Supplement of Earth Syst. Sci. Data, 17, 719–740, 2025
https://doi.org/10.5194/essd-17-719-2025-supplement

Supplement of

A global monthly 3D field of seawater pH over 3 decades: a machine learning approach

Guorong Zhong et al.

Correspondence to: Xuegang Li (lixuegang@qdio.ac.cn) and Jinming Song (jmsong@qdio.ac.cn)

S1. Uncertainty and construction method of selected ocean products

A group of products related to the physical, chemical, and biological activities that influence the ocean carbonate system were collected as potential pH predictors (Table 1). These products were constructed using different methods in previous research. The seawater temperature and salinity product were constructed based on measurements from the World Ocean Database (WOD) using the ensemble optimal interpolation method with the dynamic ensemble (EnOI-DE) provided by CMIP5 historical simulations (Cheng et al., 2016; Cheng et al., 2020). The temperature product was claimed with an uncertainty of about ±0.05°C in the recent few decades, and the uncertainty of salinity product was about ±0.001~±0.005 at different depths (present as figures in Cheng et al., 2016 and Cheng et al., 2020; *https://journals.ametsoc.org/view/journals/clim/33/23/full-jcliD200366-f5.jpg* and *https://journals.ametsoc.org/view/journals/clim/29/15/full-jcli-d-15-0730.1-f8.jpg*). The climatological Alk product was constructed from Global Ocean Data Analysis Project version 2.2019 (GLODAPv2019) measurements using a neural network (NNGv2) method, with the RMSE of 3–6.2 μmol kg^{-1} (Broullón et al., 2019). The climatological DIC product was constructed from GLODAPv2019 and the Lamont–Doherty Earth Observatory (LDEO) datasets using a feedforward neural network (dubbed NNGv2LDEO) method, with a RMSE of 3.6–13.2 μmol kg^{-1} (Broullón et al., 2020). The climatological dissolved oxygen, nitrate, phosphate, and silicate product was constructed based on measurements from the World Ocean Database, using an objective analysis method that generated a first-guess field and then carried out a correction at all gridpoints as a distance-weighted mean of all gridpoint difference values that lie within the area around the gridpoint defined by the influence radius (Garcia et al., 2019a; Garcia et al., 2019b). The producer claimed an average DO bias of 0.4±4.7 μmol kg^{-1} below 500 m depth and 1.4±10.9 μmol kg^{-1} above 500 m depth. The average biases of nutrient concentration were -0.02±0.07 μmol kg^{-1} for phosphate, -0.22±0.95 μmol kg^{-1} for nitrate, and -0.3±3.8 μmol kg^{-1} for silicate below 500 m depth, and were 0.01±0.12 μmol kg^{-1} for phosphate, 0.2±1.8 μmol kg^{-1} for nitrate, and 0.8±3.6 μmol kg^{-1} for silicate above 500 m depth. The Sea surface height (SSH), mixed layer depth (MLD), and W velocity of ocean current from the ECCO2 cube92 product were constructed by least squares fit of a global full-depth-ocean and sea-ice configuration of the Massachusetts Institute of Technology general circulation model to the available satellite and in-situ data (Menemenlis et al., 2008). The basin-wide median bias error of the MLD product is -6.6 m and the RMSE is 40 m, and the RMSE of the SSH product is 9.2 cm. The ERA5 sea level pressure and surface pressure were constructed by the Integrated Forecasting System (IFS) Cy41r2 model (Hersbach et al., 2020). The standard deviation of ERA5 sea level pressure and surface pressure are within 1 hPa and 0.8 hPa in the recent three decades. The NOAA Greenhouse Gas Marine Boundary Layer Reference xCO_2 product is constructed by extending measurements from a subset of sites from the NOAA Cooperative Global Air Sampling Network, with an uncertainty within 1 μmol mol^{-1} in most regions (Lan et al., 2023, https://gml.noaa.gov/ccgg/mbl/mbl.html). The bi-monthly Multivariate El Niño/Southern Oscillation index (MEI) was calculated by the first seasonally varying principal component of six atmosphere–ocean (COADS) variable fields in the tropical Pacific basin (Wolter et al., 2011). The Arctic Oscillation index was calculated as the first leading mode from the Emperical Orthogonal Function analysis of monthly mean height anomalies at 1000-hPa of the Northern Hemisphere or 700-hPa of the Southern Hemisphere (CPC, 2002). The Southern Oscillation Index was calculated based on the differences in air pressure anomaly between Tahiti and Darwin, Australia (CPC, 2005). The specific uncertainty of these index products is not provided. The GEBCO global bathymetric data was constructed using predicted depths based on the V32 gravity model (Sandwell et al., 2019). The monthly surface ocean pCO_2 was constructed using the SOM-FFNN method based on regional-specific predictors selected by the stepwise FFNN algorithm, with a global RMSE of 17.99 μatm (Zhong et al., 2022). A climatological pCO_2 product constructed by another SOM-FFNN model was also used, with the RMSE of 18.3 μatm (Landschützer et al., 2020). The Euphotic Depth product was constructed from remote sensing reflectance (RRS) data derived inherent optical properties using Lee algorithm (Lee et al., 2007), with an average percentage error of 13.7%. The chlorophyll concentration product was constructed based on RRS at 2-4 wavelengths between 440 and 670 nm with an uncertainty of 1-2%, using the algorithm of Hu et al. (2019) that combines an empirical band difference approach at low chlorophyll concentrations with a band ratio approach at higher chlorophyll concentrations. The photosynthetically available radiation (PAR) product was based on the observed Top-of-Atmosphere (TOA) radiances in the 400-700nm range that do not saturate over clouds using the algorithm of Frouin et al. (2002), with an RMSE of 3.6 Einstein/m^2/day. The product of the diffuse attenuation coefficient at 490 nm (Kd490) was calculated using an empirical relationship derived from in situ measurements of Kd490 and blue-to-green band ratios of RRS. The remote sensing reflectance product was derived from ocean color sensors based on the spectral distribution of reflected visible solar radiation upwelling from below the ocean surface and passing through the sea-air interface. The total absorption and backscattering products were calculated using the default global configuration of the Generalized Inherent Optical Property (GIOP) model (Werdell et al., 2013).

