Supplement A: Defining the mixed-layer

Derived by wind stress and air-sea heat exchange, the mixed-layer depth (MLD) describes the maximum penetration depth of the quasi-homogeneous region of surface water (Kara et al., 2003). Typically ranging from 20m in summer months, to 500m during the winter season in some parts of the ocean (de Boyer et al., 2004), including MLD measurements is an important additional constraint on carbon dynamics that is added from bottle measurements.

Discriminating mixed-layer measurements for each cast was conducted via a bivariant linear interpolation from a regular 2° by 2° gridded MLD climatology developed by (de Boyer et al., 2004). Their methodology was based on a change in potential density from a 10m reference measurement of 0.03 kg m^{-3}, approximately 900,000 CTD profiles including Argo data up to September 2008 were used to constrain their MLD climatology.

Supplement B: Identifying coastal data

Carbon biogeochemical dynamics in coastal zones have been shown to be divorced from the open ocean system due to terrigenous influences (e.g., Cotrim da Cunha et al., 2007; Gibbs et al., 2006; Jickells, 1998; Seitzinger et al., 2005). Sediment upwelling, anthropogenic influences on coastal ecosystems, and nutrification and carbon delivery from rivers have been identified as processes perturbing coastal biogeochemical dynamics from the open ocean. To eliminate these biases from our oceanic dataset, all casts with a seafloor bathymetry of 500m or less were removed from the mixed-layer training dataset. The bathymetric depth for each cast was linearly interpolated from NOAA’s 1 arcminute global relief product, re-gridded to 10 arcseconds (Amante and Eakins, 2009). Eliminating coastal influences reduces the dataset by ~9%, but is important when applying the NN approach.

Supplement C: Anthropogenic correction for C_T measurements

The Revelle factor (*R*) quantifies the relationship between the fractional changes of ocean pCO_2 and C_T concentrations in an otherwise static system (Eq. C1), and is therefore a well suited empirical means to account for anthropogenic biases in C_T measurements.

$$R = \frac{C_T}{pCO_2}\frac{\Delta pCO_2}{\Delta C_T} \quad (C1)$$

Constraining the anthropogenic C_T component (ΔC_T) requires in-situ measurements of C_T and pCO_2, the anthropogenic change in pCO_2 (ΔpCO_2), and Revelle factor (Eq. C2).

$$\Delta C_T = \frac{C_T}{R}\frac{\Delta pCO_2}{pCO_2} \quad (C2)$$

In-situ Revelle factor and pCO_2 concentrations were calculated using the CO2SYS program developed by (Pierrot et al., 2006). Selection of the (Mehrbach et al., 1973) constants as refitted by (Dickson and Millero, 1987) was based on findings by (Lee et al., 2000a; McNeil et al., 2007; Millero et al., 2002; Wanninkhof et al., 1999) and maintained consistency with GLODAP and CARINA products, (Key et al., 2004; Pierrot et al., 2010). Here, we assume the anthropogenic rate of increase in mixed-layer pCO_2 is in equilibrium with the atmosphere, which facilitates the ability to constrain ΔpCO_2 through atmospheric CO_2 measurements from the Mauna Loa observation site (Dr. Pieter Tans, NOAA/ESRL, www.esrl.noaa.gov/gmd/ccgg/trends and Dr. Ralph Keeling, Scripps Institute of Oceanography scrippsco2.ucsd.edu/). The final empirical equation to correct C_T measurements to the reference year 2000 is

$$C_{T(sw,2000)} = C_{T(sw,\text{in-situ year})} + \left(\frac{CO_{2(atm,2000)} - CO_{2(atm,\text{in-situ year})}}{pCO_{2(sw,\text{in-situ year})}}\right)\frac{C_{T(sw,\text{in-situ year})}}{R} \quad (C3)$$

where subscripts sw and atm represent sea-water and atmosphere respectively.

Two key assumptions underlying this methodology include a constant Revelle factor over the correction period, and a global representation of atmospheric CO_2 changes from the Mauna Loa site. As the ocean absorbs more anthropogenic CO_2 the Revelle factor will increase, however recent studies have estimated R to have only slightly changed over the past 2 centuries (Egleston et al., 2010), which validates our assumption of a constant R value over a maximum 20 year correction period. To evaluate the applicability of the Mauna Loa ΔCO_2 on a global scale, we compared the net change in atmospheric CO_2 as observed at the Mauna Loa site to a global estimate derived from multiple stations (Thomas Conway and Pieter Tans, NOAA/ESRL, www.esrl.noaa.gov/gmd/ccgg/trends) (Fig. C1). Here, we find a high degree of similarity between the two estimates, and when taking into consideration an uncertainty in these estimates of 0.1 μatm yr^{-1}, the differences between the two approaches is negligible.

Calculation of Revelle factors and pCO_2 concentrations using the CO2SYS program required in-situ measurements of temperature, salinity, A_T and C_T. Of the total mixed-layer C_T measurements, 8,711 (or ~28%) were missing at least one of these additional parameters required to constrain the anthropogenic correction using the proposed technique. Rather than discarding this data, 22,727 corrected C_T measurements were employed to constrain the anthropogenic correction using a 4-D linear interpolation in latitude, longitude, in-situ pressure and the calculated annual anthropogenic rate of C_T increase. To evaluate the skill of the interpolation approach, we divided the 22,727 measurements into 10 equal subsets and independently interpolated the anthropogenic rate of increase. We found the approach captured the increase to within 0.08 μmol yr^{-1} (or 8% for the mean value).

The global rate of increase in mixed-layer C_T concentration was found to be 0.996 μmol kg^{-1} yr^{-1} (Fig. C2), which is consistent with the 1 μmol kg^{-1} yr^{-1} anthropogenic C_T correction rate used by (Lee et al., 2000b) for measurements between 30°N and 30°S and is also consistent with reported rates of increase observed at the HOT (Winn et al., 1998) and BATS (Bates, 2007) time-series stations.

Supplement D: Significance of anthropogenic C_T correction

To test the significance of anthropogenic C_T corrections we applied the global independent test approach (GIT, see Sect. 3) to models trained using data with and without anthropogenic C_T corrections. The global RSE for a C_T model trained using C_T measurements without applying anthropogenic corrections was 13.2 μmol kg^{-1}, or ~26% higher than the global RSE for a model trained with anthropogenic corrections (10.8 μmol kg^{-1}). This difference of 2.4 μmol kg^{-1} between the two approaches signifies the low impact of anthropogenic corrections in the models ability to constrain global C_T.

To objectively illustrate the importance of this anthropogenic correction we plotted the difference between non-corrected and corrected C_T models RSE values (Eq. D1) for data in each year spanning the 30 year measurement period (Fig. D1).

$$\Delta \mathrm{RSE}_{(yr)} = \mathrm{RSE}_{(yr)}(\text{not corrected}) - \mathrm{RSE}_{(yr)}(\text{corrected}) \qquad \text{(D1)}$$

where yr spans the global dataset year range (i.e. 1981-2010).

