LIMNOLOGY and OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 19, 2021, 785–809

doi: 10.1002/lom3.10461

New and updated global empirical seawater property estimation routines

Brendan R. Carter [1,2]* Henry C. Bittig [3] Andrea J. Fassbender,[2] Jonathan D. Sharp,[1,2] Yuichiro Takeshita [4] Yuan-Yuan Xu,[5,6] Marta Álvarez,[7] Rik Wanninkhof,[6] Richard A. Feely,[2] Leticia Barbero [5,6]

[1]Cooperative Institute for Climate, Ocean, and Ecosystem Studies, University of Washington, Seattle, Washington
[2]Pacific Marine Environmental Laboratory, Seattle, Washington
[3]Department of Marine Chemistry, Leibniz Institute for Baltic Sea Research Warnemünde, Rostock, Warnemünde, Germany
[4]Monterey Bay Aquarium Research Institute, Moss Landing, California
[5]Cooperative Institute for Marine and Atmospheric Studies, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida
[6]Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida
[7]Instituto Español de Oceanografía, CSIC, A Coruña, Spain

Abstract

We introduce three new Empirical Seawater Property Estimation Routines (ESPERs) capable of predicting seawater phosphate, nitrate, silicate, oxygen, total titration seawater alkalinity, total hydrogen scale pH (pH_T), and total dissolved inorganic carbon (DIC) from up to 16 combinations of seawater property measurements. The routines generate estimates from neural networks (ESPER_NN), locally interpolated regressions (ESPER_LIR), or both (ESPER_Mixed). They require a salinity value and coordinate information, and benefit from additional seawater measurements if available. These routines are intended for seawater property measurement quality control and quality assessment, generating estimates for calculations that require approximate values, original science, and producing biogeochemical property context from a data set. Relative to earlier LIR routines, the updates expand their functionality, including new estimated properties and combinations of predictors, a larger training data product including new cruises from the 2020 Global Data Analysis Project data product release, and the implementation of a first-principles approach for quantifying the impacts of anthropogenic carbon on DIC and pH_T. We show that the new routines perform at least as well as existing routines, and, in some cases, outperform existing approaches, even when limited to the same training data. Given that additional training data has been incorporated into these updated routines, these updates should be considered an improvement over earlier versions. The routines are intended for all ocean depths for the interval from 1980 to ~2030 c.e., and we caution against using the routines to directly quantify surface ocean seasonality or make more distant predictions of DIC or pH_T.

Anthropogenic impacts on the environment are changing the physical and chemical state of the ocean. The accumulation of excess ocean heat (Roemmich et al. 2012; Purkey and Johnson 2013) and carbon (Sabine et al. 2004; Khatiwala et al. 2013; Carter et al. 2017, 2019*a*; Gruber et al. 2019) and the redistribution of freshwater between regions of the ocean (Durack et al. 2012) and geological reservoirs are modifying ocean circulation pathways and causing sea level rise (Nerem et al. 2018), ocean acidification (OA; Feely et al. 2004, 2009; Doney et al. 2009; Jiang et al. 2019), and ocean deoxygenation (Sasano et al. 2018). These changes are fundamentally shifting the physical and chemical environments of marine organisms and threatening ocean ecosystems and services (Gattuso et al. 2015; Doney et al. 2020).

Global climate change poses a challenge for ocean monitoring, necessitating sustained high-quality measurements across timescales and across the vast and remote global ocean. A variety of approaches and platforms have been developed for ocean monitoring (e.g., autonomous surface vehicles, profiling floats,

*Correspondence: brendan.carter@gmail.com

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

and fixed moorings), each of which has a niche for examining a range of temporal and spatial scales (Bushinsky et al. 2019) and each of which has strengths and weaknesses for addressing aspects of global change (Carter et al. 2019*b*). The cost and difficulty of measurements is a limiting factor for all approaches, so it is impossible as of today to have extensive high-quality and high-frequency measurements everywhere they are desired. Given this limitation, an emerging approach involves using algorithms that have been trained to reproduce measurements of seawater properties from co-located measurements of other seawater properties. These algorithms take advantage of strong regional correlations between seawater properties that result from oceanographic processes that shape the distributions of many different seawater properties in similar ways (e.g., organic matter cycling with nearly constant stoichiometric ratios between macronutrients, and freshwater cycling that linearly dilutes or concentrates most chemical concentrations in seawater). Once trained, the algorithms can be used to predict the desired properties from other properties that are more routinely measured either remotely by satellite or using available in situ sensors. This strategy has seen use for more than two decades (e.g., Goyet et al. 2000; Lee et al. 2006), though recent advances in skill, flexibility, and diversity of the algorithms available (Carter et al. 2016, 2018; Sauzède et al. 2017; Bittig et al. 2018; Landschützer et al. 2019; Gregor and Gruber 2021) have made it possible to create climatologies (Broullón et al. 2019, 2020; Jiang et al. 2019), calibrate and monitor drift-adjustments for sensors on autonomous sensor platforms (Johnson et al. 2017; Takeshita et al. 2018), create novel global data products (Carter et al. 2021), and fill holes in data sets when the final analysis is not strongly sensitive to estimate errors, for example, when silicate and phosphate are estimated for use in seawater carbonate chemistry calculations (e.g., van Hueven et al. 2011) or when total alkalinity (TA) is needed to convert pH_T between temperatures (Jiang et al. 2019; Carter et al. 2019*a*).

The growing number of use cases for seawater property estimation algorithms means it is important to refine the algorithms to the extent possible, especially given that some observing approaches depend on these algorithms for sensor calibration and validation. As a notable example, biogeochemical Argo floats calibrate pH_T and nitrate sensors using algorithm estimates in the comparatively stable mid-depths of the ocean (Johnson et al. 2017), and additionally rely on estimated seawater alkalinity at all depths to calculate dissolved inorganic carbon (DIC) and the partial pressure of CO_2 (pCO_2) (Gray et al. 2018; Williams et al. 2018).

Increasing ocean DIC content from anthropogenic carbon (C_{ant}) storage and decreasing pH_T values from OA provide an ongoing challenge to the accuracy of these algorithms: the algorithms are trained, or fit, to data collected over the last three decades, but will be used primarily to estimate seawater properties specific to recent years and the coming years until improved algorithms become available. How then should we deal with the changes from, for example, OA? Three notable existing algorithms for pH_T have simplistic and empirical treatments of the effects of OA. One has no parameterization for OA, but instead provides a suggested time-span for the algorithm (Williams et al. 2016); another uses a simple density interpolation of empirically derived global changes that, for example, does not distinguish the rapidly changing intermediate North Atlantic from the comparatively static intermediate subpolar North Pacific (Carter et al. 2018); and the one last uses a regional empirical approach that risks misattributing long-term change and natural variability in pH_T (Bittig et al. 2018). Broullón et al. (2020) also use an empirical relationship to capture the effects of OA for their DIC algorithm. These algorithms are expected to become increasingly biased under future OA conditions.

In this paper, we improve upon existing algorithms with new methods and new observational data products and encode them into a package of software routines in the MATLAB language. We also introduce a new neural-network approach that can return estimates from more diverse combinations of predictors than previous efforts. We also improve how the algorithms handle C_{ant} impacts on DIC and pH_T, and the new approach should allow future projections of these properties to be useful over longer time horizons while avoiding bias from empirical fits to interannual variability.

Methods

Basics, updates, new methods, and new features

The first of two products in this effort is an improvement upon the Locally Interpolated Regression (LIR) strategy for global and full-water column seawater alkalinity estimation that was implemented by Carter et al. (2016) and is similar to a method described by Velo et al. (2013). This approach was later updated and extended to estimating seawater pH_T and nitrate (Carter et al. 2018: LIRv2) and was most recently expanded to oxygen, phosphate, and silicate estimates (Carter et al. 2021). The new improvements in LIR-based empirical seawater property estimation routines (called here: ESPER_LIR, equivalent to LIRv3), relative to LIRv2, include:

1. Use of the 2020 release of the GLObal Data Analysis Project data product (GLODAPv2.2020: Olsen et al. 2020), for predictor variables with many thousands of new measurements, particularly in the North Pacific, relative to the GLODAPv2 version used for earlier versions of the global algorithms.
2. Numerous additional data sets from the Gulf of Mexico and the Mediterranean Sea as training data, fixing large and important data gaps in LIRv2.

3. The ability to return estimates of DIC.
4. Simple and improved estimation of anthropogenic perturbations to pH_T and DIC based on first principles, allowing better predictions of future changes in seawater carbonate chemistry.
5. Implementation of a distance weighting for the fit in ESPER_LIR, allowing more data to be used for each of the many regressions.
6. Ease-of-use changes that allow the insights from the LIR routines to be more easily adapted for regional applications.

In addition to LIR updates, we introduce new neural-network-based routines (ESPER_NN) to take advantage of the strengths of neural networks including the ability to model nonlinear relationships between predictors and estimated quantities (Tu 1996). In several important ways this new algorithm imitates the design of the "Carbonate system and Nutrients concentration from hYdrological properties and Oxygen using a Neural-network version B" (CANYON) algorithms designed by Sauzède et al. (2017) and updated by Bittig et al. (2018). The significant differences between ESPER_NN and the existing algorithms are as follows:

1. Inclusion of new data from the GLODAPv2.2020 data product (as with the LIR updates).
2. Like ESPER_LIR, ESPER_NN uses a new first-principles-based approach to estimate the impacts of long-term trends for pH_T and DIC.
3. ESPER_NN can function with 16 combinations of seawater properties requiring at minimum salinity and coordinate information, while alternative neural network approaches also require oxygen and temperature. While the temperature, salinity, and oxygen are often available and are frequently an ideal predictor combination, there remain applications where oxygen measurements are not available (due to absent, failed, or fouled sensors) or not desired as predictors (such as when estimating preformed properties from only conservative seawater properties, e.g., Carter et al. 2021).

By most validation metrics, the ESPER_NN routines perform comparably to ESPER_LIR routines and, in some places, they perform better (*see* "Assessment" section). Nevertheless, we contend there are reasons to maintain both approaches. First, the LIR routines offer a degree of simplicity and estimate explicability that lends them additional value. To highlight the explicability of the LIR estimates, we have added the ability to return the coefficients of the equations that were used to produce each estimate as an additional optional routine output. This may be useful when querying the LIR routines for an equation that could be used for a regional study in another application. Similarly, regional coefficients could be added into the ESPER_LIR coefficient files to produce a modified routine that seamlessly transitions to using regional relationships within a specific area such as a marginal sea, while still using the relationships derived for the open ocean outside of that region. Also, as we discuss later, there is merit to having and using multiple routines when the errors in the estimates appear to be partially independent, as appears to be the case with ESPER_LIR and ESPER_NN.

Both new routines are freely available as MATLAB functions at Zenodo (Carter 2021) and updates will be made available at the GitHub repository (*see*"Code Availability" section). Several changes have been made to the LIR function behavior that are noted alongside the reasoning behind the changes in Supporting Information S2.

Data products, training data, and test data

The primary data product used to train these algorithms is the GLODAPv2.2020 data product update (Olsen et al. 2020). In addition, we added data sets that will be included in the CARbon, tracer, and ancillary data In the MEDiterranean Sea (CARIMED) and that are included in the Coastal Ocean Data Analysis Project for North America (CODAP-NA; Jiang et al. 2021) data products. These data from the Mediterranean Sea (46 cruises spanning from 1976 to 2018 and covering all the sub-basins in the Mediterranean Sea) and the Gulf of Mexico (three cruises spanning 2007 to 2012) are included to ensure these important regions are well-constrained and the cruise information is provided in Supporting Information S1.1. These data products are focused on internal consistency and are inclusive for carbonate system measurements. We do not make a special effort in this study to incorporate high-resolution data from profiling sensors (e.g., 1 m oxygen values) or measurements from data products that focus on macronutrients or oxygen, but note that this could be an area of focus for future development.

As with previous versions of LIRs, we excluded data from GLODAPv2 that has not had secondary quality control checks (QC), and further omitted several sets of cruises that had large adjustments or appeared to have noisy measurements at depth (detailed in Supporting Information S1). We also excluded measurements from any bottle that lacked measurements for temperature, salinity, oxygen, and macronutrients (phosphate, silicate, and nitrate).

Homogenization of the variety of pH measurement types and calculations in GLODAPv2.2020 remains a challenge (*see* Supporting Information S1.2). As with LIRv2, the ESPERs return in situ pH_T estimates that are intended to be consistent by default with pH_T measured spectrophotometrically with purified m-cresol purple indicator dye and converted to in situ conditions, but can be made to return values that are intended to be consistent with pH_T calculated from DIC and TA at in situ conditions (as CANYON-B does by default) using an optional flag. These approaches for arriving at pH_T values have a documented disagreement (Carter et al. 2013, 2018; Williams et al. 2017; Fong and Dickson 2019; Álvarez et al. 2020), and we rely on the relationships developed by Carter

et al. (2018) to interconvert between these pH_T estimates. New observations are challenging the assumptions inherent to this approach (Takeshita et al. 2021), but currently there is insufficient data or mechanistic understanding to refine the relationships we use for interconversion.

For assessment purposes, we must separate validation data from training data and withhold the validation data from the versions of the algorithms used for assessment. It is better to withhold data from entire cruises to avoid obtaining unrealistically high skill estimates when reconstructing data from a synoptic cruise based on algorithms trained with other data from the same cruise. In past versions of LIRs, this assessment was conducted by creating algorithms that iteratively omitted each cruise while reconstructing data from the omitted cruises. However, this strategy would be too computationally intensive to employ with the ESPER_NN and would not provide a clear comparison to the CANYON-B neural network, which was trained with the original GLODAPv2 release. Instead, all data in GLODAPv2.2020 that were added following the original GLODAPv2 release (i.e., all cruises with GLODAPv2 cruise numbers ≥1000 and those incorporated from the Gulf of Mexico and the Mediterranean Sea) are used as test data for the validation versions of the algorithms that were trained only with the data in the original GLODAPv2 release. For general use, a release version of the ESPER_LIR and ESPER_NN algorithms was trained with the total data set to benefit from the recent data, and this release version is the only version provided at Zenodo and GitHub. Data within several marginal seas (the Gulf of Mexico, the Sea of Japan/East Sea, and the Mediterranean Sea) are omitted from the bulk global open-ocean assessment statistics because these are regions where the validation versions of the algorithms have insufficient training data (i.e., none) to produce estimates. Similarly, data from the Arctic (here: north of 67.5°N) are withheld from the global assessment step because the Arctic is a problematic region for algorithms (see "Regional tests" section). Instead, algorithm performance is separately assessed in these regions to explore the limitations of the approaches used ("Regional tests" section). The numbers of valid, quality-controlled measurements available for each algorithm version in each subset of the data are given in Table 1.

Table 1. Numbers of viable measurement combinations available for each property within the indicated data product subsets. The "total" column reflects the training data for the released routines, whereas the "GLODAPv2" column reflects the training data for the validation routines used to assess the algorithms against new/assessment data.

Property	GLODAPv2	New/assessment	Total
Phosphate	540,511	146,263	711,347
Nitrate	540,511	146,263	711,347
Silicate	540,511	146,263	711,347
Oxygen	540,511	146,263	711,347
TA	203,502	71,832	286,080
pH	162,783	53,615	222,822
DIC	244,062	71,326	323,328

Anthropogenic impacts on carbonate chemistry

The LIPHR (i.e., LIRv2 for pH_T) and CANYON-B algorithms use "estimate year" (i.e., for LIPHR, this is the calendar year expressed as a decimal, where the midpoint of the year 2020 would be given as 2020.5) as a predictor for seawater properties (or their reconstruction errors in the case of LIPHRv2) to capture the impacts of long-term trends on pH_T estimates and the training data. However, recent research suggests that decadal variability in seawater property trends can rival, regionally, the magnitudes of the secular trends. This is true even for C_{ant} which exhibits a large secular trend (Woosley and Millero 2016; DeVries et al. 2017; Carter et al. 2019*a*). This finding implies that empirical fits risk projecting trends from cyclical natural variability into the future. LIPHR avoids some biases from regional natural variability by using global empirical fits over density intervals, but, as a result, the routine is unable to distinguish between regions with rapid (e.g., the North Atlantic) vs. slow (e.g., the North Pacific) C_{ant} accumulation. In addition, LIPHR assumes a fixed OA rate over time, but OA rates might be expected to accelerate due to the approximately exponential increase in atmospheric CO_2. Therefore, while algorithms like LIPHR seem to accurately predict contemporaneous deep pH_T, it is likely that biases will emerge over the coming years, particularly in regions where C_{ant} penetration is large such as the North Atlantic (Gruber et al. 2019). The risks of natural variability biasing empirical trend projections are perhaps more acute for the properties that have weaker secular trends than DIC and pH_T, such as nutrients and oxygen, although the empirical trends in these properties are usually smaller components of the overall variability in their estimates.