S2. Validation of cross-boundary method

The cross-boundary method reduced the pH predicting error slightly, but improved the discontinuity problem in the SOM boundary effectively (Figure S1 a-d). However, the discontinuity problem was not completely solved and some boundary line existed in the spatial distribution, especially in the deeper ocean that pH measurements are much sparser (Figure S1 e-f). Even so, the performance of FFNN predicting was better when the cross-boundary method was applied. Compared with taking average in the boundary area, the cross-boundary method avoided subjectively modifying the boundary data. Correspondingly, this method may not solve the discontinuity problem perfectively in some situations. The cross-boundary method also decreased the predicting error slightly in vertical boundary areas (2 layers near the mixed layer depth). However, the improvement was minor in the vertical distribution, due to the natural existing substantial vertical gradient of seawater pH near the mixed layer depth (Figure S2). Overall, the cross-boundary method increases information about seawater pH variation out of boundaries in the neural network learning process, reducing the outliers near the SOM boundary and vertical boundary.

S3. Comparison of performance between FFNNs training based on pH and [H^+]

Due to the logarithmic relationship between pH value and [H^+] concentration, results obtained from training FFNN with pH and from training FFNN with [H^+] then converting outputs into pH may differ. A comparison of predicting errors was conducted between these two training methods. The results show a nearly consistent pH RMSE between the FFNN training with pH and with [H^+] (Figure 7). As the pH measurements of all GLODAP samples are closer to a normal distribution than the [H^+], the predicting error was slightly lower in most regions when the FFNN was trained with pH, but the difference in predicting errors was extremely small. In addition, the FFNN trained using [H^+] occasionally produced negative [H^+] in regions with extremely low [H^+]. Therefore, it is better to train FFNN using pH rather than using [H^+] in the reconstruction process of the pH product.

The distribution patterns of regional pH RMSE and $[H^+]$ RMSE are inconsistent whenever the FFNN was trained using pH or $[H^+]$. In fact, the pH RMSE of the intermediate layer in regions such as the subarctic North Pacific and the equatorial Pacific is significantly lower than that in the intermediate layer of the Arctic Ocean, but their $[H^+]$ RMSE is higher than that of the intermediate layer in the Arctic Ocean (Figure 7a and 7b). This is caused by the effect of the logarithmic relationship. If the pH values are different for the same pH RMSE, the corresponding $[H^+]$ RMSE will be different. Therefore, the uncertainty of the pH product is calculated based on the $[H^+]$ RMSE and pH value, rather than solely based on the pH RMSE.