The positive and increasing $\Delta RSE_{(yr)}$ as year diverges from the reference year 2000 demonstrates that the enhanced skill of the model with anthropogenic C_T corrects is a direct result of removing anthropogenic biases. This result does not advocate that applied corrections were globally accurate, it simple confirms the importance of correcting C_T measurements to better constrain the global C_T system.

Supplement E: Robust forward MLR based on hypothesis tests

Following the schematic in figure E1, the routine initiates by ranking predictor variables $(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n, \ldots, \boldsymbol{p}_N)$ according to their degree of linear correlation to the response variable $(\boldsymbol{y})$, where $\boldsymbol{p}_{n,1}$ has the highest correlation. The primary model ($\boldsymbol{M_1}$) is then established by applying a least-squares multiple-linear regression (MLR) between variables $\boldsymbol{p}_{n,1}$ and $\boldsymbol{y}$ to constrain the regression coefficients β_0 and $\beta_{n,1}$. In step 3, the routine expands on $\boldsymbol{M_1}$ by modelling the top two correlation ranked predictor variables (m = 2), where m is the ranking of the modelled predictor variable with the lowest correlation to $\boldsymbol{y}$.

To determine if multi-collinearity (MCL) exists in the expanded model ($\boldsymbol{M}_m$), we calculate the variance inflation factor (VIF) for each modelled variable in $\boldsymbol{M}_m$ and compare them to VIF values calculated for the same variables modelled in $\boldsymbol{M}_{m-1}$. If the VIF value for any predictor variable $\boldsymbol{p}_{n,i}$ (where $i < m$) increased by 5, it indicates existence of MLC. To reduce influences of MLC, the model is updated with interaction terms between the newly added predictor variable $\left(\boldsymbol{p}_{n,m}\right)$ and any modelled variable that has a VIF increase greater than 5. To evaluate the significance of the newly added predictor variable and interaction terms, an analysis of variance (ANOVA) between the previous model ($\boldsymbol{M}_{m-1}$) and expanded model ($\boldsymbol{M}_m$) is applied. If the expanded model is found statistically better in constraining the system with a 95% confidence interval, the updates are accepted and the routine returns to step 3 where the model is again expanded with the next lowest correlation ranked predictor variable (i.e., $m = m + 1$).

If MCL is not detected, a null-hypothesis test based on the t-statistic is applied to determine if the coefficient of the new predictor variable $(\beta_{n,m})$ is significantly different from 0 (i.e., the new predictor is important in constraining the system). If it does not differ from 0 with a 95% confidence interval, the new predictor variable is defined as not significant and is subsequently rejected from the model before returning to step 3 to again expand on $\boldsymbol{M}_m$ with the next lowest ranked predictor variable.

Once each predictor variable has had a opportunity to update the model (i.e., $m = I$), any desired higher order variable terms are incorporate into the model on the provision the first order term was found to be statistically significant. The routine then prunes the model through an iterative process removing insignificant terms based on the t-test. Once all terms are statistically significant, the final stage of the routine applies a robust MLR to the set of significant terms to reduce potential influences of outliers.

Supplement F: Principal Component Regression

Principal Component Regression (PCR) is an empirical approach when multi-collinearity exists between predictor variables. The process (outlined in figure F1) first calculates the principal components $(\boldsymbol{n}_1,\ldots,\boldsymbol{n}_i,\ldots,\boldsymbol{n}_I)$ of the predictor variables $(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n,\ldots,\boldsymbol{p}_{\boldsymbol{N}})$. Then a least-squares multiple-linear regression is established between a subset of the principal components and the response variable ($\boldsymbol{y}$). The subsets begin with just the first principal component, then the first two, through to all principal components. The PCR deemed optimal is simply the regression with the lowest residual standard error (RSE).

Supplement G: Optimal MLR equations

Supplement H: Supervised SOM

A supervised form of the SOM that additionally incorporates response variable information in the clustering phase was first suggested by (Kohonen, 2001) and later developed by (Melssen et al., 2006). In their approach, a second neuron map of identical size to the predictor variable map established in Sect. 5.2 (wherein after referred to as the X-map) is constructed for the response variable (Y-map). Initialization of weights for the X-map remain identical to the un-supervised form, whilst the Y-map neurons are each randomly assigned a weight from within the response variable range.

Identification of the winning neuron in the X-map for data sample $\left(\boldsymbol{x}_p, y_p\right)$ is determined using a distance measure that uses both the X-map and the Y-map

$$j(\boldsymbol{x}_p, y_p) = \min_j \left(\left(1-\alpha(\tau)\right) \left[\sum_{n=1}^{N} \left(x_{p,n} - \omega_{j,n}\right)^2 \right]^{0.5} + \alpha(\tau) \left| y_p - \omega_{j(\text{Y-map})} \right| \right) \quad \text{(H1)}$$

where $0 < \alpha(\tau) < 1$ is responsible for regulating the relative weight of the similarity measures of the X and Y maps. By initially setting $\alpha(\tau)$ to 0.75, more weight is given to the neurons in the Y-map in adjusting the X-map. As $\alpha(\tau)$ reduces linearly with iteration to 0.5, both maps are given equal weighting in identifying the winning neuron. Once the winning neuron is established, the X-map weighting vectors are updated using the same approach as presented in Sect. 5.3.

For every iteration step (τ), each sample is presented to the SOM model twice. In the first pass the winning neuron in the X-map is determined and weighting vectors adjusted, whilst the second pass establishes the winning neuron in the Y-map using

$$j(\boldsymbol{x}_p, y_p) = \min_j \left(\alpha(\tau) \left[\sum_{n=1}^{N} \left(x_{p,n} - \omega_{j,n}\right)^2 \right]^{0.5} + \left(1-\alpha(\tau)\right) \left| y_p - \omega_{j(\text{Y-map})} \right| \right) \quad \text{(H2)}$$

and subsequently adjusts Y-map weighing numbers.

After the training phase is complete, response variable prediction using the X-map and any input data vector $(\boldsymbol{x}_q)$ is conducted in the same manner as presented in Sect. 5.5.

Supplement I: Geographical representation

Global position representation through latitude and longitude is problematic due to the mid-Pacific discontinuity of longitude at ±180°, and shortening of geographical distance between degrees of longitude towards the poles. As a measure to reduce the influence of discontinuity in the model, longitude values were shifted by 160° West (or 20° East), thereby setting the 180° discontinuity at a position that bisects continental Africa and Europe (Fig. H1). We also tested a three dimensional n-vector calculated from latitude and longitude that eliminates both these issues.