Given the challenges associated with accurately quantifying secular changes with short-term, empirical information, ESPER_LIR and _NN rely on a first-principles-based estimate of C_{ant} and its impacts on pH_T. This approach assumes that exponential increases in atmospheric anthropogenic CO_2 should eventually result in marine C_{ant} concentrations that increase at rates proportional to atmospheric anthropogenic CO_2 concentrations. In other words, this approach relies on the assumption that C_{ant} is in transient steady state (Gammon et al. 1982; Tanhua et al. 2007); this is an assumption used to adjust data to reference years in the most recent global C_{ant} distribution change estimates for the 1994–2007 period (Gruber et al. 2019). This implies that, locally, the "shape" of the C_{ant} vertical profile (or C_{ant} vertical gradient) should remain constant over time while

atmospheric CO_2 and ocean C_{ant} values are increasing exponentially according to:

$$C_{ant\_year\_location} = C_{ant\_2002\_location} e^{0.018989(year-2002)}. \quad (1)$$

Therefore, if a C_{ant} value is known for a location in a reference year (e.g., $C_{ant\_2002\_location}$ in 2002 c.e.), then C_{ant} can be estimated for that location in a desired year ($C_{ant\_year\_location}$). The coefficient within the exponent is derived by solving Eq. (1) to match Gruber et al.'s (2019) assumption of a ~28% C_{ant} increase over the 13 years from 1994 to 2007 (*see* their methods supplement). We note that this approach is not able or intended to resolve non-steady-state variations in C_{ant} (Gruber et al. 2019), and the errors in the estimates that result from this deficiency are included implicitly in the assessed overall uncertainty estimates.

For the ESPERs, we utilize a gridded C_{ant} product referenced to the year 2002 (Lauvset et al. 2016). This product was created using the transit time distribution (TTD) method (Waugh et al. 2006), and gridded to the same 1° × 1° latitude/longitude resolution with 33 depth surfaces as the Global Data Analysis Project (GLODAPv2) gridded data product. This reference 2002 field can be used with Eq. 1 to estimate the difference between C_{ant} in 2002 and C_{ant} in the year in which a measurement was made, or an estimate is desired. Therefore, rather than having a time dependent prediction of pH_T or DIC, we take the following steps to address anthropogenic trends (Fig. 1):

1. Start with the unmodified training data set.
2. Transform all training data to the year 2002 by adding/removing the missing/excess C_{ant} if they are measured before/after 2002.
3. Train the pH_T or DIC algorithms on this modified training data.
4. Predict pH_T or DIC without a time dependence for 2002.
5. Transform the C_{ant} to the desired year (if other than 2002), recalculating DIC and pH_T with the new C_{ant} total accordingly.

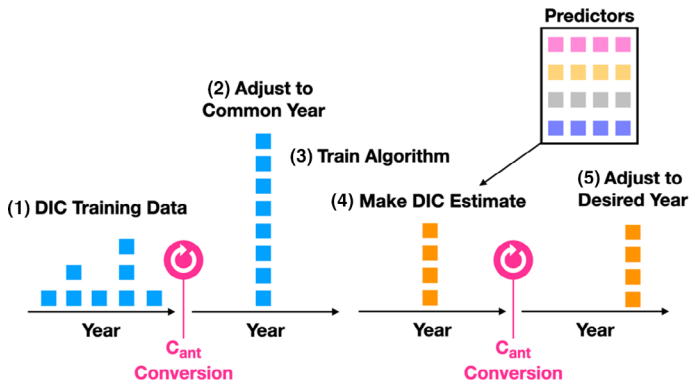

Fig 1. A schematic showing the approach for adjusting training data and estimates for effects of anthropogenic carbon accumulation. The "common year" is 2002.

Steps 1 through 3 were performed before training the routines, while steps 4 and 5 are performed by the ESPER code each time it is called. Supporting Information S1.3 provides more detail for the pH_T recalculations noted in step 5.

There are uncertainties associated with the assumptions underlying both the 2002 gridded C_{ant} data product and the transient steady state approach—particularly in regions where there are limited measurements of chlorofluorocarbons and other tracers used to calibrate the TTD approach. We therefore assert that Eq. 1 should not be used to estimate C_{ant} distributions for any application where C_{ant} is of primary interest. However, uncertainties in the adjustments that come from changes in these C_{ant} estimates over time should be modest for a window of time around the year 2002 c.e., the year in which the adjustments are zero by definition. Eq. (1) implies that adjustment errors will be smaller than errors in the underlying 2002 C_{ant} distributions for any estimate before 2039 (i.e., the C_{ant} doubling time after 2002). As the training data are also adjusted in step 2, the effective magnitudes of the adjustments are related to the difference between the years of the estimates and the average measurement years of the training data used for those algorithms (which for most regions and algorithms is close to 2002 c.e.). These ESPERs should therefore be used with increasing caution for DIC and pH_T after ~ 2030. Regardless of these challenges, this parameterization of OA rates should be more accurate moving forwards than that used by LIPHR, and any improvements in the C_{ant} estimates should directly reduce estimate bias in the modern era and the near future. Notably, implementing this approach decreased overall training data reconstruction root mean squared error for DIC by > 10% and decreased the trend in the DIC reconstruction error from ~ 0.49 μmol kg^{-1} yr^{-1} to less than 0.03 μmol kg^{-1} yr^{-1}. We caution that these assumptions do not explicitly consider declines in ocean carbon uptake efficiency and the assumption of exponential growth can lead to very large DIC accumulations when used for distant projections. Future atmospheric CO_2 concentrations are highly uncertain, and user discretion is advised for any projections.

There is no time variance for ESPER estimates of quantities other than pH_T and DIC.

ESPER_LIR construction

ESPER_LIR broadly functions similarly to LIRv2, which is described in detail by Carter et al. (2018). As with LIRv2, the ESPER_LIR algorithms use regression coefficients (*C*) that are specific to each of 16 equations and 44,957 locations on a 5° latitude × 5° longitude × 33 depth ocean interior grid subsampled from the World Ocean Atlas gridded product grid. These coefficients are interpolated in 3D space to the locations where regression coefficients are desired. The algorithm then uses the coefficients with user-provided seawater property predictor information (*P*) to produce property estimates.

Table 2. The combinations of predictors used to estimate each property for each of the 16 ESPER equations. Rows with a checkmark indicate the predictors (listed above by property) are included in that equation for that property.

Property	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5
Phosphate	S	θ	Nitrate	Oxygen	Silicate
Nitrate	S	θ	Phosphate	Oxygen	Silicate
Silicate	S	θ	Phosphate	Oxygen	Nitrate
Oxygen	S	θ	Phosphate	Nitrate	Silicate
TA	S	θ	Nitrate	Oxygen	Silicate
pH	S	θ	Nitrate	Oxygen	Silicate
DIC	S	θ	Nitrate	Oxygen	Silicate
ESPER Equation #					
1	✓	✓	✓	✓	✓
2	✓	✓	✓		✓
3	✓	✓		✓	✓
4	✓	✓			✓
5	✓	✓	✓	✓	
6	✓	✓	✓		
7	✓	✓		✓	
8	✓	✓			
9	✓		✓	✓	✓
10	✓		✓		✓
11	✓			✓	✓
12	✓				✓
13	✓		✓	✓	
14	✓		✓		
15	✓			✓	
16	✓				

The LIR algorithms are constructed by fitting 16 different regressions that relate the properties of interest, X (e.g., silicate, nitrate, phosphate, oxygen, TA, DIC, and pH_T), to combinations of up to five predictor properties, P (including salinity, potential temperature, nitrate, phosphate, oxygen, and silicate), which are specific to each property of interest (Table 2). Each equation uses between one and five predictor properties and the generalized predictor equation is:

$$X = C_0 + \sum_{i=1}^{n} C_i P_i. \quad (2)$$

The 16 variants on Eq. 2 are referred to as "ESPER equations" 1 through 16. Unlike LIRv2, depth is never used as a predictor for ESPER_LIR and is only used as a coordinate for regression coefficient interpolation. Versions with depth included as a predictor performed similarly or worse than versions with depth omitted during early testing.

The regression coefficients C_i and C_0 are fit 44,957 times for each of the 7 estimated properties and each of the 16 ESPER equations. At each grid location, "local" data are selected from the subset of all data that are within 15° in latitude, 30°/cosine (latitude) in longitude, and within either $(100 + z/10)$ m depth or 0.1 kg m^{-3} of the estimated density of seawater at that coordinate location. Here z is the coordinate depth in meters. As with LIRv2, these window dimensions are iteratively doubled when fewer than 100 measurements fall within the windows. These data selection windows are initially twice as wide as the windows used in LIRv2 in all dimensions. Doubling the baseline size of these windows is intended to include more data on average for the regression fits, introduce more modes of oceanographic variability into the fitting data, and thereby reduce multicollinearity. The average absolute values of regression coefficients in ESPER_LIR are only 80% of the average absolute values of the coefficients in LIRv2, suggesting ESPER_LIR is subject to less multicollinearity than LIRv2. However, widening the windows risks making the regressions less appropriate locally, so a weighting term is used that is equal to:

$$W = \max\left(5, \left(\frac{10(\Delta z)}{100+z}\right)^2 + (\cos(\text{lat})(\Delta\text{lon}))^2 + 4(\Delta\text{lat})^2\right)^{-2}. \quad (3)$$

The weighting term W reduces the cost of regression misfits to data that are distant or at significantly different depths from

the regression coordinate location, and the maximum function caps the weights (at a value equivalent to the weight found when 5° latitude away) to ensure the regressions are not overly fit to data very near the coordinate where the denominator approaches 0. The Δz term is the difference between the regression coordinate depth (z) and the depth of the measurements. The Δlon is the minimum difference in the measurement and coordinate longitudes when using either the −180° to 180° or 0° to 360° conventions, and Δlat is the difference between the measurement and coordinate latitudes. The regression coefficients (C_0 and C_{Pi}) are then fit using a regression of the form:

$$XW = \left(C_0 + \sum_{i=1}^{n} C_{Pi} P_i \right) W. \quad (4)$$

As with LIRv2, data outside of the Atlantic, Mediterranean, and Arctic are excluded when fitting Northern Hemisphere regression coordinates within the Atlantic, Mediterranean, or Arctic—and vice versa—in order to prevent use of data from across Central America or the Bering Strait. The widths of the data inclusion windows and the coefficients in the weighting function were optimized by selecting the variant of eight combinations that had the best validation statistics. However, some of the combinations yielded comparable results for some predictors, so this parameter tuning process should not be considered exhaustive.

ESPER_NN construction

ESPER_NN relies upon a collection of feed-forward neural networks to estimate seawater properties with a similar operation to the LIR algorithm and a similar structure to the CANYON-B algorithm: ESPER_NN uses the same combination of predictor measurements as ESPER_LIR to produce estimates of the same properties, and does so with a function call that has similar syntax. Unlike ESPER_LIR, in addition to the predictors noted in Table 2, the ESPER_NN algorithm uses latitude, depth, cos(longitude-20°E), and cos(longitude-110°E) as predictors in each equation, making the estimates somewhat more analogous to a mapping approach than the ESPER_LIR estimates. Similar, but not identical, parameters are used in CANYON (Sauzède et al. 2017) and CANYON-B (Bittig et al. 2018): unlike the original CANYON, ESPER_NN offsets the 0 longitude for the reasons noted by Bittig et al. (2018), specifically that cos(lon) loses explanatory power at the prime meridian, which is a region of oceanographic significance. Offsetting longitudes to 20°E (and 110°E) puts these regions of minimum explanatory power over land masses to the extent possible.

ESPER_NN uses 896 neural networks in total: 8 neural networks (4 in each of 2 large ocean regions: see later) are used for each of the 16 combinations of predictors used for each of the 7 property estimates. ESPER_NN averages estimates from a "committee" or ensemble of four neural networks with different combinations of neurons and hidden layers to minimize the impact of errors from any one neural network. These four neural networks include a single one-hidden-layer network with 40 neurons, and three two-hidden-layer networks with 30/10, 25/15, and 20/20 neurons in the 1st/2nd hidden layers. One committee of neural networks is used in the Indo-Pacific-Southern Ocean regions and an additional committee used in the Atlantic Ocean, Arctic Ocean, and Mediterranean Sea. The ESPER_NN algorithm linearly interpolates between the outputs of these two committees of neural networks by latitude across the Southern Atlantic and the Bering Sea, being fully in the Indo-Pacific-Southern Ocean network by 44°S in the Southern Atlantic and fully in the Atlantic, Arctic, and Mediterranean network by 34°S. Similarly, the North-Pacific-to-Arctic transition occurs between 62.5°N and 70°N along Pacific longitudes. After this meridional blending step, there is a zonal transition implemented in the Southern Atlantic between these blended values and the Indo-Pacific-Southern Ocean network starting at 19°E and being completely transitioned at 27°E.

Techniques exist for illuminating the relative importance of predictor variables in machine learning approaches (e.g., Olden and Jackson 2002), but the exact equations used by the ESPER_NN algorithm are nevertheless more opaque and less explainable than the LIR equations. The networks are fit using the MATLAB r2017 Machine Learning Toolbox "feedforwardnet" and "train" function defaults, which include Levenberg Marquardt optimization with 15% of input data reserved for assessment during iterative fitting steps. However, the neural networks have been encoded as functions, so users do not require the Machine Learning Toolbox to operate ESPER_NN.

Mixed estimates

Bittig et al. (2018) showed that linear regression and neural network estimates frequently have independent error fields. From this observation, they proposed that it might be advantageous to combine estimates from both approaches. We test this idea and find that it has merits in many circumstances. We therefore also release a wrapper function "ESPER_Mixed.m" that calls both routines, ESPER_LIR and ESPER_NN, and averages the estimates. We do not provide a similar wrapper function for CANYON-B, but we note that our assessment suggests the findings for the mixed approach could also apply to a mixed version of CANYON-B and ESPER_LIR Eq. 7. The ESPER_Mixed routine is assessed alongside the other algorithms in "Assessment" section.

Uncertainty estimation

The routines can return uncertainties for every property estimate, and the uncertainty values vary with input depth and salinity. These uncertainties are estimated at the 1σ (i.e., one standard uncertainty) level, so we would expect

$\sim 95\%$ of new measurements that have been through the GLODAPv2 QC process to fall within windows of $\pm$ twice the ESPER estimated uncertainties. The LIRv2 uncertainty estimation strategy for TA (Carter et al. 2018) is slightly modified and then implemented for all properties estimated by the two ESPERs. As before, this approach interpolates baseline error estimates (E_{X_Est}) in depth and salinity space. The interpolated values are based on the root mean squared errors (RMSEs) of all predictions from the validation versions of the routines within bins of salinity and depth. As with LIRv2, ESPER_LIR also scales these methodological uncertainties using user-provided predictor uncertainty estimates. The following equation is used when the user provides uncertainties for the predictors (E_{Pi_Provided}) that exceed the default assumed input uncertainties (Table 3):

$$E_{X\_\mathrm{Output}} = \sqrt{E_{X\_\mathrm{Est}}^2 - \sum_{i=1}^{n}\left(\frac{\partial X}{\partial P_i}E_{Pi\_\mathrm{Default}}\right)^2 + \sum_{i=1}^{n}\left(\frac{\partial X}{\partial P_i}E_{Pi\_\mathrm{Provided}}\right)^2}. \quad (5)$$

If the optional E_{Pi_Provided} input is omitted then it is assumed that E_{Pi_Provided} equals E_{P_Default} (Table 3), and the two summed terms in this equation cancel. Here $\frac{\partial X}{\partial P_i}$ is the sensitivity of the property estimate X to the ith predictor P_i and the E_{Pi} terms are the default and the user-provided predictor uncertainties. For the ESPER_LIRs, the $\frac{\partial X}{\partial P_i}$ values equal the C_{Pi} terms. For ESPER_NN calculations, the algorithm determines the sensitivities by iteratively perturbing the input predictors if and only if the user specifies larger-than-default predictor uncertainties. The uncertainties in Table 3 are the minimum uncertainties allowed by the calculations because these are the assumed uncertainties in the best open ocean training data available, so these uncertainties reflect one of the upper limits on the quality of estimates achievable with the algorithms regardless of the quality of the predictor measurements. The sole difference from the approach used for LIRv2 TA estimates is that the interpolated uncertainties now include the component of uncertainty that originates from potential errors in the training data. This saves a step in the calculations while providing numerically equivalent results.