Figure S1. Statistical distribution of GLODAP samples used for training and testing in each province. Iteration 1-4: repeated evaluation with different training and testing samples dividing by years. Samples in 1992, 1996, …, 2020 were used for testing and the rest were used for training in iteration 1; samples in 1993, 1997, …, 2017 were used for testing and the rest were used for training in iteration 2.

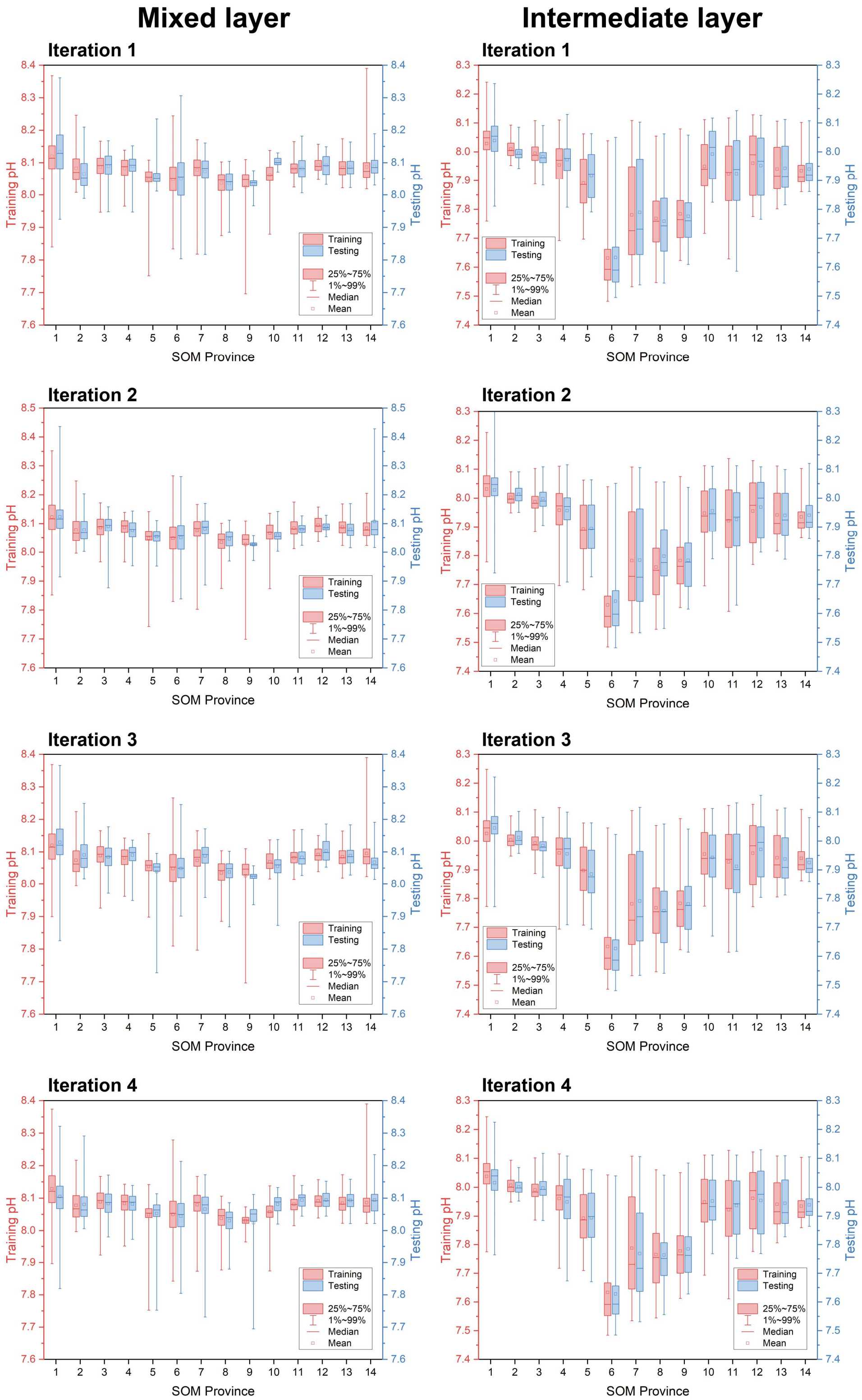

Figure S2. Validation of cross-boundary method for pH predicting in the SOM boundary. a-b): comparison of FFNN predicted pH with GLODAP in all SOM boundary areas; c-f): comparison of spatial distribution at 0 m and 1000 m in January 2020.