The normal vector to the Earth ellipsoid (n-vector), transforms the 2-D latitude/longitude position system into a 3-D vector, whilst maintaining unique vectors for every geographical position. Employing a version of the n-vector presented by (Gade, 2010), latitude and longitude values were transformed using

$$\boldsymbol{n} = \begin{bmatrix} \sin(\text{latitude}) \\ \sin(\text{longitude})\cos(\text{latitude}) \\ \cos(\text{longitude})\cos(\text{latitude}) \end{bmatrix} \quad \text{(I1)}$$

Supplement J: 14 region separation approach

Supplement K: Identifying poorly constrained coastal and marginal seas

Of the 425 combined C_T and A_T predicted measurements that have a residual error greater than ±50 μmol kg^{-1}, we found that 70% (298) were located within 300 km of a major coastline. Identifying that the coastal zone around New Zealand extends up to 345 km from the shore (Gibbs et al., 2006), these anomalous independently predicted measurements are likely a result of terrestrial influences perturbing the processes affecting the carbon and hydrographic concentrations.

To evaluate the appropriateness of identifying coastal data based on a bathymetric depth approach, RSE values were calculated for coastal (within 300 km of a major coastline) and open ocean zones using the global independent test predictions, however excluding the 298 measurements already identified as terrestrially influenced and data above 70°N (Table K1). The global models ability to capture open ocean A_T measurements is ~14% (or 1.5 μmol kg^{-1}) higher than for coastal measurements, and ~11% for C_T. This result suggests that identification of coastal measurements under a bathymetric depth approach may not be effective for ocean regions where coastal biogeochemical processes and terrestrial influences are not coupled to a shelf break, but may rather be dependent on biotic distributions. Future attempts to identify coastal measurements should therefore not solely rely on bathymetric depth.

Supplement L: Are the neurons capturing the system?

Optimal model configurations may be biased to the three independent datasets that constitute only 30% of the global data (see Sect. 6.1.1 Table 4). To ensure the SOM captures all important features of the global carbon system, and minimises the potential influence of grouping biases, the GIT approach was applied globally using the optimal model configuration and with an increase in the SOM neuron size (Table L1).

The independent test RSE values for data below 70°N increased by 0.1-0.4 μmol kg^{-1} for each step up in neuron map size (Table L1). This suggests that all important features were constrained using the three independent datasets, and that the optimal configurations remain valid on a global scale.

Supplement M: SOMLO model without Arctic measurements

Uniqueness of parameter concentrations in the Arctic region (above 70°N), in particular that of salinity due to intense freshening of the water body, results in classification of Arctic measurements into features that are near exclusive to the region (Figure M1). This observation suggests Arctic measurements have little influence in constraining the remaining system.

To test this hypothesis, we compared the skill of SOMLO models trained with and without Arctic Ocean data using the GIT approach (Table M1). The skill in capturing the global carbon systems below 70°N differed by 0.1% and 2% between the two C_T and A_T models respectively, confirming that Arctic data has very little influence in the models ability to constrain the global system. This result also suggests that no bias exists in comparing the skill of the global SOMLO model to the traditional MLR approach that excluded Arctic data in the regressions.

Supplement N: Stochastic nature of the SOM

Initialization of neuron weights in the SOM model is a stochastic process (See Sect. 5) and can therefore lead to results that are not reproducible. In this study, the influence of this facet is dampened due to small neuron to in-situ measurement ratios (1:430 for C_T), and 800 training iteration steps converging on similar distributions of measurements among neurons for every model under static conditions.

As a test to explore stochastic influences in our model, the three independent subsets (see Sect. 6.1.1 Table 4) were each predicted 100 times using models trained under optimal configurations and the resulting RSE values examined for reproducibility (Table N1). The very small 1st standard deviation of 0.2 μmol kg^{-1} (or 1.6%) around the mean RSE value for C_T demonstrates reproducibility in our SOMLO approach for constraining the carbon system and suggests a negligible influence from stochastic SOM initialization.

Acknowledgements

We would like to thank the developers of the SciPy software from which the 4D interpolations were constrained (Jones, E., Oliphant, T., Peterson, P., and others: SciPy: Open Source scientific tools for python, 2001, http://www.scipy.org/).

References

Amante, C., and Eakins, B. W.: ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis, NOAA Technical Memorandum NESDIS NGDC-24, 19 pp, March, 2009.

Bates, N. R.: Interannual variability of the oceanic CO_2 sink in the subtropical gyre of the North Atlantic Ocean over the last 2 decades, J. Geophys. Res., 112, C09013, DOI: 10.1029/2006jc003759, 2007.

Cotrim da Cunha, L., Buitenhuis, E. T., Le Quéré, C., Giraud, X., and Ludwig, W.: Potential impact of changes in river nutrient supply on global ocean biogeochemistry, Global Biogeochem. Cycles, 21, GB4007, DOI: 10.1029/2006gb002718, 2007.

de Boyer, M. C., Madec, G., Fischer, A. S., Lazar, A., and Iudicone, D.: Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, J. Geophys. Res., 109, C12003, DOI: 10.1029/2004jc002378, 2004.

Dickson, A. G., and Millero, F. J.: A comparison of the equilibrium constants for the dissociation of carbonic acid in seawater media, Deep Sea Research Part A. Oceanographic Research Papers, 34, 1733-1743, DOI: 10.1016/0198-0149(87)90021-5, 1987.

Egleston, E. S., Sabine, C. L., and Morel, F. M. M.: Revelle revisited: Buffer factors that quantify the response of ocean chemistry to changes in DIC and alkalinity, Global Biogeochem. Cycles, 24, GB1002, DOI: 10.1029/2008gb003407, 2010.

Gade, K.: A Non-singular Horizontal Position Representation, Journal of Navigation, 63, 395 - 417, DOI: 10.1017/S0373463309990415, 2010.

Gibbs, M. T., Hobday, A. J., Sanderson, B., and Hewitt, C. L.: Defining the seaward extent of New Zealand's coastal zone, Estuarine, Coastal and Shelf Science, 66, 240-254, DOI: 10.1016/j.ecss.2005.08.015, 2006.

Jickells, T. D.: Nutrient Biogeochemistry of the Coastal Zone, Science, 281, 217-222, DOI: 10.1126/science.281.5374.217 1998.

Kara, A. B., Rochford, P. A., and Hurlburt, H. E.: Mixed layer depth variability over the global ocean, J. Geophys. Res., 108(C3), 3079, DOI: 10.1029/2000jc000736, 2003.

Key, R. M., Kozyr, A., Sabine, C. L., Lee, K., Wanninkhof, R., Bullister, J. L., Feely, R. A., Millero, F. J., Mordy, C., and Peng, T. H.: A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP), Global Biogeochem. Cycles, 18, GB4031, DOI: 10.1029/2004gb002247, 2004.

Kohonen, T.: Self-Organizing Maps, 3 ed., Number 30 in Springer Series in Information Sciences, Springer-Verlag, Berlin, 2001.

Lee, K., Millero, F. J., Byrne, R. H., Feely, R. A., and Wanninkhof, R.: The recommended dissociation constants for carbonic acid in seawater, Geophys. Res. Lett., 27, 229-232, DOI: 10.1029/1999gl002345, 2000a.