Table 3. Assumed default measurement uncertainties, or E_{Pi_Default} or E_{X_Default} as defined in the text.

Property	Uncertainty	Units
S	0.003	
θ	0.003	°C
Phosphate	2%	μmol kg^{-1}
Nitrate	2%	μmol kg^{-1}
Silicate	2%	μmol kg^{-1}
Oxygen	1%	μmol kg^{-1}

The uncertainty for an ESPER_Mixed estimate is assessed simplistically as the minimum uncertainty assessed for the two component ESPER_LIR and ESPER_NN estimates ("Mixed ESPER" section).

Assessment

Routines are validated using versions of the algorithms trained only with the data that were present in the original GLODAPv2 release (Table 1). This cutoff was chosen to make the validation algorithms for ESPER_LIR and ESPER_NN comparable to the LIRv2 and CANYON-B routines to the degree possible. These "validation" versions of the algorithms are then used to recreate the "validation data set," or the newly added data in the GLODAPv2.2019 and GLODAPv2.2020 updates plus the other cruises from the Mediterranean Sea and the Gulf of Mexico. The reconstruction errors for these new measurements are used to derive error statistics for the five routines that we assess (LIRv2, ESPER_LIR, ESPER_NN, CANYON-B, and ESPER_Mixed). The validation data set is in some ways not ideal, in that it is not evenly distributed globally and there is spatial overlap between the test and the training data sets (Fig. 2). An alternate approach to assessing prediction errors involves omitting all training data from regions of the ocean representative of data gaps between cruises, and then estimating the errors within these gaps. This approach has been used previously by Sauzède et al. (2017) and Carter et al. (2018), but was found to generally yield smaller uncertainty estimates in the open ocean than approaches that omit entire cruises (Carter et al. 2018), so we conservatively rely on the cruise-omission assessments. The additional data sets from the Gulf of Mexico and the Mediterranean Sea that were incorporated into this paper were omitted from the global-average validation data set because neither had undergone secondary QC and because a small subset of the Mediterranean Sea data from GLODAPv2 had been previously incorporated into the training data product for some algorithms but not others. New measurements from the Sea of Japan/East Sea, a biogeochemically distinct region where no previous measurements existed in the original GLODAPv2 product, are also omitted from bulk validation statistics. However, validation statistics for these regions are given separately ("Regional tests" section).

The reported validation statistics are bias (average reconstruction error), RMSE, and the number of new measurements used for each assessment (N). The 10$^{\mathrm{th}}$, 50$^{\mathrm{th}}$, and 90$^{\mathrm{th}}$ error percentiles were examined as potential additional statistics, but these statistics were within expectations when assuming normally distributed errors with the given RMSE and bias statistics.

Macronutrients

The routines work well for macronutrients (i.e., phosphate, nitrate, and silicate) when given at least two predictors, reproducing the validation data with low average bias and a

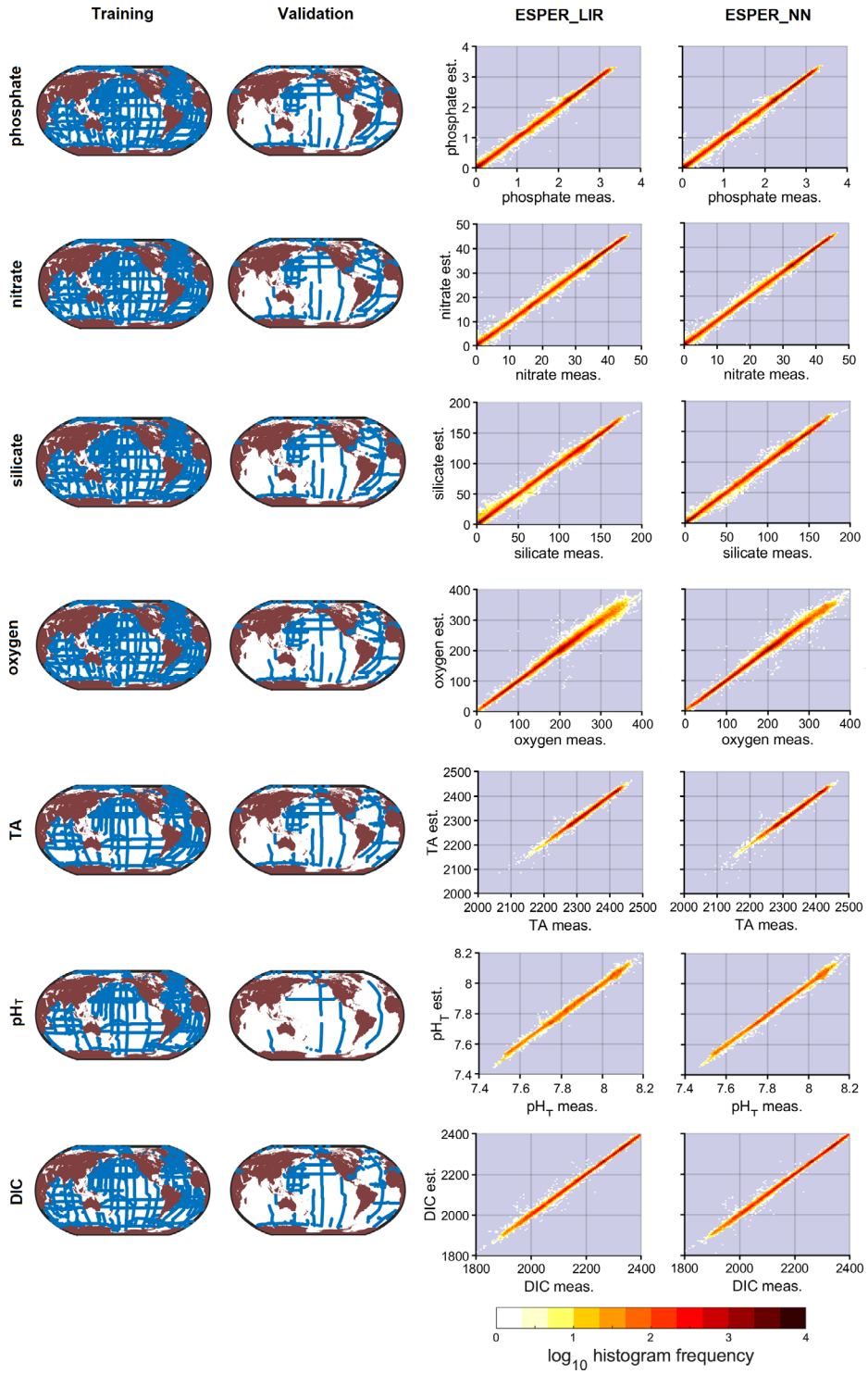

Fig 2. The first column contains maps of the measurement locations used to train the ESPER_LIR_validation and ESPER_NN_validation algorithms. The second column maps the validation data used to assess these versions of the algorithms. The final ESPER_NN and ESPER_LIR algorithms are trained with data shown in both rows of maps. Panels in the right two columns are two-dimensional histograms showing the number of measurements that fall within bins of measured (*x*-axes) and estimated (with ESPER Eq. 1 from Table 2, *y*-axes) values of the indicated properties for ESPER_LIR. Color indicates the number of measurements in each bin (bins are small enough as to appear to be pixels), with darker colors indicating more measurements. The rightmost column is the same as the 3rd column from the left, but for ESPER_NN property estimates. An ideal algorithm would have darker colored boxes along the 1 : 1 lines in the first two rows.

RMSE that is comparable to the measurement uncertainties (Tables 4–6). Phosphate and nitrate have a strong and well-documented covariance in the ocean (Redfield et al. 1963). This covariance results in low RMSE statistics for the equations relating these properties to one another (e.g., ESPER Eqs. 1 and 2 in Table 2), but reduces the value of adding the other as a predictor when one is already included. This covariance is less strong between silicate and either phosphate or nitrate, and oxygen is comparably useful to the macronutrients when predicting silicate. Unsurprisingly, the equations with more fitting parameters tended to perform better, and the RMSE ranged from being comparable to nominal $\sim 2\%$ measurement uncertainty at best (or ± 0.04 μmol kg^{-1} for a phosphate measurement of 2 μmol kg^{-1}; Olsen et al. 2016) to 3–4 times worse when only S and coordinate information is used in the prediction. All algorithms assessed perform comparably for the ESPER equations using T, S, and oxygen as predictors (i.e., ESPER Eq. 7), but LIRv2 performs slightly worse for silicate. LIRv2 performs comparably to alternatives for many macronutrient estimates, but alternatives outperform LIRv2 for the equations with the largest RMSE values and fewest predictors (e.g., ESPER Eqs. 12 and 16), suggesting that the modifications in ESPER_LIR have resulted in an improvement in the least-accurate estimates. Likely, this is due to the larger number of measurements available for each regression in ESPER_LIR relative to LIRv2. Unlike the ESPER_LIR_validation routine assessed here, the released version of ESPER_LIR benefits from including the newly added data in the recent updates to GLODAP, and is therefore preferred to LIRv2 even when the validation statistics are comparable.

Oxygen

Validation statistics are reasonable for oxygen though persistently greater than the nominal 1% measurement uncertainty (i.e., 3 μmol kg^{-1} for a 300 μmol kg^{-1} measurement; Olsen et al. 2016), ranging from 4.5 to 13.2 μmol kg^{-1} in the global ocean for ESPER_NN_validation and ESPER_LIR_validation (Table 7). LIRv2 is also comparable, but again shows worse validation statistics for equations with fewer predictors and larger RMSE values. The statistics are markedly better at intermediate depths, and range from 2.7 to 6.0 μmol kg^{-1} between 1000 and 1500 m depth for ESPER_NN_validation. Below the well-lit surface ocean there is no gas exchange and essentially no primary production of organic matter, and the algorithms are therefore better able to capture the fewer processes controlling oxygen distributions. As a result, the oxygen algorithms perform less well at higher oxygen concentrations, which is evident in the larger error statistics globally than in the intermediate depth statistics, as well as in the comparatively diffuse cloud of estimates in the upper right of the oxygen histograms in Fig. 2. Interestingly, the neural network estimates in Fig. 2 appear less diffuse than the LIR-based estimates: the RMSE for ESPER Eq. 1 for only the top 200 m is 8.6, 7.6, and 8.0 μmol kg^{-1} for the LIR, NN, and Mixed validation ESPER variants, respectively. This suggests that the neural network framework is more skillful at capturing the nonlinear relationships between properties that can result in the presence of gas exchange and primary production in the surface ocean. Oxygen estimates show a non-negligible bias, overestimating oxygen by an average 0.9 μmol kg^{-1} for all three algorithms across ESPER equations. It should be noted that a large amount of the validation data used for this assessment are located within the North Pacific where oxygen concentrations are low, so this could reflect a small regional bias in the algorithms, a tendency to overestimate lower oxygen concentrations, or differences between the test and the training data products. Supporting this idea, the released versions of the algorithms—which use all data as training data—still have a 0.6 μmol kg^{-1} bias for the ESPER_Mixed_validation test data reconstructions while having a -0.1 μmol kg^{-1} bias for the ESPER_Mixed_validation training data reconstructions (i.e., GLODAPv2) and no significant bias for both data subsets combined.

Total titration seawater alkalinity

Seawater alkalinity continues to show strong predictability even with comparatively few predictors (Table 8) and has the smallest relative range in RMSE values with the least precise estimates having a RMSE that is less than double the RMSE of the most precise estimates (ranging from 3.7 to 5.2 μmol kg^{-1} for TA for ESPER_NN_validation estimates). The small range in assessed RMSE values is expected because all ESPER equations use S, and freshwater cycling is a major driver explaining variability in both S and TA. The excellent validation metrics for new and existing algorithms for TA likely reflect particularly precise TA measurements in the newly added cruises in GLODAPv2.2020, in part due to increased use of certified reference materials for TA (Dickson et al. 2003).

Interestingly, there is an estimate bias averaging 0.5 to 1 μmol kg^{-1} across ESPER equations for the various routines. It is difficult to identify the cause of these average mismatches when considering that the GLODAP secondary QC effort already adjusted several cruises to be in line with the existing GLODAPv2 data product. However, Olsen et al. (2019) note that many of the newly added cruises in the North Pacific show a negative bias against earlier cruises, consistent with this observation. Also, many of these cruises use single-point spectrophotometric TA titration endpoint detections, which Bockmon and Dickson (2015) previously noted could be a source of disagreement with TA values from full-pH-range titration fits. Interestingly, Sharp and Byrne (2020) have provided a mechanistic explanation that would account for these analytical disagreements if alkaline organic molecules were present in open-ocean seawater. While this discussion highlights the challenges of creating a consistent data product across research groups, the high precision and modest bias of this TA reconstruction nevertheless demonstrates the high quality of the underlying measurements and the importance of the GLODAPv2 secondary QC process.

In situ pH on the total scale

There is some difficulty comparing across pH_T algorithms because the training data for earlier pH_T algorithms were supplemented with several additional cruises (Bittig et al. 2018; Carter et al. 2018), many of which were since added to the GLODAPv2 data product in annual updates. This means that some algorithms would benefit from overlap between the training and validation data products in this comparison. The comparison cannot simply be limited to the truly new cruises because there are not many additional cruises where purified spectrophotometric dye measurements were made that were not used to train earlier algorithms; we limit our comparison to cruises with these spectrophotometric measurements because it has been shown that there are consistent disagreements between measured and calculated pH_T (Carter et al. 2018; Álvarez et al. 2020). Moreover, measurements made with purified dyes are consistent with measurements made by sensors that have been shown to have the expected Nernstian response to pH_T changes (Takeshita et al. 2020) lending support to the use of spectrophotometric pH_T values over the disagreeing calculated values. Complicating the comparison further, the three new cruises that were not included in LIRv2 or CANYON-B pH_T training data that do meet our criteria had large adjustments applied during the GLODAP secondary QC. Therefore, for this study we do not re-assess LIRv2 or CANYON-B, and instead show that the ESPERs have similar validation statistics (Table 9) to those published by earlier validation efforts for these algorithms (Bittig et al. 2018; Carter et al. 2018). We do note however, that the statistics obtained when we assess all four algorithms using *T*, *S*, and oxygen with the same data (not shown) are quite close to each other despite the partial overlap between training and validation data sets. This suggests all four algorithms are valid for pH_T.

It is difficult to read into pH_T validation statistics too much given the comparatively small number of valid assessment data points. However, one pattern in pH_T assessment statistics that is apparent is that pH reconstructions benefit significantly from the use of either nitrate or oxygen as predictors, as these predictors provide information regarding organic matter remineralization. The ESPER equations with neither quantity have higher RMSE values, even when silicate is included as a predictor.

Total dissolved inorganic carbon

The routines reproduce DIC measurements with good skill and a small positive average bias, with RMSE values ranging from 4.8 to 16.7 μmol kg^{-1} globally and 3.2 to 7.0 μmol kg^{-1} at intermediate depths for the various validation versions (Table 10). Assessment statistics are comparable across the three routines that estimate DIC (LIRv2 does not). We caution that DIC does not have seasonal resolution in the surface ocean in most regions of its training data product. Therefore, estimates within the surface ocean should be treated with caution, and we recommend avoiding interpreting seasonality in the ESPER estimates. This caution applies to all property estimates, but is important to note for DIC specifically because of the high sensitivity of DIC to most modes of seasonal variability and the large scientific interest in seasonal DIC cycling. DIC calculations from measured pH or pCO_2 and estimated TA are expected to be less challenged by the lack of seasonal resolution than direct DIC estimates, as TA seasonality is usually less pronounced than DIC seasonality. These two approaches to DIC seasonality reconstruction can return quite different results in the surface ocean (Supporting Information S1.4). There are empirical routines for global DIC estimation (Broullón et al. 2020) and surface DIC estimation (Gregor and Gruber 2021) that are also trained with the surface pCO_2 measurements. In the many regions where surface pCO_2 has better seasonal data coverage than GLODAPv2, these routines are likely to better resolve DIC surface seasonality than ESPER or other DIC algorithms trained primarily with discrete DIC measurements.