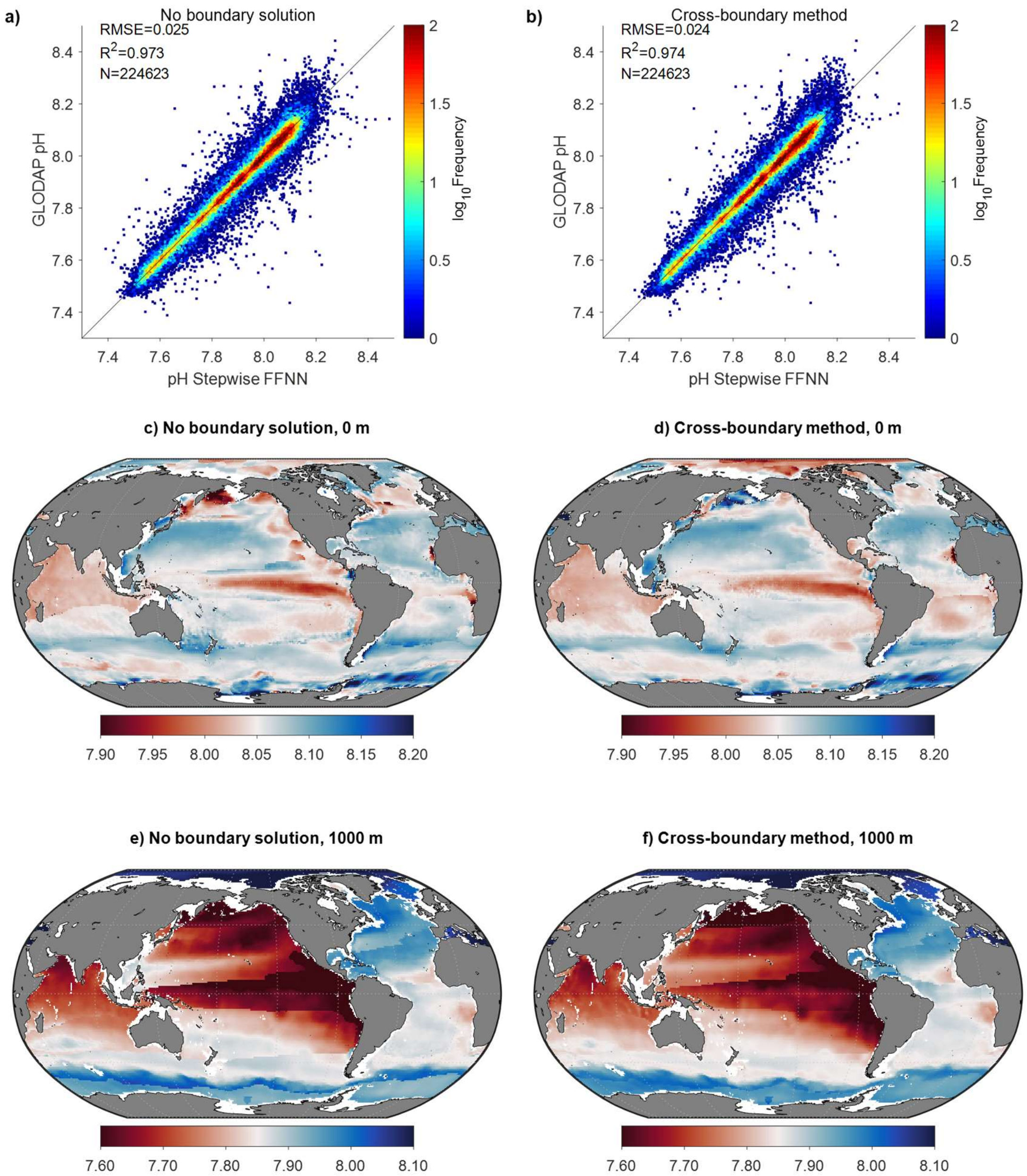

Figure S3. Validation of cross-boundary method for pH predicting in the vertical boundary. a) and b): comparison of FFNN predicted pH with GLODAP in all vertical boundary areas (2 layers near the mixed layer depth); c) and d): comparison of vertical distribution at different basin in January 2020.