Lee, K., Wanninkhof, R., Feely, R. A., Millero, F. J., and Peng, T. H.: Global relationships of total inorganic carbon with temperature and nitrate in surface seawater, Global Biogeochem. Cycles, 14, 979–994, DOI: 10.1029/1998GB001087, 2000b.

McNeil, B. I., Metzl, N., Key, R. M., Matear, R. J., and Corbiere, A.: An empirical estimate of the Southern Ocean air-sea CO_2 flux, Global Biogeochem. Cycles, 21, GB3011, DOI: 10.1029/2007gb002991, 2007.

Mehrbach, C., Culberson, C. H., Hawley, J. E., and Pytkowicz, R. M.: Measurement of the Apparent Dissociation Constants of Carbonic Acid in Seawater at Atmospheric Pressure, Limnology and Oceanography, 18, 897-907, 1973.

Melssen, W., Wehrens, R., and Buydens, L. M. C.: Supervised Kohonen networks for classification problems, Chemometrics and Intelligent Laboratory Systems, 83, 99-113, DOI: 10.1016/j.chemolab.2006.02.003, 2006.

Millero, F. J., Pierrot, D., Lee, K., Wanninkhof, R., Feely, R. A., Sabine, C. L., Key, R. M., and Takahashi, T.: Dissociation constants for carbonic acid determined from field measurements, Deep Sea Research Part I: Oceanographic Research Papers, 49, 1705-1723, DOI: 10.1016/s0967-0637(02)00093-6, 2002.

Pierrot, D., Lewis, E., and Wallace, D. W. R.: MS Excel Program Developed for CO_2 System Calculations, ORNL/CDIAC-105a. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tennessee., DOI: 10.3334/CDIAC/otg.CO2SYS_XLS_CDIAC105a, 2006.

Pierrot, D., Brown, P., Van Heuven, S., Tanhua, T., Schuster, U., Wanninkhof, R., and Key, R. M.: CARINA TCO_2 data in the Atlantic Ocean, Earth Syst. Sci. Data, 2, 177-187, DOI: 10.5194/essd-2-177-2010, 2010.

Seitzinger, S. P., Harrison, J. A., Dumont, E., Beusen, A. H. W., and Bouwman, A. F.: Sources and delivery of carbon, nitrogen, and phosphorus to the coastal zone: An overview of Global Nutrient Export from Watersheds (NEWS) models and their application, Global Biogeochem. Cycles, 19, GB4S01, DOI: 10.1029/2005gb002606, 2005.

Wanninkhof, R., Lewis, E., Feely, R. A., and Millero, F. J.: The optimal carbonate dissociation constants for determining surface water pCO_2 from alkalinity and total inorganic carbon, Marine Chemistry, 65, 291-301, DOI: 10.1016/s0304-4203(99)00021-3, 1999.

Winn, C. D., Li, Y.-H., Mackenzie, F. T., and Karl, D. M.: Rising surface ocean dissolved inorganic carbon at the Hawaii Ocean Time-series site, Marine Chemistry, 60, 33-47, DOI: 10.1016/s0304-4203(97)00085-6, 1998.

Table G1. Ad-hoc and universal C_T regression equations with interaction terms (Int.).

	Intercept	T	S	DO	N	Si	P	Int.	Int.	Int.	Int.
North Pacific summer	1066.9	–	24.16	0.38	5.07	–	–	–	–	–	–
North Pacific winter	868.73	-7.95	36.54	–	–	4.73	–	-0.01*Si*DO	–	–	–
Southern Ocean summer	698.74	–	35.84	0.25	–	0.42	83.28	–	–	–	–
Southern Ocean winter	1494.14	-48.81	22.3	-0.31	–	0.48	–	0.03*T*DO	0.92*T*S	–	–
Northwest Atlantic summer	1709.15	–	9.13	–	9.14	–	–	–	–	–	–
Northwest Atlantic winter	1128.79	-5.93	28.71	–	17.31	–	–	-0.05*N*DO	–	–	–
Northeast Atlantic summer	1625.17	–	12.35	–	6.13	–	–	–	–	–	–
Northeast Atlantic winter	1013.95	61.48	32.75	–	-61.93	-2.05	-12.54	0.4*N*Si	-0.3*T*DO	-1.67*T*S	1.78*N*S
Equatorial Pacific	467.55	-7.11	48.4	–	2.34	1.44	38.85	–	–	–	–
Sub-tropical North Pacific summer	519.77	-9.98	48.97	–	20.25	–	1.92	–	–	–	–
Sub-tropical North Pacific winter	236.82	-5.32	50.65	0.38	-1.5	–	139.22	–	–	–	–
Sub-tropical South Pacific summer	147.12	-4.61	52.63	0.34	7.48	1.67	72.68	–	–	–	–
Sub-tropical South Pacific winter	643.05	-12.01	46.68	–	–	-1.28	107.88	–	–	–	–
Indian Ocean summer	551.82	-6.59	45.21	–	–	–	27.16	0.14*S*N	–	–	–
Indian Ocean winter	1733.55	-1.84	–	-4.78	18.13	2.64	67.1	0.17*DO*S	–	–	–
Sub-tropical North Atlantic summer	619.34	-8.18	46.94	-0.37	-31.07	–	–	–	–	–	–
Sub-tropical North Atlantic winter	765	-7.36	39.89	–	–	-4.88	109.27	–	–	–	–
Equatorial Atlantic	163.5	-8.91	59.32	-0.17	–	–	–	–	–	–	–
Sub-tropical South Atlantic	2277.89	-6.15	–	-7.48	-5.08	–	74.92	0.2*DO*S	–	–	–
Universal model	596.77	-8.21	45.5	-0.17	1.12	0.45	17.83	0.01*T*DO	1.52*T*P	–	–

Table G2. Ad-hoc and universal A_T regression equations with interaction terms (Int.).

	Intercept	T	S	S^2	DO	Si	P	Int.	Int.	Int.
Sub-tropical Oceans	2064.66	-0.3	-47.57	1.54	0.13	-1.12	10.1	–	–	–
Equatorial Pacific	1142.6	-1.39	–	0.96	0.14	–	-3.51	–	–	–
North Atlantic	1543.52	-4.78	–	0.64	-0.04	-0.29	-9.04	0.13*S*T	–	–
North Pacific	721.6	–	44.31	–	0.09	-7.81	9.97	0.24*S*Si	–	–
Southern Ocean	7661.04	-1.46	-362.53	5.86	0.54	-12.17	-6.56	0.08*S*T	0.44*S*Si	-0.01*DO*Si
Global	1972.44	-12.78	-33.44	1.19	0.16	0.39	6.89	0.37*S*T	–	–

Table K1. Skill comparison between coastal and open ocean predicted measurements.