Regional tests

We assess the performance of the algorithms in eight regions independently (Fig. 3). Some of these regions are where biogeochemical Argo floats are currently being deployed (i.e., the North Atlantic, California Current, Equatorial Pacific, and the Southern Ocean) and therefore where there is additional interest in the performance of the algorithms. Other regions are biogeochemically distinct places where there were no training data used for the CANYON-B and/or LIRv2 algorithms (i.e., Sea of Japan/East Sea, Gulf of Mexico, and the Mediterranean). These regions therefore allow tests of the likely errors one can expect when applying global algorithms to biogeochemically distinct regions where there were no available training data. Finally, the Arctic is a problematic region for the algorithms that warrants special attention.

We first consider the Southern Ocean, the Equatorial Pacific, the California Current, and the Northern Atlantic. The validation statistics in these regions where there are active ongoing biogeochemical float deployment efforts are, for the most part, consistent with the global average statistics. The Northern Atlantic shows validation statistics that are somewhat worse than global averages for macronutrients and oxygen and the California Current shows oxygen RMSE values that are equally elevated. Given the active physical processes and biogeochemical cycling in these regions of interest (and the comparatively small validation data set in the California Current), none of these sets of validation statistics are unexpected. We therefore conclude that the algorithms should function within expectations in these important regions and suggest Table 11 can be used to get a sense for how the global validation statistics might vary on a regional level.

The Sea of Japan/East Sea provides an excellent case study to assess the use of algorithms in regions without training data for three reasons: (1) this region had no data in the first GLODAPv2 release, and thus is a region where neither LIRv2 nor CANYON-B had training data; (2) a large quantity of high-

Table 4. Assessment statistics, reported as bias (± RMSE) in μmol kg^{-1}, for various phosphate estimation routines presented both globally (top rows) and for intermediate ocean depths (bottom rows, provided for comparison only as there are no float-based phosphate sensors calibrated using algorithms). The equation numbers are specific to the LIR approach, but the equivalent seawater property predictors are used for the other algorithms in the same row.

Global

	LIRv2	ESPER_LIR	ESPER_NN	CANYON-B	Mixed
N	146,263	146,263	146,263	146,263	146,263
Eq. 1	0.002 (± 0.035)	0.001 (± 0.036)	0.001 (± 0.036)	—	0.003 (± 0.039)
Eq. 2	0.001 (± 0.039)	0.000 (± 0.038)	0.001 (± 0.037)	—	0.002 (± 0.039)
Eq. 3	0.003 (± 0.044)	0.001 (± 0.044)	0.001 (± 0.040)	—	0.003 (± 0.042)
Eq. 4	−0.001 (± 0.061)	−0.006 (± 0.060)	−0.003 (± 0.053)	—	0.000 (± 0.045)
Eq. 5	0.002 (± 0.036)	0.001 (± 0.037)	0.002 (± 0.036)	—	0.003 (± 0.039)
Eq. 6	0.001 (± 0.041)	−0.001 (± 0.039)	0.001 (± 0.038)	—	0.002 (± 0.039)
Eq. 7	0.005 (± 0.052)	0.004 (± 0.051)	0.003 (± 0.043)	0.004 (± 0.043)	0.004 (± 0.045)
Eq. 8	−0.003 (± 0.089)	−0.003 (± 0.086)	−0.002 (± 0.075)	—	0.001 (± 0.053)
Eq. 9	0.003 (± 0.036)	0.002 (± 0.037)	0.002 (± 0.036)	—	0.003 (± 0.039)
Eq. 10	0.002 (± 0.040)	0.000 (± 0.039)	0.001 (± 0.039)	—	0.002 (± 0.038)
Eq. 11	0.005 (± 0.048)	0.002 (± 0.049)	0.002 (± 0.044)	—	0.003 (± 0.043)
Eq. 12	−0.003 (± 0.079)	−0.006 (± 0.065)	−0.003 (± 0.057)	—	0.001 (± 0.046)
Eq. 13	0.004 (± 0.037)	0.002 (± 0.038)	0.003 (± 0.037)	—	0.003 (± 0.039)
Eq. 14	0.002 (± 0.043)	0.000 (± 0.040)	0.002 (± 0.040)	—	0.003 (± 0.039)
Eq. 15	0.011 (± 0.069)	0.008 (± 0.067)	0.007 (± 0.059)	—	0.005 (± 0.051)
Eq. 16	0.008 (± 0.152)	0.005 (± 0.141)	0.004 (± 0.129)	—	0.004 (± 0.078)

Intermediate depth only **(i.e., > 1000 m and < 1500 m depth)**

N	14,397	14,397	14,397	14,397	14,397
Eq. 1	0.009 (± 0.030)	0.007 (± 0.030)	0.007 (± 0.028)	—	0.007 (± 0.029)
Eq. 2	0.009 (± 0.031)	0.006 (± 0.030)	0.008 (± 0.030)	—	0.008 (± 0.029)
Eq. 3	0.011 (± 0.032)	0.008 (± 0.032)	0.009 (± 0.030)	—	0.008 (± 0.030)
Eq. 4	0.012 (± 0.040)	0.007 (± 0.038)	0.006 (± 0.036)	—	0.007 (± 0.031)
Eq. 5	0.010 (± 0.029)	0.007 (± 0.029)	0.008 (± 0.029)	—	0.008 (± 0.029)
Eq. 6	0.009 (± 0.030)	0.006 (± 0.030)	0.008 (± 0.030)	—	0.008 (± 0.029)
Eq. 7	0.011 (± 0.031)	0.008 (± 0.031)	0.010 (± 0.030)	0.011 (± 0.031)	0.009 (± 0.030)
Eq. 8	0.012 (± 0.044)	0.003 (± 0.041)	0.005 (± 0.046)	—	0.007 (± 0.034)
Eq. 9	0.009 (± 0.030)	0.007 (± 0.030)	0.007 (± 0.029)	—	0.008 (± 0.029)
Eq. 10	0.009 (± 0.031)	0.006 (± 0.030)	0.005 (± 0.029)	—	0.006 (± 0.028)
Eq. 11	0.011 (± 0.032)	0.008 (± 0.032)	0.009 (± 0.031)	—	0.008 (± 0.030)
Eq. 12	0.012 (± 0.046)	0.005 (± 0.038)	0.005 (± 0.038)	—	0.007 (± 0.032)
Eq. 13	0.010 (± 0.030)	0.007 (± 0.029)	0.007 (± 0.029)	—	0.007 (± 0.029)
Eq. 14	0.009 (± 0.031)	0.006 (± 0.030)	0.007 (± 0.030)	—	0.007 (± 0.028)
Eq. 15	0.012 (± 0.033)	0.008 (± 0.031)	0.010 (± 0.032)	—	0.009 (± 0.031)
Eq. 16	0.013 (± 0.056)	0.000 (± 0.049)	0.002 (± 0.053)	—	0.005 (± 0.037)

quality data from the Sea of Japan/East Sea were included with the GLODAPv2.2020 release; and (3) the Sea of Japan/East Sea is biogeochemically distinct from the open ocean to the east of Japan, providing a challenge for the predictive capabilities of the approaches. Neither of the earlier generation of algorithms work well there with large average biases and RMSE values that are approximately nine times greater on average than in the first set of regions considered, but with significant variance between properties and routines (Table 11). LIRv2 is especially problematic in this region, and the marked improvement in ESPER_LIR_validation relative to LIRv2 suggests the wider data inclusion windows did indeed reduce variance inflation in this region. The release versions of the ESPERs that do include data from the Sea of Japan/East Sea as training data indeed reproduce

Table 5. Assessment statistics, reported as bias (± RMSE) in μmol kg^{-1}, for various nitrate estimation routines presented both globally (top rows) and for the intermediate ocean where float-based sensor measurements are often checked against algorithm-based estimates (bottom rows).

Global					
	LIRv2	**ESPER_LIR**	**ESPER_NN**	**CANYON-B**	**Mixed**
N	**146,263**	**146,263**	**146,263**	**146,263**	**146,263**
Eq. 1	0.03 (± 0.52)	0.02 (± 0.48)	0.00 (± 0.42)	—	0.03 (± 0.49)
Eq. 2	0.01 (± 0.56)	0.00 (± 0.52)	−0.01 (± 0.47)	—	0.03 (± 0.49)
Eq. 3	0.04 (± 0.61)	0.01 (± 0.59)	0.00 (± 0.50)	—	0.03 (± 0.55)
Eq. 4	−0.02 (± 0.86)	−0.09 (± 0.82)	−0.07 (± 0.72)	—	0.00 (± 0.59)
Eq. 5	0.03 (± 0.54)	0.03 (± 0.49)	0.02 (± 0.43)	—	0.04 (± 0.50)
Eq. 6	0.00 (± 0.58)	−0.01 (± 0.55)	−0.01 (± 0.50)	—	0.03 (± 0.50)
Eq. 7	0.06 (± 0.72)	0.06 (± 0.70)	0.03 (± 0.56)	0.03 (± 0.56)	0.04 (± 0.59)
Eq. 8	−0.06 (± 1.26)	−0.04 (± 1.21)	−0.05 (± 1.04)	—	0.01 (± 0.73)
Eq. 9	0.03 (± 0.54)	0.02 (± 0.50)	0.01 (± 0.44)	—	0.04 (± 0.50)
Eq. 10	0.00 (± 0.58)	−0.01 (± 0.54)	−0.01 (± 0.49)	—	0.03 (± 0.50)
Eq. 11	0.05 (± 0.67)	0.02 (± 0.65)	0.00 (± 0.57)	—	0.03 (± 0.56)
Eq. 12	−0.08 (± 1.21)	−0.10 (± 0.89)	−0.06 (± 0.77)	—	0.00 (± 0.60)
Eq. 13	0.05 (± 0.57)	0.04 (± 0.52)	0.03 (± 0.48)	—	0.05 (± 0.52)
Eq. 14	0.00 (± 0.62)	0.00 (± 0.57)	−0.01 (± 0.53)	—	0.03 (± 0.51)
Eq. 15	0.12 (± 0.96)	0.11 (± 0.91)	0.08 (± 0.81)	—	0.07 (± 0.69)
Eq. 16	0.06 (± 2.22)	0.06 (± 2.00)	0.02 (± 1.83)	—	0.04 (± 1.08)
Intermediate depth only **(i.e., > 1000 m and < 1500 m depth)**					
N	**14,397**	**14,397**	**14,397**	**14,397**	**14,397**
Eq. 1	−0.01 (± 0.32)	−0.01 (± 0.31)	−0.01 (± 0.29)	—	0.01 (± 0.30)
Eq. 2	−0.04 (± 0.36)	−0.05 (± 0.34)	−0.04 (± 0.34)	—	−0.01 (± 0.30)
Eq. 3	0.03 (± 0.33)	0.02 (± 0.32)	0.02 (± 0.31)	—	0.02 (± 0.31)
Eq. 4	0.05 (± 0.45)	0.01 (± 0.40)	−0.01 (± 0.44)	—	0.00 (± 0.34)
Eq. 5	−0.01 (± 0.33)	−0.02 (± 0.32)	0.00 (± 0.30)	—	0.01 (± 0.30)
Eq. 6	−0.05 (± 0.38)	−0.08 (± 0.38)	−0.07 (± 0.38)	—	−0.02 (± 0.31)
Eq. 7	0.04 (± 0.34)	0.02 (± 0.33)	0.04 (± 0.33)	0.04 (± 0.33)	0.03 (± 0.32)
Eq. 8	0.05 (± 0.54)	−0.05 (± 0.53)	−0.01 (± 0.58)	—	0.01 (± 0.40)
Eq. 9	−0.01 (± 0.32)	−0.02 (± 0.32)	0.00 (± 0.30)	—	0.01 (± 0.30)
Eq. 10	−0.05 (± 0.37)	−0.07 (± 0.37)	−0.07 (± 0.34)	—	−0.02 (± 0.30)
Eq. 11	0.03 (± 0.34)	0.02 (± 0.32)	0.03 (± 0.32)	—	0.02 (± 0.31)
Eq. 12	0.04 (± 0.55)	−0.03 (± 0.45)	−0.02 (± 0.46)	—	0.00 (± 0.35)
Eq. 13	−0.01 (± 0.34)	−0.02 (± 0.33)	−0.01 (± 0.32)	—	0.01 (± 0.31)
Eq. 14	−0.06 (± 0.40)	−0.10 (± 0.39)	−0.07 (± 0.40)	—	−0.03 (± 0.32)
Eq. 15	0.05 (± 0.37)	0.02 (± 0.34)	0.03 (± 0.36)	—	0.03 (± 0.33)
Eq. 16	0.06 (± 0.73)	−0.09 (± 0.65)	−0.06 (± 0.71)	—	−0.02 (± 0.45)

these data with comparable fidelity to the global statistics (Supporting Information S1.4). We conclude this region is not a special challenge for algorithms when training data are included. The release versions of these algorithms updated with the new data should therefore work in the now-measured portions of the Sea of Japan/East Sea.

Two additional marginal seas deserve mention. GLODAPv2 does not yet include data from the Gulf of Mexico or the Mediterranean Sea that have been subjected to the GLODAPv2 secondary quality control process (some data from the Mediterranean Sea are included, but with QC flags of 0). However, due to the large errors expected within marginal seas (and now demonstrated for the Sea of Japan) when training data are absent or omitted, data from two cruises to the Mediterranean were included in the training data for CANYON-B despite the lack of secondary QC. We now do similarly in the

Table 6. Assessment statistics, reported as bias (± RMSE) in μmol kg^{-1}, for various silicate estimation routines presented both globally (top rows) and for the intermediate ocean (bottom rows, provided for comparison only as there are no float-based sensors for phosphate that are calibrated using algorithms).

Global

	LIRv2	ESPER_LIR	ESPER_NN	CANYON-B	Mixed
N	146,263	146,263	146,263	146,263	146,263
Eq. 1	−0.3 (± 2.4)	0.0 (± 2.2)	0.0 (± 1.8)	—	0.1 (± 1.9)
Eq. 2	−0.3 (± 2.5)	−0.1 (± 2.5)	−0.1 (± 2.1)	—	0.0 (± 2.0)
Eq. 3	−0.2 (± 2.4)	0.0 (± 2.2)	0.1 (± 2.0)	—	0.1 (± 2.0)
Eq. 4	−0.3 (± 2.6)	−0.1 (± 2.5)	0.0 (± 2.0)	—	0.1 (± 2.0)
Eq. 5	−0.2 (± 2.4)	0.0 (± 2.3)	0.1 (± 1.8)	—	0.1 (± 1.9)
Eq. 6	−0.3 (± 2.7)	−0.2 (± 2.6)	−0.1 (± 2.1)	—	0.0 (± 2.0)
Eq. 7	−0.2 (± 2.7)	0.1 (± 2.3)	0.1 (± 2.0)	0.1 (± 1.9)	0.1 (± 2.0)
Eq. 8	−0.3 (± 3.6)	−0.1 (± 3.3)	−0.1 (± 2.7)	—	0.0 (± 2.2)
Eq. 9	0.0 (± 4.1)	0.1 (± 3.0)	0.1 (± 2.6)	—	0.1 (± 2.2)
Eq. 10	−0.1 (± 5.0)	0.1 (± 3.1)	0.0 (± 2.6)	—	0.0 (± 2.2)
Eq. 11	0.0 (± 4.3)	0.1 (± 3.0)	0.1 (± 2.6)	—	0.1 (± 2.1)
Eq. 12	0.0 (± 4.9)	0.1 (± 3.1)	0.0 (± 2.7)	—	0.1 (± 2.2)
Eq. 13	0.1 (± 4.6)	0.1 (± 3.2)	0.1 (± 2.7)	—	0.1 (± 2.2)
Eq. 14	−0.1 (± 5.2)	0.0 (± 3.3)	−0.1 (± 2.8)	—	0.0 (± 2.2)
Eq. 15	0.3 (± 5.5)	0.3 (± 3.4)	0.2 (± 3.2)	—	0.2 (± 2.4)
Eq. 16	0.4 (± 6.9)	0.1 (± 5.4)	−0.1 (± 5.3)	—	0.0 (± 3.2)