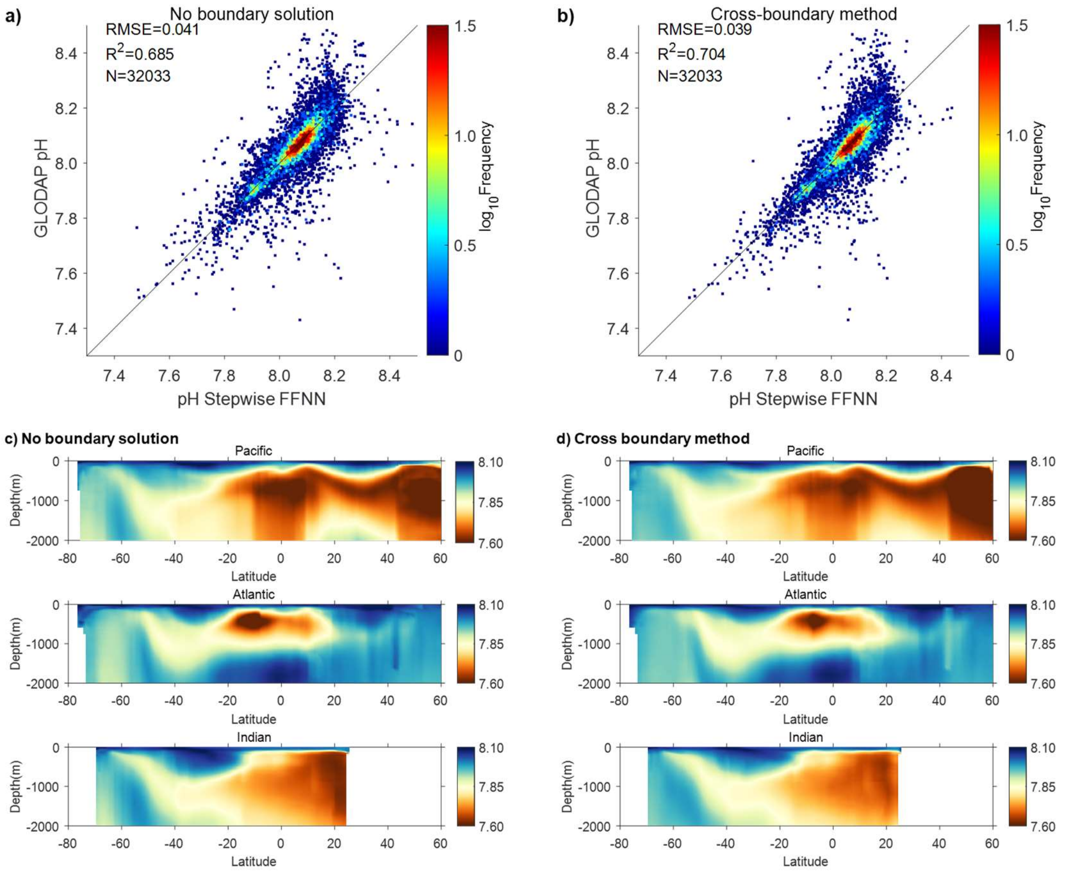

Figure S4. Comparison of pH RMSE and [H$^+$] RMSE from training FFNN using pH and using [H$^+$]. a): pH RMSE of FFNN trained using pH and [H$^+$] in each biogeochemical province, the predicted [H$^+$] from FFNN trained using [H$^+$] was converted to pH for estimating pH RMSE. b): [H$^+$] RMSE of FFNN trained using pH and [H$^+$] in each biogeochemical province; c): [H$^+$] RMSE of FFNN trained using pH and [H$^+$] in each vertical layer; the predicted pH from FFNN trained using pH was converted to [H$^+$] for estimating [H$^+$] RMSE. The numbers shown in the X-axis represent the SOM province in Figure 1.