Model	RSE[a] (N[b]) Coastal	RSE[a] (N[b]) Open Ocean	% difference
C_T	11.9 (4338)	10.6 (18875)	10.9
A_T	10.4 (2856)	8.9 (14014)	14.4

[a] Residual Standard Error (µmol kg^{-1})

[b] N = number of in-situ measurements

Table L1. RSE values (μmol kg^{-1}) for models under optimal configurations and two increases in neuron map size.

	C_T model		A_T model	
	Number of neurons	RSE	Number of neurons	RSE
Optimal	64	12.45	25	9.78
Step 1	72	12.59	30	10.16
Step 2	81	12.82	36	10.28

Table M1. Independent test RSE values for data below 70°N.

	RSE (μmol kg^{-1})		
	Model with Arctic data	Model without Arctic data	% difference
C_T	12.45	12.44	0.1%
A_T	9.71	9.9	2%

Table N1. RSE results for stochastic initialization test.

	Mean RSE (μmol kg^{-1})	1^{st} Standard Deviation (μmol kg^{-1})	% of mean
C_T	12.2	0.2	1.6%
A_T	8.2	0.1	1.2%

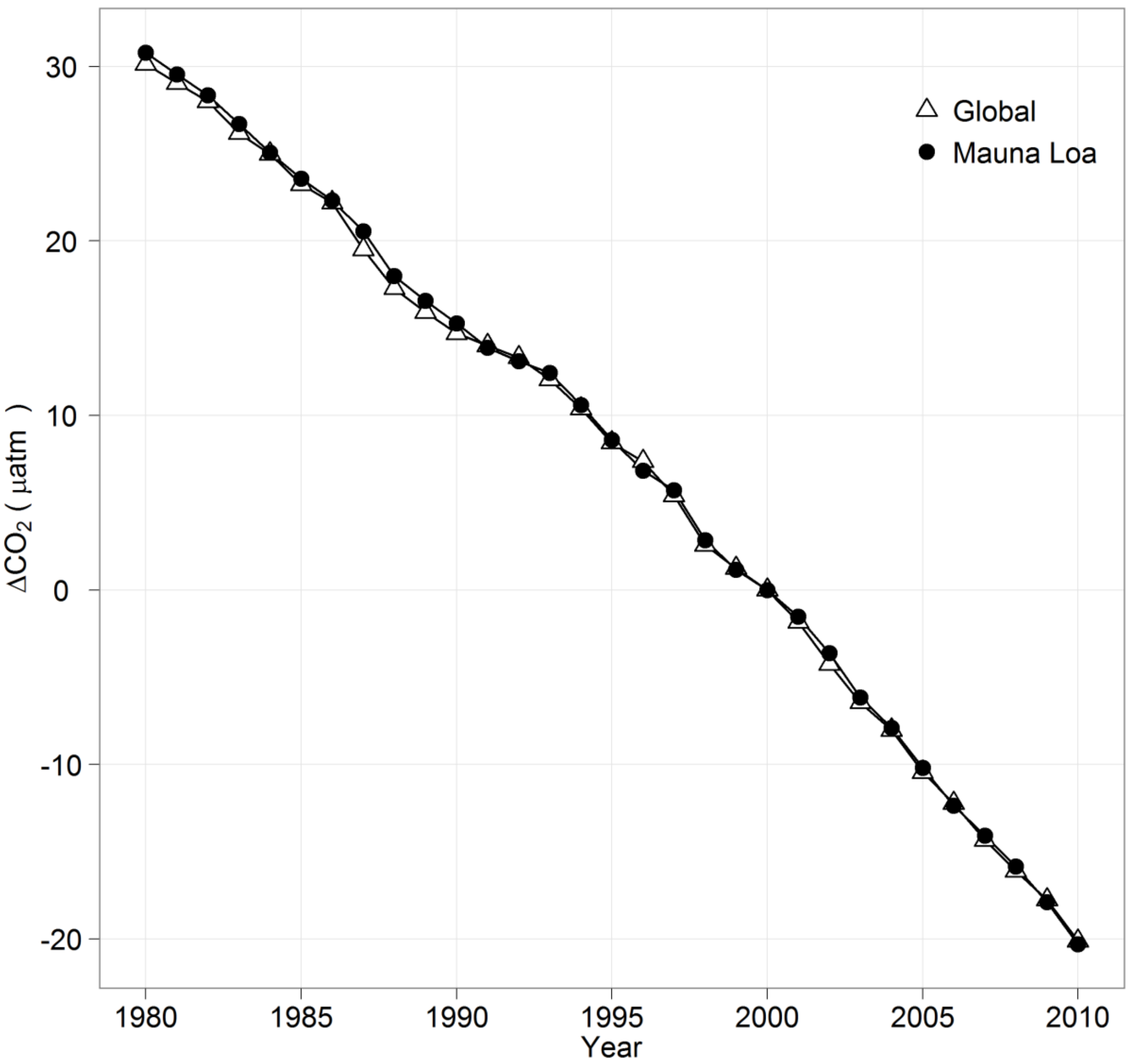

Fig. C1. Global and Mauna Loa site CO_2 difference between in-situ year and the year 2000.