Intermediate depth only **(i.e., > 1000 m and < 1500 m depth)**

N	14,397	14,397	14,397	14,397	14,397
Eq. 1	−0.3 (± 2.0)	−0.2 (± 1.7)	−0.1 (± 1.6)	—	−0.1 (± 1.5)
Eq. 2	−0.3 (± 2.1)	−0.3 (± 2.1)	−0.2 (± 2.0)	—	−0.2 (± 1.6)
Eq. 3	−0.3 (± 2.0)	−0.1 (± 1.6)	−0.1 (± 1.7)	—	−0.1 (± 1.5)
Eq. 4	−0.3 (± 2.1)	−0.2 (± 1.9)	−0.1 (± 1.9)	—	−0.1 (± 1.6)
Eq. 5	−0.3 (± 2.1)	−0.2 (± 1.8)	−0.1 (± 1.6)	—	−0.1 (± 1.5)
Eq. 6	−0.3 (± 2.3)	−0.5 (± 2.4)	−0.3 (± 2.0)	—	−0.2 (± 1.6)
Eq. 7	−0.3 (± 2.1)	−0.1 (± 1.6)	−0.2 (± 1.7)	0.0 (± 1.5)	−0.2 (± 1.5)
Eq. 8	−0.1 (± 2.7)	−0.3 (± 2.6)	−0.1 (± 2.4)	—	−0.1 (± 1.7)
Eq. 9	0.0 (± 3.4)	−0.1 (± 3.3)	−0.2 (± 3.3)	—	−0.2 (± 2.2)
Eq. 10	0.0 (± 5.7)	−0.1 (± 3.2)	−0.2 (± 3.3)	—	−0.2 (± 2.1)
Eq. 11	0.0 (± 3.7)	−0.1 (± 2.9)	−0.1 (± 3.4)	—	−0.1 (± 2.2)
Eq. 12	0.1 (± 5.5)	0.0 (± 3.0)	0.0 (± 3.3)	—	−0.1 (± 2.1)
Eq. 13	0.0 (± 4.1)	−0.1 (± 3.7)	−0.3 (± 3.4)	—	−0.2 (± 2.2)
Eq. 14	−0.1 (± 6.4)	−0.4 (± 3.4)	−0.4 (± 3.6)	—	−0.3 (± 2.3)
Eq. 15	0.1 (± 5.3)	0.0 (± 3.2)	−0.1 (± 3.8)	—	−0.1 (± 2.3)
Eq. 16	0.2 (± 6.1)	−0.4 (± 4.0)	−0.1 (± 4.7)	—	−0.1 (± 2.7)

ESPERs and include additional data gathered as part of the CODAP-NA (Jiang et al. 2021) and ongoing CARIMED efforts (Supporting Information S1.1). The same lessons from the Sea of Japan/East Sea analysis apply to the reconstruction of measurements from the Gulf of Mexico and the Mediterranean Sea (Table 11). We caution that ESPER_LIR is challenged by the lack of data below 2000 m depth in the Mediterranean and increases its window sizes large enough to incorporate data at depth from the deep North Atlantic. This results in poor RMSE statistics even when the test data are included with the training data (Supporting Information S1.4). Until this is addressed, it is recommended that users interested in this area use ESPER_NN or CANYON_MED (Fourrier et al. 2020) in place of ESPER_LIR or ESPER_Mixed. Such regional algorithms

Table 7. Assessment statistics, reported as bias (± RMSE) in μmol kg^{-1}, for various oxygen estimation routines presented both globally (top rows) and for the intermediate ocean (bottom rows, provided for comparison only as float-based oxygen sensors are not commonly quality controlled against algorithms).

Global					
	LIRv2	**ESPER_LIR**	**ESPER_NN**	**CANYON-B***	**Mixed**
N	**146,263**	**146,263**	**146,263**	**—***	**146,263**
Eq. 1	0.5 (± 5.3)	0.6 (± 5.2)	0.5 (± 4.5)	—	0.6 (± 4.7)
Eq. 2	0.4 (± 5.7)	0.5 (± 5.6)	0.5 (± 5.0)	—	0.6 (± 4.8)
Eq. 3	0.5 (± 5.8)	0.6 (± 5.5)	0.6 (± 4.8)	—	0.6 (± 4.9)
Eq. 4	0.7 (± 8.0)	1.3 (± 7.6)	1.0 (± 7.1)	—	0.8 (± 5.6)
Eq. 5	0.6 (± 5.5)	0.8 (± 5.4)	0.7 (± 4.7)	—	0.7 (± 4.8)
Eq. 6	0.7 (± 5.9)	0.8 (± 5.8)	0.6 (± 5.3)	—	0.7 (± 4.8)
Eq. 7	0.6 (± 6.2)	0.7 (± 5.6)	0.5 (± 5.0)	—	0.6 (± 5.0)
Eq. 8	1.1 (± 10.8)	1.2 (± 10.0)	1.1 (± 9.7)	—	0.9 (± 6.6)
Eq. 9	1.1 (± 8.1)	1.0 (± 7.9)	1.1 (± 7.0)	—	0.9 (± 5.6)
Eq. 10	1.1 (± 8.8)	1.0 (± 8.3)	1.1 (± 7.6)	—	0.9 (± 5.7)
Eq. 11	1.1 (± 8.4)	1.0 (± 8.0)	1.0 (± 7.4)	—	0.9 (± 5.8)
Eq. 12	2.0 (± 14.2)	1.7 (± 9.9)	1.4 (± 9.5)	—	1.1 (± 6.5)
Eq. 13	1.4 (± 9.8)	1.3 (± 8.2)	1.1 (± 7.3)	—	0.9 (± 5.7)
Eq. 14	1.5 (± 10.4)	1.3 (± 8.4)	1.2 (± 7.7)	—	1.0 (± 5.8)
Eq. 15	1.4 (± 9.8)	1.2 (± 8.2)	1.0 (± 7.6)	—	0.9 (± 5.9)
Eq. 16	1.6 (± 18.6)	1.2 (± 13.7)	0.8 (± 13.1)	—	0.8 (± 7.9)
Intermediate depth only **(i.e., >1000 m and < 1500 m depth)**					
N	**14,397**	**14,397**	**14,397**	**—***	**14,397**
Eq. 1	0.2 (± 2.8)	0.4 (± 2.6)	0.6 (± 2.7)	—	0.5 (± 2.6)
Eq. 2	0.4 (± 3.4)	0.7 (± 2.9)	0.8 (± 3.1)	—	0.6 (± 2.6)
Eq. 3	0.0 (± 3.0)	0.2 (± 2.6)	0.1 (± 2.8)	—	0.3 (± 2.6)
Eq. 4	−0.4 (± 4.3)	0.2 (± 3.3)	0.1 (± 4.2)	—	0.3 (± 2.9)
Eq. 5	0.4 (± 3.0)	0.6 (± 2.9)	0.8 (± 3.1)	—	0.6 (± 2.8)
Eq. 6	0.6 (± 3.8)	1.1 (± 3.5)	1.1 (± 3.9)	—	0.8 (± 3.0)
Eq. 7	0.0 (± 3.2)	0.4 (± 2.9)	0.4 (± 3.1)	—	0.4 (± 2.8)
Eq. 8	−0.3 (± 5.1)	0.8 (± 4.8)	0.2 (± 5.9)	—	0.3 (± 3.7)
Eq. 9	0.4 (± 3.8)	0.8 (± 3.9)	1.0 (± 3.6)	—	0.7 (± 3.0)
Eq. 10	0.7 (± 4.2)	1.2 (± 4.7)	1.2 (± 4.1)	—	0.8 (± 3.0)
Eq. 11	0.2 (± 3.9)	0.6 (± 3.8)	0.7 (± 4.0)	—	0.6 (± 3.1)
Eq. 12	−0.2 (± 6.1)	0.7 (± 4.8)	0.4 (± 5.4)	—	0.4 (± 3.4)
Eq. 13	0.7 (± 5.4)	1.0 (± 4.0)	0.8 (± 4.0)	—	0.6 (± 3.1)
Eq. 14	1.1 (± 5.7)	1.5 (± 4.5)	1.2 (± 4.4)	—	0.8 (± 3.1)
Eq. 15	0.4 (± 5.5)	0.8 (± 4.0)	0.6 (± 4.3)	—	0.5 (± 3.2)
Eq. 16	0.0 (± 7.6)	1.4 (± 6.2)	0.2 (± 6.0)	—	0.3 (± 3.7)

*This routine does not estimate this quantity.

can be meaningfully better for regional efforts, and work in progress on a regional algorithm for the Gulf of Mexico shows promise for reducing the RMS misfit to the observations from this region. The Gulf of Mexico challenges the ESPERs because this is a region where the underlying TTD-based C_{ant} data product does not contain estimates, so C_{ant} is crudely triangulated between the Pacific and Atlantic in this region. A regional algorithm could address this limitation with a more sophisticated approach.

Finally, with intense seasonality, strong freshwater cycling and riverine inputs, seasonal ice cover, and broad continental shelves, the Arctic is an interesting "worst case scenario" for

Table 8. Assessment statistics, reported as bias (± RMSE) in μmol kg^{-1}, for various TA estimation routines presented both globally (top rows) and for the intermediate ocean (bottom rows, provided for comparison only as TA sensors have yet to be widely deployed on floats).

Global					
	LIRv2	**ESPER_LIR**	**ESPER_NN**	**CANYON-B**	**Mixed**
N	**71,832**	**71,832**	**71,832**	**71,832**	**71,832**
Eq. 1	0.8 (± 3.6)	0.8 (± 3.6)	0.8 (± 3.7)	—	0.8 (± 3.5)
Eq. 2	0.7 (± 3.6)	0.8 (± 3.6)	0.8 (± 3.7)	—	0.8 (± 3.5)
Eq. 3	0.7 (± 3.7)	0.8 (± 3.6)	0.8 (± 3.7)	—	0.7 (± 3.5)
Eq. 4	0.7 (± 3.7)	0.9 (± 3.6)	0.9 (± 3.8)	—	0.8 (± 3.6)
Eq. 5	0.5 (± 3.9)	0.6 (± 3.7)	0.7 (± 3.7)	—	0.7 (± 3.6)
Eq. 6	0.4 (± 4.0)	0.5 (± 3.8)	0.7 (± 3.9)	—	0.7 (± 3.6)
Eq. 7	0.5 (± 4.0)	0.7 (± 3.7)	0.8 (± 3.8)	0.4 (± 4.2)	0.7 (± 3.6)
Eq. 8	0.5 (± 4.3)	0.6 (± 4.0)	0.8 (± 4.1)	—	0.7 (± 3.7)
Eq. 9	0.7 (± 3.7)	0.8 (± 3.7)	0.9 (± 3.7)	—	0.8 (± 3.5)
Eq. 10	0.7 (± 3.7)	0.9 (± 3.7)	0.9 (± 3.7)	—	0.8 (± 3.5)
Eq. 11	0.7 (± 3.7)	0.9 (± 3.7)	0.9 (± 3.6)	—	0.8 (± 3.5)
Eq. 12	0.8 (± 3.9)	1.0 (± 3.7)	0.9 (± 3.7)	—	0.8 (± 3.5)
Eq. 13	0.7 (± 4.4)	0.8 (± 3.9)	0.8 (± 4.0)	—	0.7 (± 3.6)
Eq. 14	0.7 (± 4.9)	0.7 (± 4.1)	0.8 (± 4.0)	—	0.7 (± 3.6)
Eq. 15	1.0 (± 4.8)	0.9 (± 4.0)	0.9 (± 4.0)	—	0.8 (± 3.6)
Eq. 16	1.2 (± 6.5)	0.9 (± 5.0)	0.7 (± 5.2)	—	0.7 (± 4.0)
Intermediate depth only **(i.e., >1000 m and < 1500 m depth)**					
N	***6797***	***6797***	***6797***	***6797***	***6797***
Eq. 1	0.9 (± 3.0)	0.8 (± 2.9)	1.0 (± 3.0)	—	0.8 (± 2.8)
Eq. 2	0.9 (± 2.9)	0.8 (± 2.9)	0.9 (± 2.9)	—	0.8 (± 2.8)
Eq. 3	0.9 (± 2.9)	0.8 (± 2.9)	0.9 (± 3.0)	—	0.8 (± 2.8)
Eq. 4	0.8 (± 3.0)	0.8 (± 2.9)	0.9 (± 3.0)	—	0.8 (± 2.9)
Eq. 5	0.6 (± 3.2)	0.6 (± 2.9)	0.7 (± 3.1)	—	0.7 (± 2.9)
Eq. 6	0.6 (± 3.2)	0.5 (± 2.9)	0.8 (± 3.2)	—	0.7 (± 2.9)
Eq. 7	0.6 (± 3.2)	0.6 (± 2.9)	0.7 (± 3.1)	0.5 (± 3.2)	0.7 (± 2.9)
Eq. 8	0.7 (± 3.2)	0.6 (± 2.9)	0.8 (± 3.3)	—	0.7 (± 3.0)
Eq. 9	0.9 (± 3.0)	0.9 (± 2.9)	0.9 (± 3.0)	—	0.8 (± 2.9)
Eq. 10	0.8 (± 3.0)	0.8 (± 2.9)	1.0 (± 3.1)	—	0.8 (± 2.9)
Eq. 11	0.9 (± 3.0)	0.9 (± 2.9)	1.0 (± 3.0)	—	0.8 (± 2.9)
Eq. 12	0.8 (± 3.0)	0.9 (± 2.9)	1.0 (± 3.1)	—	0.8 (± 2.9)
Eq. 13	0.7 (± 3.8)	0.6 (± 3.2)	0.6 (± 3.6)	—	0.6 (± 3.1)
Eq. 14	0.7 (± 4.1)	0.5 (± 3.2)	0.6 (± 3.6)	—	0.6 (± 3.1)
Eq. 15	0.8 (± 3.8)	0.6 (± 3.2)	0.8 (± 3.7)	—	0.7 (± 3.1)
Eq. 16	0.9 (± 4.4)	0.6 (± 3.4)	0.7 (± 4.5)	—	0.6 (± 3.3)

the algorithms, even when training data are available. The validation statistics in this region are significantly worse than the global statistics (RMSEs average ~ 2.3 times greater, though again with variance between properties and routines, Table 11). These larger uncertainties found in the Arctic could perhaps be generalized to other problematic regions such as shallow coastal areas, small marginal seas, areas with significant riverine inputs, or other areas with seasonal ice cover.

Mixed ESPER

As proposed by Bittig et al. (2018), averaging the estimates from ESPER_LIR_validation and ESPER_NN_validation indeed seems to improve the global average prediction statistics, though the improvement is sometimes small and often the individual residuals are greater with the ESPER_Mixed estimate than for the better of the two estimates. For equations with few predictors (e.g., ESPER Eq. 16, using *S* as the only seawater

Table 9. Assessment statistics, reported as bias (± RMSE), for various pH estimation routines presented both globally (top rows) and for the intermediate ocean where float-based sensor measurements are often checked against algorithm-based estimates (bottom rows). Only measurements made with purified dyes were used in these assessments to ensure the validation data had no adjustments beyond those applied in the GLODAPv2.2020 secondary quality control process.