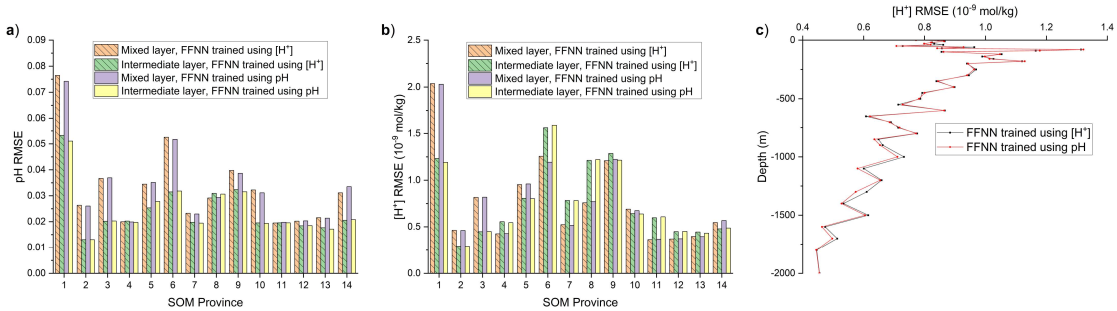

Figure S5. Mean standard deviation between FFNN pH with different initial status.

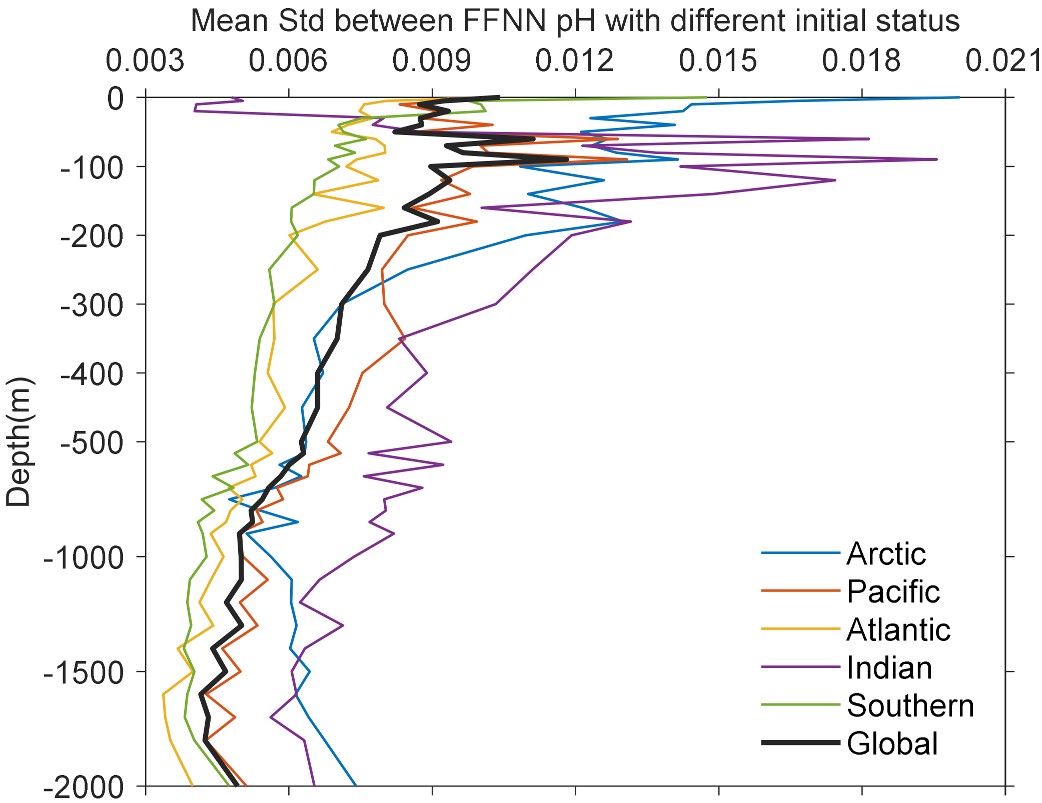

Figure S6. Station map of used delayed-mode BGC-Argo pH-adjusted data with quality control flag 1.