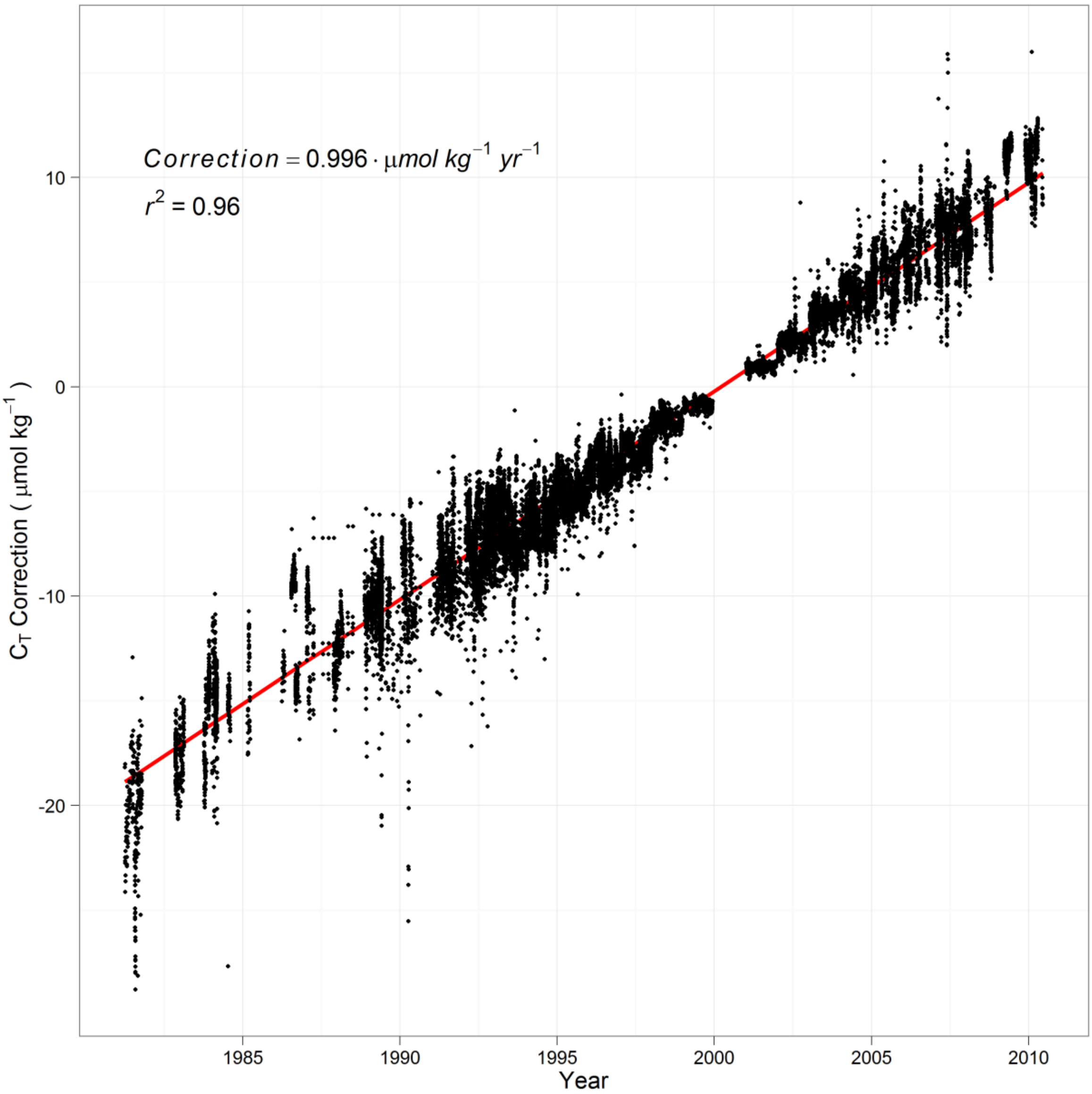

Fig. C2. Correction factor applied to C_T measurements defined by $C_T\text{correction} = C_{T(2000)} - C_{T(\text{in-situ year})}$

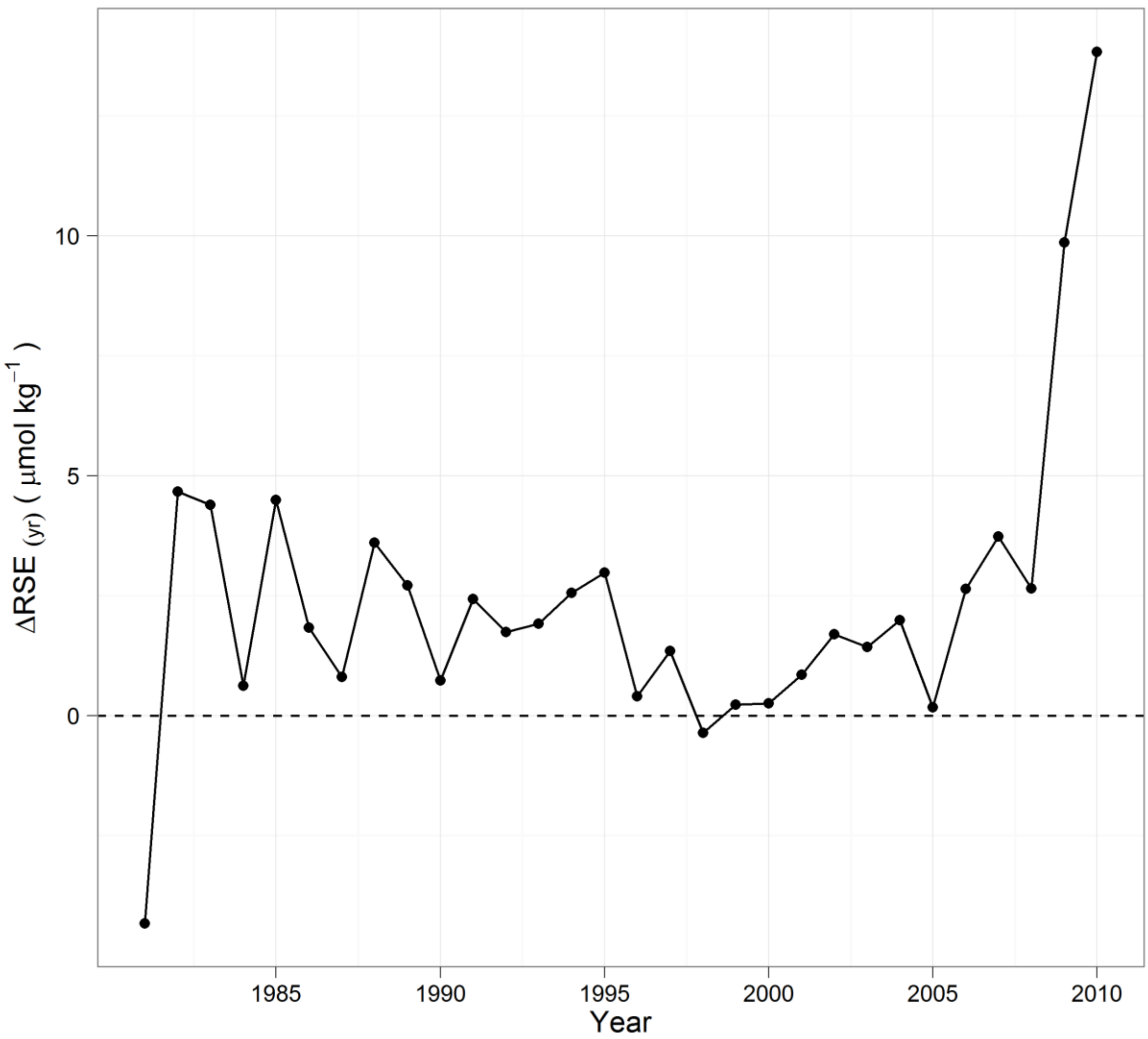

Fig. D1. Annual ΔRSE between C_T models trained with and without anthropogenic corrections.