Global

	LIRv2	ESPER_LIR	ESPER_NN	CANYON-B	Mixed
N	20,181	20,181	20,181	20,181	20,181
Eq. 1	−0.007 (± 0.012)	−0.004 (± 0.013)	−0.004 (± 0.011)	—	−0.004 (± 0.011)
Eq. 2	−0.006 (± 0.015)	−0.002 (± 0.014)	−0.002 (± 0.013)	—	−0.003 (± 0.011)
Eq. 3	−0.007 (± 0.013)	−0.004 (± 0.013)	−0.004 (± 0.011)	—	−0.004 (± 0.011)
Eq. 4	−0.005 (± 0.022)	−0.001 (± 0.017)	−0.002 (± 0.016)	—	−0.003 (± 0.012)
Eq. 5	−0.007 (± 0.012)	−0.004 (± 0.012)	−0.004 (± 0.011)	—	−0.004 (± 0.011)
Eq. 6	−0.005 (± 0.015)	−0.001 (± 0.014)	−0.002 (± 0.014)	—	−0.003 (± 0.011)
Eq. 7	−0.007 (± 0.013)	−0.004 (± 0.013)	−0.004 (± 0.011)	—*	−0.004 (± 0.011)
Eq. 8	−0.005 (± 0.026)	0.000 (± 0.020)	0.000 (± 0.021)	—	−0.002 (± 0.014)
Eq. 9	−0.007 (± 0.013)	−0.004 (± 0.014)	−0.003 (± 0.012)	—	−0.004 (± 0.011)
Eq. 10	−0.005 (± 0.016)	−0.002 (± 0.015)	−0.001 (± 0.014)	—	−0.003 (± 0.012)
Eq. 11	−0.007 (± 0.013)	−0.004 (± 0.014)	−0.003 (± 0.012)	—	−0.004 (± 0.011)
Eq. 12	−0.004 (± 0.023)	−0.001 (± 0.018)	−0.001 (± 0.018)	—	−0.003 (± 0.013)
Eq. 13	−0.006 (± 0.013)	−0.004 (± 0.013)	−0.003 (± 0.012)	—	−0.004 (± 0.011)
Eq. 14	−0.004 (± 0.017)	−0.001 (± 0.015)	−0.001 (± 0.014)	—	−0.003 (± 0.012)
Eq. 15	−0.006 (± 0.013)	−0.004 (± 0.014)	−0.004 (± 0.012)	—	−0.004 (± 0.012)
Eq. 16	−0.005 (± 0.033)	−0.001 (± 0.026)	−0.002 (± 0.027)	—	−0.003 (± 0.017)

Intermediate depth only **(i.e., >1000 m and < 1500 m depth)**

N	2352	2352	2352	2352	2352
Eq. 1	−0.008 (± 0.011)	−0.002 (± 0.008)	−0.002 (± 0.007)	—	−0.002 (± 0.006)
Eq. 2	−0.007 (± 0.013)	−0.001 (± 0.008)	−0.001 (± 0.008)	—	−0.001 (± 0.006)
Eq. 3	−0.008 (± 0.011)	−0.002 (± 0.007)	−0.001 (± 0.006)	—	−0.001 (± 0.006)
Eq. 4	−0.008 (± 0.024)	−0.001 (± 0.009)	−0.002 (± 0.011)	—	−0.002 (± 0.008)
Eq. 5	−0.008 (± 0.011)	−0.002 (± 0.007)	−0.002 (± 0.007)	—	−0.002 (± 0.006)
Eq. 6	−0.007 (± 0.013)	0.001 (± 0.008)	0.000 (± 0.007)	—	−0.001 (± 0.006)
Eq. 7	−0.008 (± 0.011)	−0.002 (± 0.007)	−0.002 (± 0.007)	—*	−0.002 (± 0.006)
Eq. 8	−0.008 (± 0.024)	0.001 (± 0.009)	0.000 (± 0.014)	—	−0.001 (± 0.008)
Eq. 9	−0.007 (± 0.011)	−0.002 (± 0.008)	−0.001 (± 0.006)	—	−0.002 (± 0.006)
Eq. 10	−0.007 (± 0.013)	0.001 (± 0.008)	0.000 (± 0.008)	—	−0.001 (± 0.006)
Eq. 11	−0.007 (± 0.011)	−0.002 (± 0.007)	−0.001 (± 0.007)	—	−0.002 (± 0.006)
Eq. 12	−0.008 (± 0.024)	0.000 (± 0.010)	0.000 (± 0.013)	—	−0.001 (± 0.008)
Eq. 13	−0.007 (± 0.011)	−0.002 (± 0.007)	−0.002 (± 0.007)	—	−0.002 (± 0.007)
Eq. 14	−0.007 (± 0.014)	0.001 (± 0.007)	0.000 (± 0.008)	—	−0.001 (± 0.006)
Eq. 15	−0.007 (± 0.011)	−0.001 (± 0.007)	−0.001 (± 0.007)	—	−0.001 (± 0.006)
Eq. 16	−0.008 (± 0.028)	0.002 (± 0.010)	0.001 (± 0.015)	—	−0.001 (± 0.009)

*No viable comparison in this effort due to overlap between training and validation data subsets.

property predictor) the improvement in the global open-ocean average RMSE is pronounced for all seven properties estimated by the routines. We therefore recommend using ESPER_Mixed over ESPER_LIR or ESPER_NN unless there is reason to prefer one approach over another due to, for example, the results of a regional validation exercise in the region of interest.

Discussion and summary statements

Several patterns hold across the various properties. For example, including more predictors leads to better estimates on average (Fig. 4, showing an average across all properties for both ESPERs) when the predictor measurements are high quality (i.e., comparable to the measurements in GLODAPv2).

Table 10. Assessment statistics, reported as bias (± RMSE) in μmol kg^{-1}, for various DIC estimation routines presented both globally (top rows) and for the intermediate ocean (bottom rows, provided for comparison only as DIC sensors have yet to be widely deployed on floats).

Global					
	LIRv2*	**ESPER_LIR**	**ESPER_NN**	**CANYON-B**	**Mixed**
N	**—***	**71,326**	**71,326**	**71,326**	**71,326**
Eq. 1	—	0.4 (± 5.1)	0.4 (± 4.9)	—	0.4 (± 4.8)
Eq. 2	—	0.2 (± 5.8)	0.4 (± 5.7)	—	0.4 (± 4.9)
Eq. 3	—	0.3 (± 4.9)	0.4 (± 4.8)	—	0.4 (± 4.8)
Eq. 4	—	−0.2 (± 6.6)	0.0 (± 6.6)	—	0.2 (± 5.2)
Eq. 5	—	0.3 (± 5.1)	0.4 (± 5.1)	—	0.4 (± 4.9)
Eq. 6	—	0.0 (± 6.1)	0.3 (± 6.4)	—	0.3 (± 5.2)
Eq. 7	—	0.4 (± 5.3)	0.4 (± 5.1)	−1.3 (± 5.8)	0.4 (± 5.0)
Eq. 8	—	−0.4 (± 8.7)	−0.1 (± 8.6)	—	0.1 (± 6.0)
Eq. 9	—	0.6 (± 8.2)	0.6 (± 6.9)	—	0.5 (± 5.3)
Eq. 10	—	0.3 (± 9.0)	0.4 (± 7.3)	—	0.4 (± 5.3)
Eq. 11	—	0.5 (± 7.4)	0.6 (± 6.7)	—	0.5 (± 5.3)
Eq. 12	—	−0.2 (± 9.3)	0.1 (± 8.5)	—	0.3 (± 5.7)
Eq. 13	—	0.6 (± 7.9)	0.7 (± 7.3)	—	0.5 (± 5.5)
Eq. 14	—	0.1 (± 8.7)	0.3 (± 8.0)	—	0.3 (± 5.6)
Eq. 15	—	0.8 (± 8.9)	0.8 (± 8.4)	—	0.6 (± 6.1)
Eq. 16	—	0.6 (± 16.7)	0.3 (± 15.7)	—	0.4 (± 8.9)
Intermediate depth only **(i.e., >1000 m and < 1500 m depth)**					
N	**—***	***6740***	***6740***	***6740***	***6740***
Eq. 1	—	−0.2 (± 3.3)	−0.1 (± 3.3)	—	−0.2 (± 3.3)
Eq. 2	—	−0.3 (± 3.5)	0.0 (± 3.7)	—	−0.1 (± 3.3)
Eq. 3	—	−0.2 (± 3.3)	0.1 (± 3.2)	—	−0.1 (± 3.2)
Eq. 4	—	−0.1 (± 3.8)	0.0 (± 4.3)	—	−0.1 (± 3.5)
Eq. 5	—	−0.2 (± 3.3)	−0.1 (± 3.4)	—	−0.2 (± 3.3)
Eq. 6	—	−0.5 (± 3.7)	−0.2 (± 4.1)	—	−0.2 (± 3.5)
Eq. 7	—	−0.2 (± 3.3)	−0.2 (± 3.5)	−0.8 (± 3.4)	−0.2 (± 3.3)
Eq. 8	—	−0.5 (± 4.5)	−0.4 (± 5.4)	—	−0.3 (± 3.9)
Eq. 9	—	0.0 (± 3.4)	0.1 (± 3.3)	—	−0.1 (± 3.2)
Eq. 10	—	−0.2 (± 3.5)	−0.1 (± 3.7)	—	−0.2 (± 3.3)
Eq. 11	—	−0.1 (± 3.4)	0.1 (± 3.3)	—	−0.1 (± 3.2)
Eq. 12	—	−0.2 (± 3.8)	0.0 (± 4.4)	—	−0.1 (± 3.5)
Eq. 13	—	−0.2 (± 3.7)	−0.1 (± 4.0)	—	−0.2 (± 3.5)
Eq. 14	—	−0.4 (± 4.0)	−0.5 (± 4.5)	—	−0.4 (± 3.6)
Eq. 15	—	−0.1 (± 3.8)	0.0 (± 4.2)	—	−0.1 (± 3.5)
Eq. 16	—	−0.7 (± 5.7)	−0.3 (± 6.8)	—	−0.3 (± 4.4)

*This routine does not estimate this quantity.

However, estimate improvements are marginal beyond four predictors. Also, ESPER Eqs. 6 and 7 do nearly as well as any equation despite having only three predictors (i.e., temperature; salinity; and either oxygen, nitrate, or phosphate, depending on the predicted property). This observation shows the predictive power of including at least one macronutrient or oxygen as a predictor for biogeochemical properties.

A second important generalization is that all predictions do better at depth (> 1000 m) though this is especially the case for gas distribution reconstructions: the intermediate-depth RMSE values average 55% of the global RMSE values for oxygen, pH_T, and DIC (Tables 7, 9, and 10, respectively) whereas they average ~ 70% of the global RMSE values for phosphate, nitrate, silicate, and TA (Tables 4–6 and 8, respectively). The

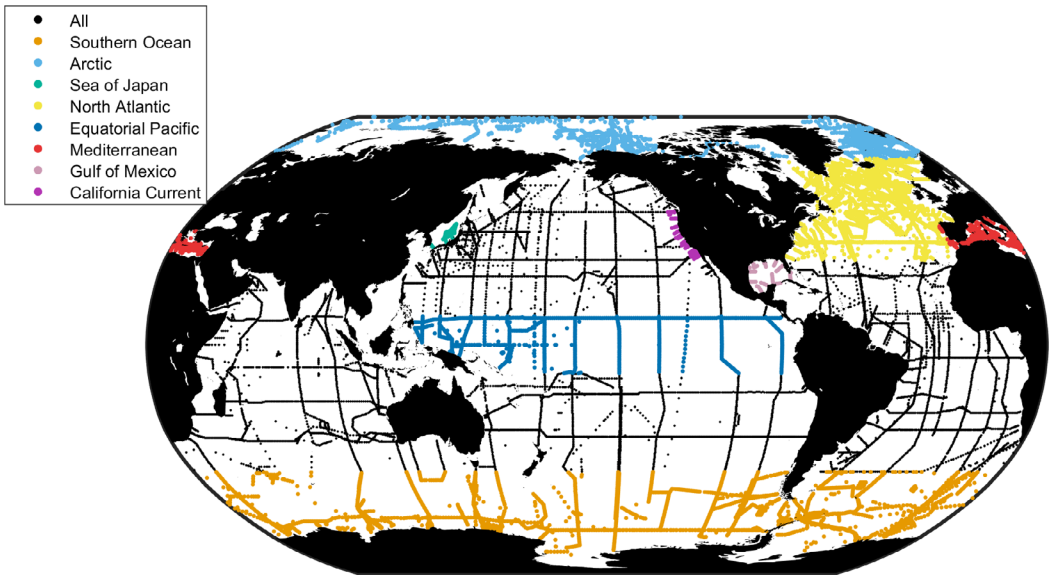

Fig 3. A map showing the regions considered independently in "Regional Tests" section.

larger, near surface estimate errors for parameters influenced by air–sea gas exchange (e.g., pH_T, DIC, and oxygen) are likely the result of their decoupling with predictor variables that are not gases (or are gases with different equilibration and residence times). These changes in parameter relationships near the surface due to air–sea exchange are also sensitive to dynamic processes (e.g., wind speed), which are not well captured by the predictor parameters, and are thus difficult to parameterize in static algorithm relationships.

Finally, regional errors are sometimes significantly larger than global open-ocean errors, and regional biases are almost always larger than the global biases. This highlights an important caution for users of these routines: the global statistics may not be appropriate for estimates over a more limited area. For this, we note both that it is important to validate the algorithm estimates for a given region/application and to consider how large of an average estimate bias is likely for a region of a given size. As an example, we have assessed how the bias decreases as the size of the latitude and longitude window considered increases for ESPER_NN_validation nitrate estimates (Fig. 5). These average regional biases are computed by iteratively averaging all estimate errors inside windows of a given size around each of the grid points used by the LIR routines. Then, for each window size considered we compute an area-weighted average of the absolute values of the bias estimates for the grid points. In the example presented, the average estimate bias is approximately half of the global RMSE when estimates are averaged over a $10^\circ \times 10^\circ$ window, and as expected the bias becomes smaller as the averaging window grows. This shows that the estimates retain significant regional bias, implying nearby algorithm estimates cannot be treated as statistically independent. For a float or mooring that stays within a small spatial region, this algorithm bias could be somewhat worse still than shown in Fig. 5. For pCO_2 calculations based on pH_T measurements that are adjusted to algorithm values, even a small average bias could lead to a meaningful change in calculated air–sea CO_2 flux.

Comments and recommendations

We have updated global algorithms for seawater biogeochemical property estimation and their associated MATLAB routines with new functionality using new methods and new data. We show that our new methods are mechanistically at least as skillful as earlier methods and are in some cases better. They also have the advantages of being trained with the latest quality-controlled data products, easy to implement in MATLAB, capable of estimating a variety of seawater properties, flexible with the choice of input parameters, and capable of adapting several aspects of their outputs to user needs (e.g., calculated-like or measured-like pH_T). Where possible, our validation statistics provide comparisons using validation versions of the algorithms with identical training and validation data sets for all versions of the routines assessed. We therefore recommend these updates even when validation metrics are comparable to those of earlier routines because the newer routines are trained from a larger data set with better temporal and spatial coverage. Two important features of our new routines are (1) the flexibility to predict many seawater properties from 16 combinations of seawater properties using either a regression approach or a neural network approach and (2) the implementation of a simple estimate of the impacts of C_{ant} on pH_T and DIC based on first principles. While the new C_{ant} estimation strategy is an improvement

Table 11. Regional assessment statistics for ESPER Eq. 7 of the validation versions of the algorithms and for CANYON-B. These statistics are obtained without including any training data from the new data added in the 2019 and 2020 GLODAPv2 data product updates; without the supplemental data in the Gulf of Mexico; and, in the case of LIRv2, ESPER_LIR, and ESPER_NN, without any measurements in the Mediterranean. The released ESPER_LIR and ESPER_NN routines should perform significantly better in the Sea of Japan/East Sea, the Gulf of Mexico, and the Mediterranean. Statistics obtained when these data are included are provided as Supporting Information.