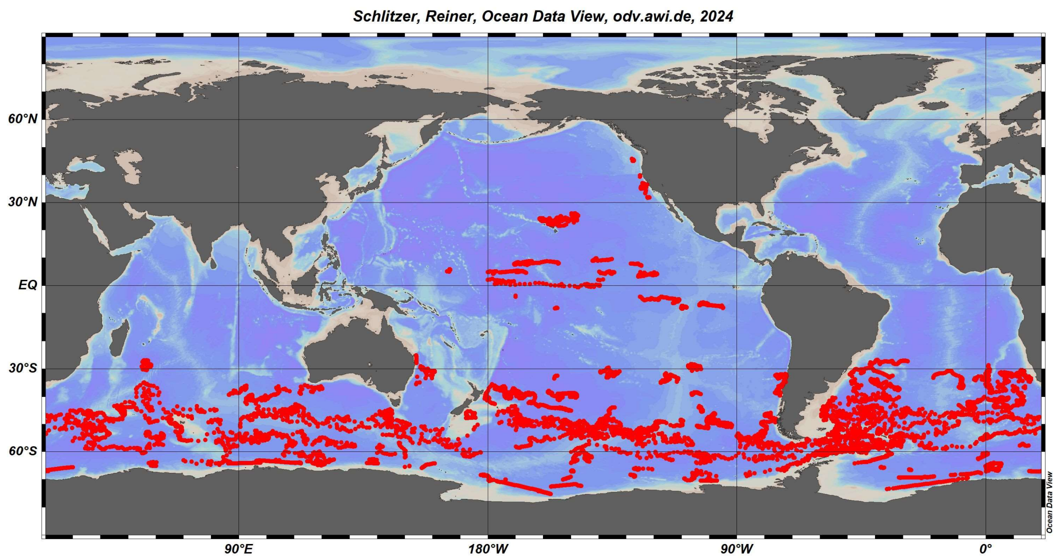

Table S1. Predictors selected by the stepwise FFNN algorithm in the Mixed layer for period before August 2002. The predictors are arranged in order of relative importance, with the variables listed at the front of each province being more effective in reducing predicting errors when used as pH predictors.

Province	FFNN neurons	Predictors
P5 Equatorial Atlantic	25	Phosphate, Temp, SLP, DIC, P_{surf}, TA, pCO_2, W_{vel}(in-situ), DO
P8 Equatorial Pacific	10	pCO_2, Depth, sLat, Temp, Sal, DIC, W_{vel}(in-situ), Nitrate
P10 Subtropical South Atlantic	20	pCO_2, Silicate, Nitrate, W_{vel}(65m), W_{vel}(in-situ), W_{vel}(195m)
P11 Subtropical South Pacific	10	Phosphate, pCO_2, Depth, sLat, Silicate, $pCO_{2\ clim}$, W_{vel}(5m), W_{vel}(105m)

References mentioned in supplementary text:

Broullón, D., Pérez, F. F., Velo, A., Hoppema, M., Olsen, A., Takahashi, T., Key, R. M., Tanhua, T., González-Dávila, M., Jeansson, E., Kozyr, A., and van Heuven, S. M. A. C.: A global monthly climatology of total alkalinity: a neural network approach, Earth Syst. Sci. Data, 11, 1109–1127, https://doi.org/10.5194/essd-11-1109-2019, 2019.

Broullón, D., Pérez, F. F., Velo, A., Hoppema, M., Olsen, A., Takahashi, T., Key, R. M., Tanhua, T., Santana-Casiano, J. M., and Kozyr, A.: A global monthly climatology of oceanic total dissolved inorganic carbon: a neural network approach, Earth Syst. Sci. Data, 12, 1725–1743, https://doi.org/10.5194/essd-12-1725-2020, 2020.

Cheng, L. and Zhu, J.: Benefits of CMIP5 multimodel ensemble in reconstructing historical ocean subsurface temperature variations, J. Clim., 29, 5393-5416, https://doi.org/10.1175/JCLI-D-15-0730.1, 2016.

Cheng, L., Trenberth, K. E., Gruber, N., Abraham, J. P., Fasullo, J. T., Li, G., Mann, M. E., Zhao, X., and Zhu, J.: Improved estimates of changes in upper ocean salinity and the hydrological cycle, J. Clim., 33, 10357-10381, https://doi.org/10.1175/JCLI-D-20-0366.1, 2020.

Climate Prediction Center: Daily Arctic Oscillation Index [data set], https://www.cpc.ncep.noaa.gov/products/precip/CWlink/daily_ao_index/ao_index.html, 2002.

Climate Prediction Center: Southern Oscillation Index [data set], https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensocycle/soi.shtml, 2005.

Frouin, R., Franz, B. A., and Werdell, P. J.: The SeaWiFS PAR product. ,In: S.B. Hooker and E.R. Firestone, Algorithm Updates for the Fourth SeaWiFS Data

Reprocessing, NASA Tech. Memo, 2003-206892, Volume 22, NASA Goddard Space Flight Center, Greenbelt, Maryland, 46-50, 2002.