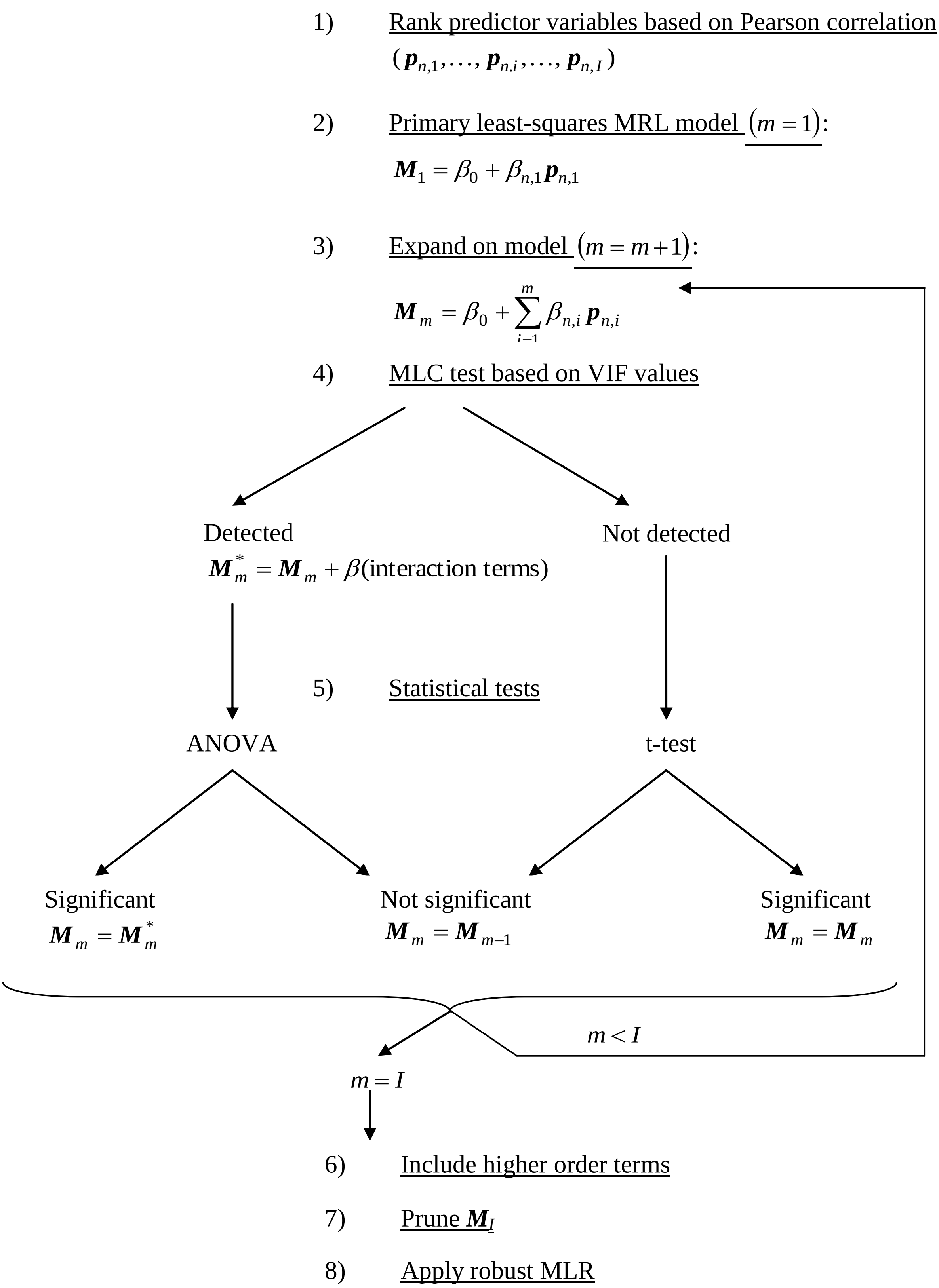

Fig. E1. Robust forward MLR routine schematic.

$$(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n, \ldots, \boldsymbol{p}_N)$$

$$\downarrow \text{ PCA}$$

$$(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_i, \ldots, \boldsymbol{n}_I)$$

$$\downarrow \text{ Constrain PCR's}$$

$$\boldsymbol{M}_1 = \beta_0 + \beta_1 \boldsymbol{n}_1$$

$$\vdots$$

$$\boldsymbol{M}_i = \beta_0 + \sum_{i=1}^{i} \beta_i \boldsymbol{n}_i$$

$$\vdots$$

$$\boldsymbol{M}_I = \beta_0 + \sum_{i=1}^{I} \beta_i \boldsymbol{n}_i$$

$$\downarrow$$

$$\boldsymbol{M}_{\text{optimal}} = \min_{\text{i}} (\text{RSE of } \boldsymbol{M}_i)$$

Fig. F1. Principle Component Regression schematic.