Southern Ocean	**Phosphate**	**Nitrate**	**Silicate**	**Oxygen**	**TA**	**pH**	**DIC**
N	**20,294**	**20,294**	**20,294**	**20,294**	**11,088**	**4094**	**11,945**
LIRv2	0.000 (± 0.059)	−0.03 (± 0.77)	−0.4 (± 5.1)	0.0 (± 6.3)	−0.3 (± 3.3)	−0.001 (± 0.011)	—
ESPER_LIR	−0.004 (± 0.062)	−0.07 (± 0.76)	0.1 (± 4.8)	0.0 (± 6.3)	0.3 (± 3.0)	−0.001 (± 0.013)	1.4 (± 4.7)
ESPER_NN	−0.003 (± 0.054)	−0.03 (± 0.69)	0.0 (± 3.9)	0.6 (± 6.1)	0.7 (± 3.1)	−0.002 (± 0.010)	1.6 (± 4.6)
CANYON-B	−0.001 (± 0.055)	−0.04 (± 0.65)	0.1 (± 3.7)	—*	−0.4 (± 3.1)	−0.002 (± 0.009)	−0.8 (± 4.3)
ESPER_Mixed	−0.003 (± 0.057)	−0.05 (± 0.71)	0.1 (± 4.1)	0.3 (± 5.8)	0.5 (± 2.9)	−0.001 (± 0.011)	1.5 (± 4.6)

Equatorial Pacific	**Phosphate**	**Nitrate**	**Silicate**	**Oxygen**	**TA**	**pH**	**DIC**
N	**23,169**	**23,169**	**23,169**	**23,169**	**8661**	**1739**	**8969**
LIRv2	−0.003 (± 0.038)	0.04 (± 0.54)	0.1 (± 1.2)	0.7 (± 4.6)	0.8 (± 3.5)	−0.012 (± 0.016)	—
ESPER_LIR	−0.002 (± 0.041)	0.09 (± 0.56)	0.3 (± 1.4)	1.0 (± 4.7)	0.9 (± 3.3)	−0.007 (± 0.017)	−0.8 (± 5.1)
ESPER_NN	−0.003 (± 0.033)	0.05 (± 0.37)	0.3 (± 1.3)	0.2 (± 3.9)	1.0 (± 3.4)	−0.007 (± 0.014)	−0.5 (± 5.2)
CANYON-B	−0.003 (± 0.033)	0.04 (± 0.38)	0.2 (± 1.2)	—*	0.1 (± 4.4)	−0.004 (± 0.011)	−1.3 (± 5.1)
ESPER_Mixed	−0.003 (± 0.034)	0.07 (± 0.43)	0.3 (± 1.3)	0.6 (± 3.9)	1.0 (± 3.2)	−0.007 (± 0.014)	−0.6 (± 5.0)

California current	**Phosphate**	**Nitrate**	**Silicate**	**Oxygen**	**TA**	**pH**	**DIC**
N	**466**	**466**	**466**	**466**	**283**	**191**	**276**
LIRv2	−0.012 (± 0.049)	0.02 (± 0.79)	−0.8 (± 3.3)	0.4 (± 9.0)	2.2 (± 3.8)	−0.008 (± 0.012)	—
ESPER_LIR	−0.004 (± 0.046)	0.00 (± 0.75)	−0.2 (± 2.4)	0.6 (± 8.2)	2.3 (± 4.9)	−0.007 (± 0.015)	−0.3 (± 4.5)
ESPER_NN	0.002 (± 0.044)	−0.02 (± 0.55)	0.7 (± 1.7)	0.5 (± 5.6)	3.0 (± 4.3)	−0.004 (± 0.011)	1.2 (± 4.6)
CANYON-B	−0.006 (± 0.042)	0.04 (± 0.58)	0.0 (± 1.9)	—*	3.6 (± 5.2)	−0.002 (± 0.010)	1.3 (± 5.1)
ESPER_Mixed	−0.001 (± 0.042)	−0.01 (± 0.54)	0.3 (± 1.7)	0.5 (± 5.6)	2.7 (± 4.1)	−0.006 (± 0.012)	0.5 (± 4.1)

Northern Atlantic	**Phosphate**	**Nitrate**	**Silicate**	**Oxygen**	**TA**	**pH**	**DIC**
N	**10,829**	**10,829**	**10,829**	**10,829**	**6619**	**1123**	**4743**
LIRv2	0.009 (± 0.070)	0.14 (± 1.16)	0.3 (± 2.5)	0.7 (± 9.8)	−0.6 (± 6.3)	0.003 (± 0.010)	—
ESPER_LIR	0.006 (± 0.071)	0.05 (± 1.23)	0.3 (± 1.2)	0.6 (± 9.2)	−0.7 (± 5.0)	−0.003 (± 0.011)	1.0 (± 7.7)
ESPER_NN	0.009 (± 0.069)	0.12 (± 0.99)	0.3 (± 1.0)	0.1 (± 7.7)	−1.0 (± 5.4)	−0.004 (± 0.009)	0.9 (± 8.3)
CANYON-B	0.012 (± 0.067)	0.09 (± 1.02)	0.2 (± 1.1)	—*	−0.3 (± 5.7)	−0.004 (± 0.008)	−1.0 (± 8.6)
ESPER_Mixed	0.008 (± 0.067)	0.09 (± 1.05)	0.3 (± 1.0)	0.4 (± 8.2)	−0.8 (± 5.0)	−0.003 (± 0.009)	1.0 (± 7.7)

Sea of Japan/East Sea	**Phosphate**	**Nitrate**	**Silicate**	**Oxygen**	**TA**	**pH**	**DIC**
N	**5995**	**5995**	**5995**	**5995**	**1450**	**0**	**1480**
LIRv2	0.431 (± 0.459)	6.20 (± 6.90)	46.2 (± 54.6)	−19.1 (± 63.2)	−31.7 (± 209.3)	—†	—
ESPER_LIR	0.101 (± 0.154)	1.63 (± 2.11)	3.0 (± 7.7)	6.6 (± 15.5)	51.4 (± 63.0)	—†	2.2 (± 17.2)
ESPER_NN	0.029 (± 0.066)	1.16 (± 1.58)	3.6 (± 4.6)	5.4 (± 10.3)	48.7 (± 55.5)	—†	16.8 (± 20.0)
CANYON-B	0.385 (± 0.409)	5.88 (± 6.42)	21.0 (± 23.6)	—*	28.3 (± 33.8)	—†	12.3 (± 18.4)
ESPER_Mixed	0.065 (± 0.094)	1.40 (± 1.66)	3.3 (± 5.3)	6.0 (± 10.8)	50.0 (± 58.8)	—†	9.5 (± 14.2)

Gulf of Mexico	**Phosphate**	**Nitrate**	**Silicate**	**Oxygen**	**TA**	**pH**	**DIC**
N	**1067**	**1067**	**1067**	**1067**	**943**	**0**	**909**
LIRv2	−0.004 (± 0.123)	0.27 (± 1.71)	0.5 (± 3.8)	8.6 (± 16.1)	−0.9 (± 11.4)	—†	—
ESPER_LIR	−0.009 (± 0.110)	0.30 (± 1.58)	−0.3 (± 2.1)	6.6 (± 16.6)	−16.3 (± 44.5)	—†	−8.7 (± 26.1)

(Continues)

Table 11. Continued

Gulf of Mexico *N*	Phosphate 1067	Nitrate 1067	Silicate 1067	Oxygen 1067	TA 943	pH 0	DIC 909
ESPER_NN	0.002 (± 0.108)	0.35 (± 1.39)	1.0 (± 3.4)	7.3 (± 16.5)	−27.5 (± 47.4)	—†	−19.6 (± 41.6)
CANYON-B	0.056 (± 0.125)	0.68 (± 1.40)	2.5 (± 5.2)	—*	4.5 (± 13.0)	—†	−5.1 (± 16.8)
ESPER_Mixed	−0.003 (± 0.099)	0.32 (± 1.38)	0.4 (± 2.4)	7.0 (± 15.8)	−21.9 (± 45.1)	—†	−14.2 (± 33.4)

Mediterranean *N*	Phosphate 11,394	Nitrate 11,394	Silicate 11,394	Oxygen 11,394	TA 5164	pH 0	DIC 2604
LIRv2	0.081 (± 0.254)	1.90 (± 4.85)	0.5 (± 7.3)	−10.4 (± 50.0)	−37.9 (± 71.2)	—†	—
ESPER_LIR	0.003 (± 0.585)	2.44 (± 7.72)	−4.0 (± 37.5)	−25.1 (± 92.5)	−43.9 (± 72.1)	—†	−105.9 (± 169.9)
ESPER_NN	0.095 (± 0.199)	−2.40 (± 6.21)	−28.6 (± 40.1)	1.8 (± 15.3)	−30.0 (± 43.9)	—†	−40.7 (± 48.9)
CANYON-B	—†	—†	—†	—*	—†	—†	−3.0 (± 26.2)
ESPER_Mixed	0.049 (± 0.325)	0.02 (± 4.82)	−16.3 (± 30.2)	−11.6 (± 45.7)	−37.0 (± 54.3)	—†	−73.3 (± 101.6)

Arctic *N*	Phosphate 6117	Nitrate 6117	Silicate 6117	Oxygen 6117	TA 3189	pH 1634	DIC 2947
LIRv2	0.036 (± 0.122)	0.28 (± 1.20)	0.5 (± 3.4)	2.7 (± 11.8)	1.5 (± 19.4)	—†	—
ESPER_LIR	0.043 (± 0.121)	0.25 (± 1.22)	0.4 (± 2.9)	3.3 (± 11.4)	0.0 (± 12.7)	0.003 (± 0.032)	−1.0 (± 18.7)
ESPER_NN	0.022 (± 0.104)	0.19 (± 0.95)	0.0 (± 2.3)	1.9 (± 11.1)	−2.9 (± 13.3)	0.021 (± 0.054)	−2.6 (± 16.0)
CANYON-B	—†	—†	—†	—†	—†	—†	—†
ESPER_Mixed	0.033 (± 0.099)	0.22 (± 1.00)	0.2 (± 2.3)	2.6 (± 10.8)	−1.5 (± 11.5)	0.012 (± 0.037)	−1.8 (± 16.4)

*This routine does notestimate this quantity.
†No viable comparison in this effort due to partial or complete overlap between training and validation data subsets or insufficient viable measurements.

over the LIRv2 approach for estimating the impacts of OA on pH, it nevertheless is quite simplistic and should not be relied upon when C_{ant} distributions are themselves of interest.

We test the practice of averaging estimates from multiple algorithms and find that it frequently improves estimates (in a global open-ocean RMSE sense). This practice is therefore recommended for most applications, and we suggest further improvements might be obtained by averaging estimates from still more algorithms such as CANYON-B or its updates. A wrapper function for averaging CANYON-B values is under development and may eventually be included at the same GitHub repository as the ESPER functions.

Our assessment also revealed/reinforced several important ideas to consider when using algorithm estimates: First it is critical to have measurements in the training data set that are near to the region in which estimates are desired. Poor reconstructions of the properties of seawater in the Sea of Japan/East Sea from the versions of the routines that did not include measurements in this Sea highlight the importance of this caution. Write-ups of earlier algorithm assessment efforts also cautioned against the use of the algorithms in coastal environments and marginal seas where the algorithms did not have training data, but this case study helps quantify the large likely errors when proceeding despite this caution, as many data-poor marginal seas remain. Second, global oxygen, DIC, and pH estimation routine validation statistics are not as strong as the equivalent statistics when limited to intermediate depths. This is likely because the current generation of algorithms lacks data with sufficient temporal resolution to capture seasonal or shorter patterns of variability associated with gas exchanges. It is possible that the algorithms could be improved by incorporating measurements from the biogeochemical Argo array or other data products that are more seasonally resolved than GLODAPv2, though care would have to be taken to avoid reinforcing the algorithms with float data that is calibrated against earlier versions of the algorithms. This could perhaps be accomplished by removing float measurements that reside below the depths that experience seasonal variability from the data products used to train these future algorithms. At least until such an improvement is made seasonal variability in the estimated fields should be treated with caution.

At intermediate depths, ESPER_LIR_validation Eq. 8 reproduces oxygen with an RMSE of 4.8 μmol kg^{-1} using only T and S as predictors (and 3.7 μmol kg^{-1} for ESPER_Mixed_validation), raising the possibility that estimates could be used to check oxygen sensor performance on in situ platforms. Currently, most float oxygen sensors are subjected to a 1-point gain calibration

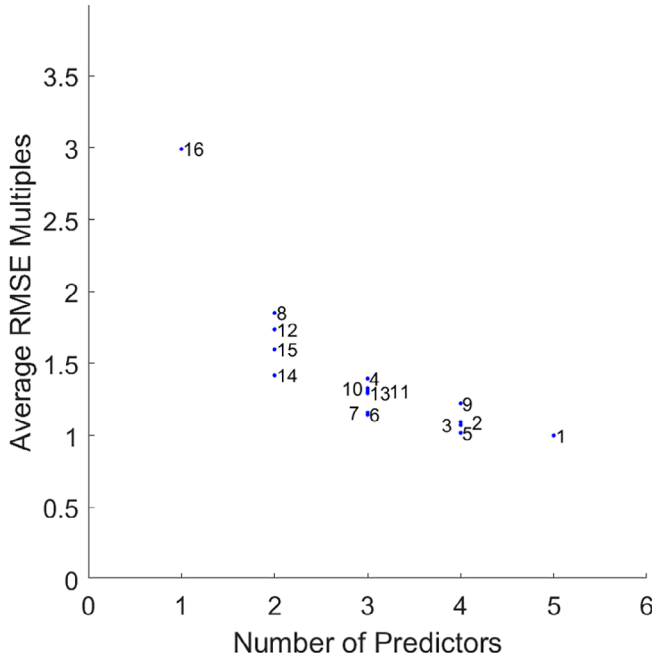

Fig 4. The average global RMSE across all property estimates for both ESPER variants normalized to the RMSE of the equation with the lowest average global RMSE (ESPER Eq. 1) and plotted against the number of predictors required for each estimate (*x*-axis). The point labels correspond to the ESPER equation numbers in Table 2. RMSE generally decreases as the number of predictors increases, but not all predictors have the same predictive power and the incremental increase in predictive power diminishes when more than three predictors are used.

against air–oxygen readings or climatological values at high oxygen concentrations, and a deep algorithm estimate could allow a 2-point check that would assess sensor performance at low oxygen saturation. Comparisons at park depths could circumvent potential issues associated with slow sensor response times.

Our use of a smaller committee of neural networks with somewhat fewer nodes/neurons than is used by CANYON-B is a pragmatic decision based on the computational costs associated with training neural networks for many combinations of predictors and regions, and we have only done a small amount of neural network structure optimization. However, it should be noted that our use of separate network committees for the Indo-Pacific and Arctic-Atlantic regions effectively doubles the complexity of our networks, and that increasing the complexity further did not seem to meaningfully improve our predictions in limited trials. It is nevertheless likely that further improvements in fit and predictive power could be obtained with additional tuning.

While the neural networks are powerful, we demonstrate that the regression-based approach of the ESPER_LIR routines can nevertheless yield comparably skillful estimates in the open ocean or under the right conditions. We contend that the LIR machinery has an advantage of being more explainable than a neural network, and therefore that the LIRs serve a valuable role among seawater prediction routines. An example of where that could prove useful would be in adapting the LIRs to work in an inland sea. A user could append their own grid of regression coefficients determined for a marginal sea such as the Baltic or Mediterranean Seas or an inland waterway such as the Puget Sound, and the routine would transition seamlessly between global estimates and regionally appropriate estimates. This is a future direction for LIR development that would require partnerships with researchers investigating such bodies of water.

The ESPER_LIR routine lacks predictors derived from coordinate information—rather, this information is used in the interpolation of regression coefficients only. As a result, the LIR routines struggle more than the neural networks when applied in regions that are dissimilar from the training data in property space but are nearby in physical space. This can be seen clearly as larger reconstruction errors in the Mediterranean, the Gulf of Mexico, and the Sea of Japan/East Sea. This was doubly true for the LIRv2 routines which tended to also be less well-constrained than the ESPER_LIR (i.e., LIRv3) routines. By contrast, the neural networks also struggle, but tend to have better RMSE statistics for these regions. We reiterate that the release versions of the ESPERs should substantially outperform the bleak assessment statistics given for such regions because the release versions of these routines are trained with data in these regions (unlike the _validation versions, which are used to highlight the dangers of using algorithms in regions where they were not trained).

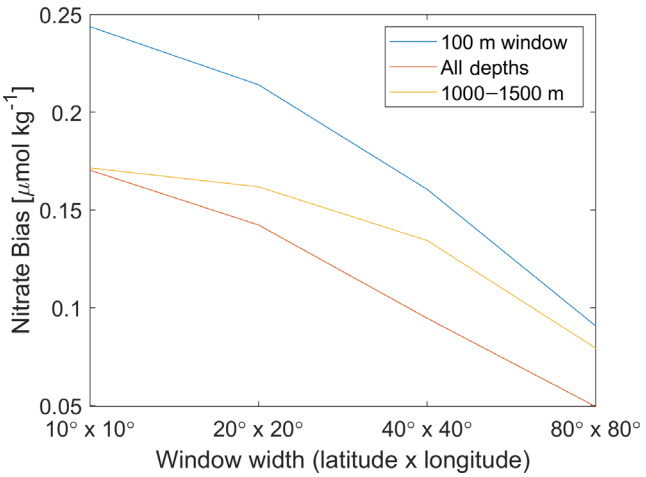

Fig 5. Average absolute bias in ESPER_NN_validation ESPER Eq. 7 nitrate estimates (*y*-axis) vs. the size of the latitude and longitude windows (*x*-axis) over which the average of the absolute biases was computed. The three lines correspond to bias estimates that were averaged over a narrow 100 m depth window (blue line), over all depths (orange), and over the 1000–1500 m depth range commonly used for float calibration (red). Biases are area-weighted average estimates for each of the grid locations used by the ESPER_NN routine. Nitrate ESPER Eq. 7 is chosen as this is one of the equations that is used to calibrate and validate nitrate sensors on biogeochemical Argo floats.

Author contributions

B.R.C. led the data compilation, coding, figure generation, and writing efforts. H.C.B., A.J.F., J.D.S., Y.T., and Y.-Y. Xu provided guidance and input on the code structure and format and aided with testing the routines and iterating on them and their documentation. A.J.F. generated several key figures. M.A., L.B., and A.J.F. identified and provided key data sets. R.W. and R.A.F. played significant roles in securing and sustaining funding for this effort. Critically, all authors aided with writing and vetting this manuscript and provided comments and feedback at multiple stages during planning and writing.

