sci. **9**: 219–242. doi:10.1007/BF00977785

Lee, K., and others. 2006. Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans. Geophys. Res. Lett. **19**: L19605. doi:10.1029/2006GL027207

Macdonald, A. 2015. Cruise report: P16N. Available from http://cchdo.ucsd.edu/cruise/33RO20150525

McNeil, B. I., N. Metzl, R. M. Key, R. J. Matear, and A. Corbiere. 2007. An empirical estimate of the Southern Ocean air-sea CO_2 flux. Global Biogeochem. Cycles **21**: GB3011. doi:10.1029/2007GB002991

Millero, F. J., K. Lee, and M. Roche. 1998. Distribution of alkalinity in the surface waters of the major oceans. Mar. Chem. **60**: 111–130. doi:10.1016/S0304-4203(97)00084-4

Morgan, P., and L. Pender. 2006. CSIRO Marine Research MATLAB Seawater Software Library.

Sasse, T. P., B. I. McNeil, and G. Abramowitz. 2013. A new constraint on global air-sea CO_2 fluxes using bottle carbon data. Geophys. Res. Lett. **40**: 1594–1599. doi:10.1002/grl.50342

Schuller, D., and others. 2015. SOCCOM float deployments from Polarstern ANTXXX_2 PS89.

Sutton, A. J., and others. 2014. A high-frequency atmospheric and seawater pCO_2 data set from 14 open-ocean sites using a moored autonomous system. Earth Syst. Sci. Data **6**: 353–366. doi:10.5194/essd-6-353-2014

Suzuki, T., and others. 2013. PACIFICA Data Synthesis Project. ORNL/CDIAC-159, NDP-092. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee. doi:10.3334/CDIAC/OTG.PACIFICA_NDP092

Talley, L. D. 2014. Cruise report: P16S. doi:10.7942/320620140320

Velo, A., F. F. Perez, P. Brown, T. Tanhua, U. Schuster, and R. M. Key. 2009. CARINA alkalinity data in the Atlantic Ocean. Earth Syst. Sci. Data **1**: 45–61. doi:10.5194/essd-1-45-2009

Velo, A., F. F. Pérez, T. Tanhua, M. Gilcoto, A. F. Ríos, and R. M. Key. 2013. Total alkalinity estimation using MLR and neural network techniques. J. Mar. Syst. **111**: 11–18. doi:10.1016/j.jmarsys.2012.09.002

Acknowledgments

We would like to thank Dr. Robert Key for helpful comments early in the process and Dr. Kelly Kearney for help with code optimization. Datasets were acquired from the Carbon Dioxide Information and Analysis Center (CDIAC) webpage, from the HOT-DOGS data portal (http://hahana.soest.hawaii.edu/hot/hot-dogs/), and from the BATS data portal (http://bats.bios.edu/). We are grateful to Kathy Tedesco of the Climate Observation Division of the NOAA Climate Program Office for funding for this research. Funding was provided by the Global Ocean Ship-based Hydrographic Investigations Program (GO-SHIP) (N8R1SE3-PRF) and the US Global Carbon Data Management and Synthesis Project (N8R3CEA-PDM). A. R. Gray was supported by a National Oceanic and Atmospheric Administration (NOAA) Climate and Global Change Postdoctoral Fellowship. This is PMEL contribution number 4390 and JISAO contribution number 2503.

Submitted 15 October 2015
Revised 18 December 2015
Accepted 26 December 2015

Associate editor: Mike DeGrandpre