Garcia, H. E., Weathers, K. W., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, R. A., Zweng, M. M., Mishonov, A. V., Baranova, O. K., Seidov, D., and Reagan, J. R.: World Ocean Atlas 2018, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Dissolved Oxygen Saturation, edited by: Mishonov, A., NOAA Atlas NESDIS 83, 38 pp., https://www.nodc.noaa.gov/OC5/woa18/pubwoa18.htm, 2019a.

Garcia, H. E., Weathers, K. W., Paver, C. R., Smolyar, I., Boyer, T. P., Locarnini, R. A., Zweng, M. M., Mishonov, A. V., Baranova, O. K., Seidov, D., and Reagan, J. R.: World Ocean Atlas 2018. Vol. 4: Dissolved Inorganic Nutrients (phosphate, nitrate and nitrate+nitrite, silicate). A. Mishonov Technical Editor, NOAA Atlas NESDIS 84, 35 pp., https://archimer.ifremer.fr/doc/00651/76336/, 2019b.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas J., Peubey C., Radu R., Schepers D., Simmons A., Soci C., Abdalla S., Abellan X., Balsamo G., Bechtold P., Biavati G., Bidlot J., Bonavita M., Chiara G. D., Dahlgren P., Dee D., Diamantakis M., Dragani R., Flemming J., Forbes R., Fuentes M., Geer A., Haimberger L., Healy S., Hogan R. J., Hólm E., Janisková M., Keeley S., Laloyaux P., Lopez P., Lupu C., Radnoti G., Rosnay P. D., Rozum I., Vamborg F., Villaume S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

Hu, C., Feng, L., Lee, Z., Franz, B. A., Bailey, S. W., Werdell, P. J., and Proctor, C. W.: Improving satellite global chlorophyll a data products through algorithm refinement and data recovery, J. Geophys. Res.-Oceans, 124(3), 1524-1543, https://doi.org/10.1029/2019JC014941, 2019.

Lan, X., Tans, P., Thoning, K., and NOAA Global Monitoring Laboratory: NOAA Greenhouse Gas Marine Boundary Layer Reference - CO_2, NOAA GML [Data set], https://doi.org/10.15138/DVNP-F961, 2023.

Landschützer, P., Laruelle, G. G., Roobaert, A., and Regnier, P.: A uniform $p$$CO_2$ climatology combining open and coastal oceans, Earth Syst. Sci. Data, 12, 2537–2553, https://doi.org/10.5194/essd-12-2537-2020, 2020.

Lee, Z., Weidemann, A., Kindle, J., Arnone, R., Carder, K. L., and Davis, C.: Euphotic zone depth: Its derivation and implication to ocean-color remote sensing, J. Geophys. Res., 112, C3, https://doi.org/10.1029/2006JC003802, 2007.

Menemenlis, D., Campin, J. M., Heimbach, P., Hill, C., Lee, T., Nguyen, A., Schodlok, M., and Zhang, H.: ECCO2: High resolution global ocean and sea ice data synthesis, *Mercat. Ocean Q. Newsl,* 31, 13-21, 2008.

Sandwell, D. T., Harper, H., Tozer, B., and Smith, W. H.: Gravity field recovery from geodetic altimeter missions, Adv. Space Res., 68(2), 1059-1072, https://doi.org/10.1016/j.asr.2019.09.011, 2021.

Werdell, P. J., Franz, B. A., Bailey, S. W., Feldman, G. C., Boss, E., Brando, V. E., Dowell M., Hirata T., Lavender S. J., Lee, Z., Loisel H., Maritorena S., Mélin F., Moore T. S., Smyth T. J., Antoine D., Devred E., d'Andon O. H. F., and Mangin, A.: Generalized ocean color inversion model for retrieving marine inherent optical properties. Appl. Optics, 52(10), 2019-2037, https://doi.org/10.1364/AO.52.002019, 2013.

Wolter, K. and Timlin, M. S.: El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI. ext), Int. J. Climatol., 31, 1074-1087, https://doi.org/10.1002/joc.2336, 2011.

Zhong, G., Li, X., Song, J., Qu, B., Wang, F., Wang, Y., Zhang, B., Sun, X., Zhang, W., Wang, Z., Ma, J., Yuan, H., and Duan, L.: Reconstruction of global surface ocean $p\mathrm{CO_2}$ using region-specific predictors based on a stepwise FFNN regression algorithm, Biogeosciences, 19, 845–859, https://doi.org/10.5194/bg-19-845-2022, 2022.