Data availability statement

The training data are available from the GLODAPv2.2020 data product (https://www.glodap.info/). The data from the Gulf of Mexico are available from the National Center for Environmental Information (https://www.ncei.noaa.gov/access/ocean-carbon-data-system/oceans/Coastal/NACP_East.html). The training data from the Mediterranean are compiled as part of the ongoing CARIMED data product synthesis that will be made public through the GLODAP information page. These cruises are listed in Supporting Information S1.1 and can be obtained online individually from the National Center for Environmental Information (e.g., https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.nodc:0214546) or, for some cruises, the Pangaea webpage (https://www.pangaea.de/). The algorithms are publicly accessible and archived as submitted at Zenodo (Carter 2021) https://doi.org/10.5281/zenodo.5512697, and updates will be maintained at the GitHub repository https://github.com/BRCScience Products/ESPER.

References

Álvarez, M., N. M. Fajar, B. R. Carter, E. F. Guallart, F. F. Pérez, R. J. Woosley, and A. Murata. 2020. Global Ocean spectrophotometric pH assessment: Consistent inconsistencies. Environ. Sci. Technol. **54**: 10977–10988. doi:10.1021/acs.est.9b06932

Bittig, H. C., T. Steinhoff, H. Claustre, B. Fiedler, N. L. Williams, R. Sauzède, A. Körtzinger, and J.-P. Gattuso. 2018. An alternative to static climatologies: Robust estimation of open ocean CO_2 variables and nutrient concentrations from T, S, and O2 data using Bayesian neural networks. Front. Mar. Sci. **5**: 328. doi:10.3389/fmars.2018.00328

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

Broullón, D., and others. 2019. A global monthly climatology of total alkalinity: A neural network approach. Earth Syst. Sci. Data **11**: 1109–1127. doi:10.5194/essd-11-1109-2019

Broullón, D., and others. 2020. A global monthly climatology of oceanic total dissolved inorganic carbon: A neural network approach. Earth Syst. Sci. Data **12**: 1725–1743. doi:10.5194/essd-12-1725-2020

Bushinsky, S. M., Y. Takeshita, and N. L. Williams. 2019. Observing changes in ocean carbonate chemistry: Our autonomous future. Curr. Clim. Chang. Rep. **5**: 207–220. doi:10.1007/s40641-019-00129-8

Carter, B. R. 2021. Empirical seawater property estimation routines. doi:10.5281/ZENODO.5512697

Carter, B. R., J. A. Radich, H. L. Doyle, and A. G. Dickson. 2013. An automated system for spectrophotometric seawater pH measurements. Limnol. Oceanogr. Methods **11**: 16–27. doi:10.4319/lom.2013.11.16

Carter, B. R., N. L. Williams, A. R. Gray, and R. A. Feely. 2016. Locally interpolated alkalinity regression for global alkalinity estimation. Limnol. Oceanogr. Methods **14**: 268–277. doi:10.1002/lom3.10087

Carter, B. R., and others. 2017. Two decades of Pacific anthropogenic carbon storage and ocean acidification along Global Ocean Ship-based hydrographic investigations program sections P16 and P02. Global Biogeochem. Cycl. **31**: 306–327. doi:10.1002/2016GB005485

Carter, B. R., R. A. Feely, N. L. Williams, A. G. Dickson, M. B. Fong, and Y. Takeshita. 2018. Updated methods for global locally interpolated estimation of alkalinity, pH, and nitrate. Limnol. Oceanogr. Methods **16**: 119–131. doi:10.1002/lom3.10232

Carter, B. R., R. A. Feely, R. Wanninkhof, S. Kouketsu, R. E. Sonnerup, P. C. Pardo, C. L. Sabine, G. C. Johnson, B. M. Sloyan, A. Murata, S. Mecking, B. Tilbrook, K. Speer, L. D. Talley, F. J. Millero, S. E. Wijffels, A. M. Macdonald, N. Gruber, and J. L. Bullister. 2019a. Pacific anthropogenic carbon between 1991 and 2017. Global Biogeochem. Cycl. **33**: 2018GB006154. doi:10.1029/2018GB006154

Carter, B. R., N. L. Williams, W. Evans, A. J. Fassbender, L. Barbero, C. Hauri, R. A. Feely, and A. J. Sutton. 2019b. Time of detection as a metric for prioritizing between climate observation quality, frequency, and duration. Geophys. Res. Lett. **46**: 3853–3861. doi:10.1029/2018GL080773

Carter, B. R., R. A. Feely, S. K. Lauvset, A. Olsen, T. DeVries, and R. Sonnerup. 2021. Preformed properties for marine organic matter and carbonate mineral cycling quantification. Global Biogeochem. Cycles **35**: e2020GB006623. doi:10.1029/2020GB006623

DeVries, T., M. Holzer, and F. Primeau. 2017. Recent increase in oceanic carbon uptake driven by weaker upper-ocean overturning. Nature **542**: 215–218. doi:10.1038/nature21068

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

Doney, S. C., V. J. Fabry, R. A. Feely, and J. A. Kleypas. 2009. Ocean acidification: The other CO_2 problem. Ann. Rev.

Mar. Sci. **1**: 169–192. doi:10.1146/annurev.marine.010908.163834

Doney, S. C., D. S. Busch, S. R. Cooley, and K. J. Kroeker. 2020. The impacts of ocean acidification on marine ecosystems and reliant human communities. Annu. Rev. Env. Resour. **45**: 83–112. doi:10.1146/annurev-environ-012320-083019

Durack, P. J., S. E. Wijffels, and R. J. Matear. 2012. Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. Science **336**: 455–458. doi:10.1126/science.1212222

Feely, R. A., S. Doney, and S. Cooley. 2009. Ocean acidification: Present conditions and future changes in a high-CO_2 world. Oceanography **22**: 36–47. doi:10.5670/oceanog.2009.95

Feely, R. A., C. L. Sabine, K. Lee, W. Berelson, J. Kleypas, V. J. Fabry, and F. J. Millero. 2004. Impact of anthropogenic CO_2 on the $CaCO_3$ system in the oceans. Science **305**: 362–366. doi:10.1126/science.1097329

Fong, M. B., and A. G. Dickson. 2019. Insights from GO-SHIP hydrography data into the thermodynamic consistency of CO_2 system measurements in seawater. Mar. Chem. **211**: 52–63. doi:10.1016/j.marchem.2019.03.006

Fourrier, M., L. Coppola, H. Claustre, F. D'Ortenzio, R. Sauzède, and J.-P. Gattuso. 2020. A regional neural network approach to estimate water-column nutrient concentrations and carbonate system variables in the Mediterranean Sea: CANYON-MED. Front. Mar. Sci. **7**: 1–20. doi:10.3389/fmars.2020.00620

Gammon, R. H., J. Cline, and D. Wisegarver. 1982. Chlorofluoromethanes in the Northeast Pacific Ocean: Measured vertical distributions and application as transient tracers of upper ocean mixing. J. Geophys. Res. **87**: 9441. doi:10.1029/JC087iC12p09441

Gattuso, J.-P., and others. 2015. Contrasting futures for ocean and society from different anthropogenic CO_2 emissions scenarios. Science **349**. doi:10.1126/science.aac4722

Goyet, C., R. Healy, J. Ryan, and A. Kozyr. 2000. Global distribution of total inorganic carbon and total alkalinity below the deepest winter mixed layer depths. Oak Ridge National Laboratory, Tennessee.

Gray, A. R., and others. 2018. Autonomous biogeochemical floats detect significant carbon dioxide outgassing in the high-latitude Southern Ocean. Geophys. Res. Lett. **45**: 9049–9057. doi:10.1029/2018GL078013

Gregor, L., and N. Gruber. 2021. OceanSODA-ETHZ: A global gridded data set of the surface ocean carbonate system for seasonal to decadal studies of ocean acidification. Earth Syst. Sci. Data **13**: 777–808. doi:10.5194/essd-13-777-2021

Gruber, N., and others. 2019. The oceanic sink for anthropogenic CO_2 from 1994 to 2007. Science. **363**: 1193–1199. doi:10.1126/science.aau5153

Jiang, L.-Q., B. R. Carter, R. A. Feely, S. K. Lauvset, and A. Olsen. 2019. Surface ocean pH and buffer capacity: Past, present and future. Sci. Rep. **9**: 18624. doi:10.1038/s41598-019-55039-4

Jiang, L. Q., and others. 2021. Coastal ocean data analysis product in North America (CODAP-NA)-an internally consistent data product for discrete inorganic carbon, oxygen, and nutrients on the north American ocean margins. Earth Syst. Sci. Data **13**: 2777–2799. doi:10.5194/ESSD-13-2777-2021

Johnson, K. S., and others. 2017. Biogeochemical sensor performance in the SOCCOM profiling float array.

Khatiwala, S., and others. 2013. Global ocean storage of anthropogenic carbon. Biogeosciences **10**: 2169–2191. doi:10.5194/bg-10-2169-2013

Landschützer, P., T. Ilyina, and N. S. Lovenduski. 2019. Detecting regional modes of variability in observation-based surface ocean pCO_2. Geophys. Res. Lett. **46**: 2670–2679. doi:10.1029/2018GL081756

Lauvset, S. K., and others. 2016. A new global interior ocean mapped climatology: The $1^\circ \times 1^\circ$ GLODAP version 2. Earth Syst. Sci. Data **8**: 325–340. doi:10.5194/ESSD-8-325-2016

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. **33**: L19605. doi:10.1029/2006GL027207

Nerem, R. S., B. D. Beckley, J. T. Fasullo, B. D. Hamlington, D. Masters, and G. T. Mitchum. 2018. Climate-change–driven accelerated sea-level rise detected in the altimeter era. Proc. Natl. Acad. Sci. USA **115**: 2022–2025. doi:10.1073/pnas.1717312115

Olden, J. D., and D. A. Jackson. 2002. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. Ecol. Model. **154**: 135–150. doi:10.1016/S0304-3800(02)00064-9

Olsen, A., and others. 2016. The Global Ocean data analysis project version 2 (GLODAPv2)—An internally consistent data product for the world ocean. Earth Syst. Sci. Data **8**: 297–323. doi:10.5194/essd-8-297-2016

Olsen, A., and others. 2019. GLODAPv2.2019—An update of GLODAPv2. Earth Syst. Sci. Data **11**: 1437–1461. doi:10.5194/essd-11-1437-2019

Olsen, A., and others. 2020. An updated version of the global interior ocean biogeochemical data product, GLODAPv2.2020. Earth Syst. Sci. Data **12**: 3653–3678. doi:10.5194/essd-12-3653-2020

Purkey, S. G., and G. C. Johnson. 2013. Antarctic bottom water warming and freshening: Contributions to sea level rise, ocean freshwater budgets, and global heat gain. J. Climate **26**: 6105–6122. doi:10.1175/JCLI-D-12-00834.1

Redfield, A. C., B. H. Ketchum, and A. F. Richards. 1963. The influence of organisms on the composition of seawater. Sea **2**: 26–77.

Roemmich, D., W. John Gould, and J. Gilson. 2012. 135 Years of global ocean warming between the challenger expedition and the Argo Programme. Nat. Clim. Chang. **2**: 425–428. doi:10.1038/nclimate1461

Sabine, C. L., and others. 2004. The oceanic sink for anthropogenic CO_2. Science **305**: 367–371. doi:10.1126/science.1097403

Sasano, D., Y. Takatani, N. Kosugi, T. Nakano, T. Midorikawa, and M. Ishii. 2018. Decline and Bidecadal oscillations of dissolved oxygen in the Oyashio region and their propagation to the Western North Pacific. Global Biogeochem. Cycl. **32**: 909–931. doi:10.1029/2017GB005876

Sauzède, R., H. C. Bittig, H. Claustre, O. Pasqueron de Fommervault, J.-P. Gattuso, L. Legendre, and K. S. Johnson. 2017. Estimates of water-column nutrient concentrations and carbonate system parameters in the Global Ocean: A novel approach based on neural networks. Front. Mar. Sci. **4**: 128. doi:10.3389/fmars.2017.00128

Sharp, J. D., and R. H. Byrne. 2020. Interpreting measurements of total alkalinity in marine and estuarine waters in the presence of proton-binding organic matter. Deep. Res. Part I Oceanogr. Res. Pap **165**: 103338. doi:10.1016/j.dsr.2020.103338

Takeshita, Y., K. S. Johnson, T. R. Martz, J. N. Plant, and J. L. Sarmiento. 2018. Assessment of autonomous pH measurements for determining surface seawater partial pressure of CO_2. J. Geophys. Res. Ocean. **123**: 4003–4013. doi:10.1029/2017JC013387

Takeshita, Y., K. S. Johnson, L. J. Coletti, H. W. Jannasch, P. M. Walz, and J. K. Warren. 2020. Assessment of pH dependent errors in spectrophotometric pH measurements of seawater. Mar. Chem. **223**: 103801. doi:10.1016/j.marchem.2020.103801

Takeshita, Y., and others. 2021. Consistency and stability of purified meta-cresol purple for spectrophotometric pH measurements in seawater. Mar. Chem. **236**: 104018. doi:10.1016/J.MARCHEM.2021.104018

Tanhua, T., A. Körtzinger, K. Friis, D. W. Waugh, and D. W. R. Wallace. 2007. An estimate of anthropogenic CO_2 inventory from decadal changes in oceanic carbon content. Proc. Natl. Acad. Sci. USA **104**: 3037–3042. doi:10.1073/pnas.0606574104

Tu, J. V. 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. **49**: 1225–1231. doi:10.1016/S0895-4356(96)00002-9

van Heuven, S. M. A. C. et al. 2011. MATLAB program developed for CO_2 system calculations. ORNL/CDIAC-105b. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee. doi:10.3334/cdiac/otg.co2sys_matlab_v1.1

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. **111–112**: 11–18. doi:10.1016/j.jmarsys.2012.09.002

Waugh, D. W., T. M. Hall, B. I. Mcneil, R. Key, and R. J. Matear. 2006. Anthropogenic CO_2 in the oceans estimated using transit time distributions. Tellus B Chem. Phys. Meteorol. **58**: 376–389. doi:10.1111/j.1600-0889.2006.00222.x

Williams, N. L., and others. 2016. Empirical algorithms to estimate water column pH in the Southern Ocean. Geophys. Res. Lett. **43**: 3415–3422. doi:10.1002/2016GL068539

Williams, N. L., and others. 2017. Calculating surface ocean pCO2 from biogeochemical Argo floats equipped with pH: An uncertainty analysis. Global Biogeochem. Cycl. **31**: 591–604. doi:10.1002/2016GB005541

Williams, N. L., L. W. Juranek, R. A. Feely, J. L. Russell, K. S. Johnson, and B. Hales. 2018. Assessment of the carbonate chemistry seasonal cycles in the Southern Ocean from persistent observational platforms. J. Geophys. Res. Ocean. **123**: 4833–4852. doi:10.1029/2017JC012917

Woosley, R. J., F. J. Millero, and R. Wanninkhof. 2016. Rapid anthropogenic changes in CO_2 and pH in the Atlantic Ocean: 2003–2014. Global Biogeochem. Cycl. **30**: 1–21. doi:10.1002/2015GB005248

Acknowledgments

Carter, Feely, and Wanninkhof thank the Global Ocean Monitoring and Observing (GOMO) program of the National Oceanic and Atmospheric Administration (NOAA) for funding algorithm development under the Carbon Data Management and Synthesis Grant (Fund ref. #100007298). Regional data contributions and validation efforts originate from NOAA National Oceanographic Partnership Program (NOPP) funding (NA19OAR4310362), the NOAA Ocean Acidification Program, and GOMO support for biogeochemical Argo. Award number 2048509 from the National Science Foundation of the United States supported development of the anthropogenic carbon adjustment strategy. We further thank the scientists and crew aboard the research vessels that collected these data, and the National Science Foundation and the NOAA GOMO program for supporting the critical Global Ocean Ship-based Hydrographic Investigations Program and other cruise programs. Bittig acknowledges funding from the DArgo2025 and C-SCOPE projects (grant No. 03F0857D and 03F0877D). Fassbender was supported by GOMO. Álvarez was supported by the RADIALES, RADPROF and MedSHIP IEO monitoring programs. This research was carried out in part under the auspices of the Cooperative Institutes for Climate, Ocean, and Ecosystem Studies (CICOES) and Marine and Atmospheric Studies (CIMAS), Cooperative Institutes between Universities of Washington and Miami (respectively) and the National Oceanic and Atmospheric Administration, cooperative agreement numbers NA15OAR4320063 and NA20OAR4320472, respectively. This is CICOES contribution number 2020-1138 and PMEL contribution number 5243.

Conflict of Interest

None declared.

Submitted 04 May 2021
Revised 01 September 2021
Accepted 11 September 2021

Associate editor: Krista Longnecker