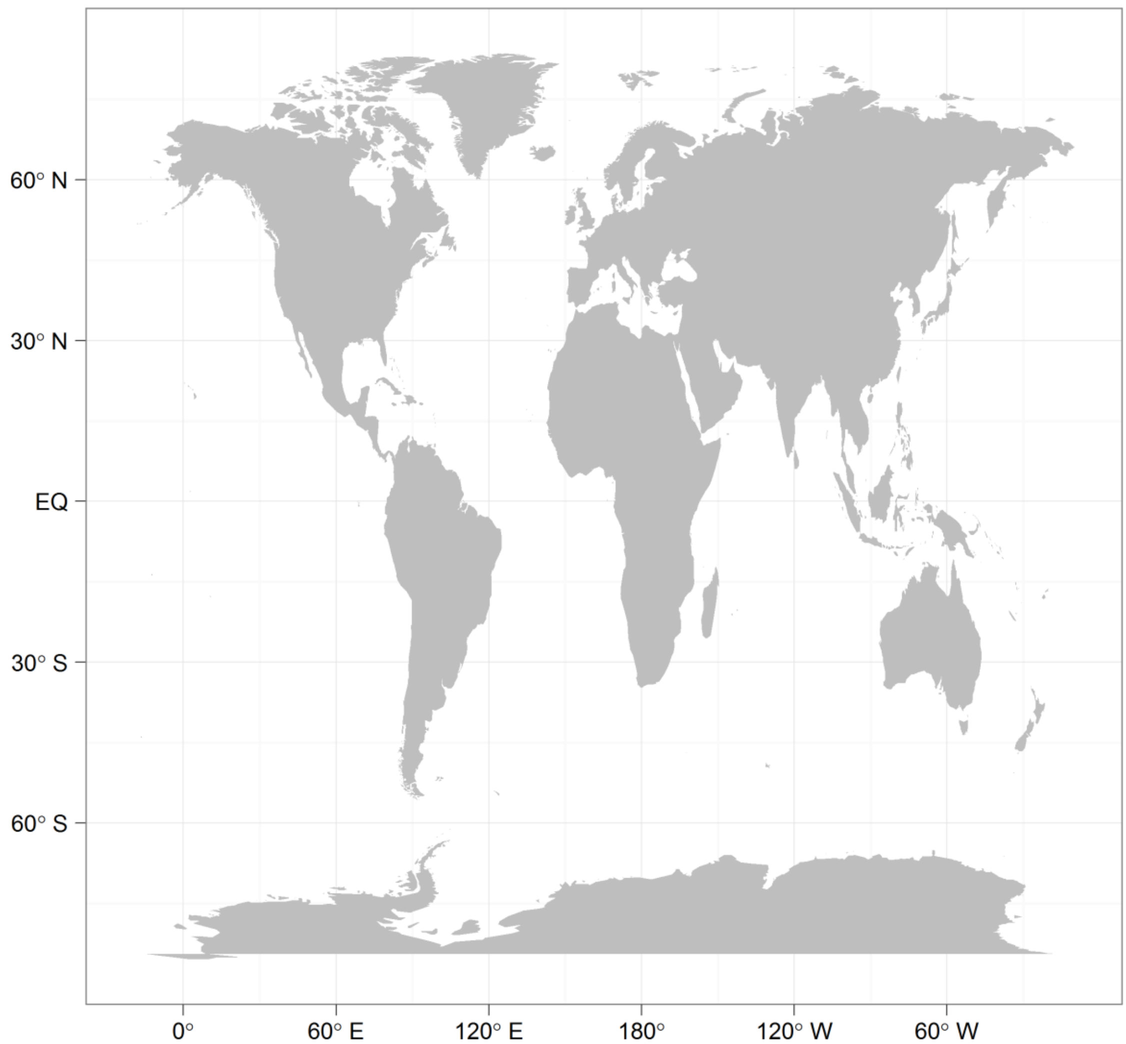

Fig. I1. Longitude values after reorganization.

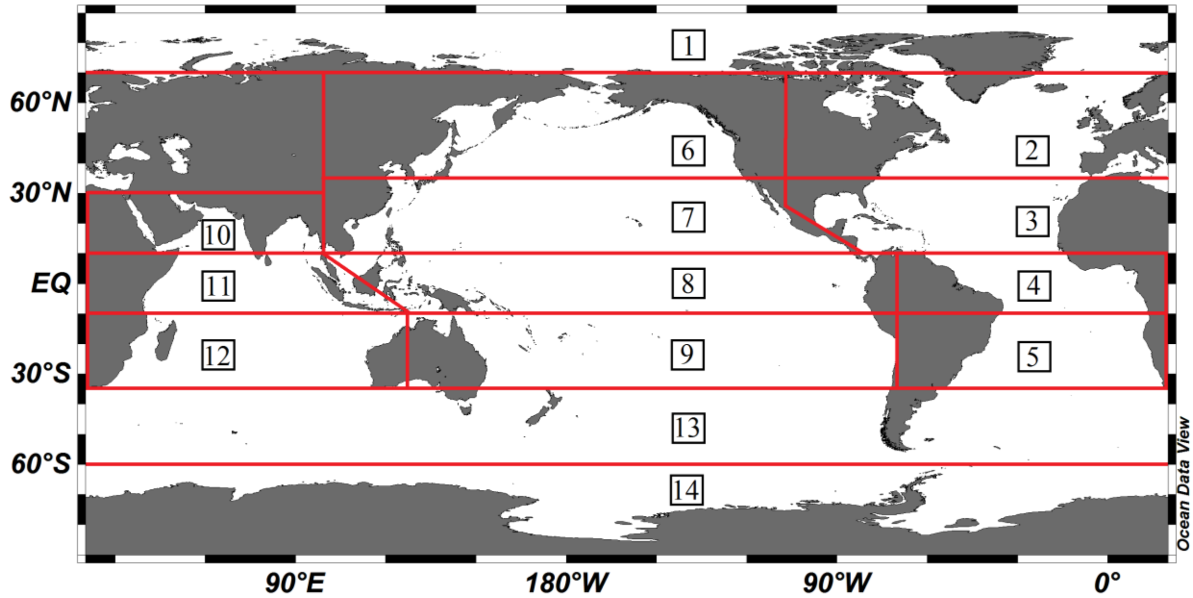

Fig. J1. Geographical separation of independently predicted dataset into 14 regions.

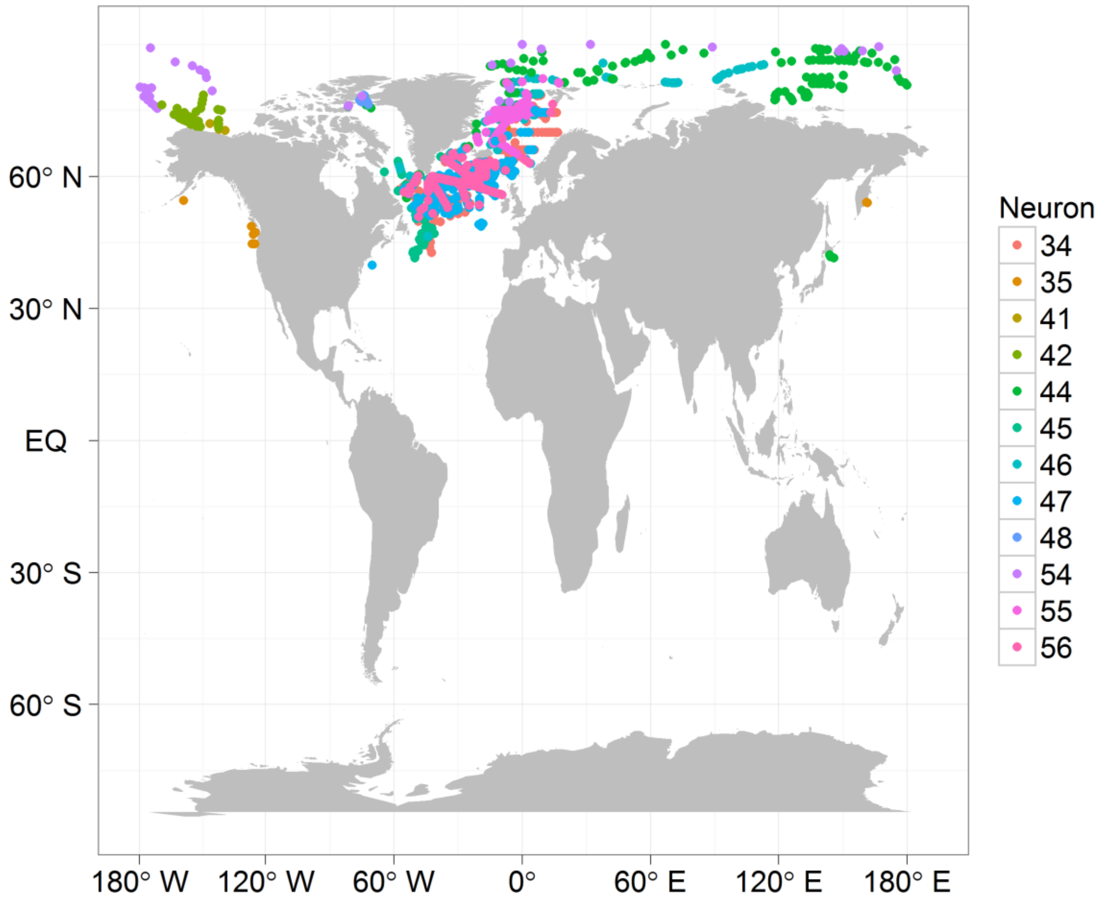

Fig. M1. Distribution of measurements assigned to a neuron containing at least one Arctic measurement.