1

Supplementary Materials

Acronym	Description
ΔpCO_2	ocean pCO_2 – atmospheric pCO_2
atm	atmospheric
BATS	Bermuda Atlantic Time-series Study
BIO23	Modified Fay and McKinley (2014) ocean CO_2 biomes
CARIOCA	CARbon Interface OCean Atmosphere
Chl-a	Chlorophyll-a
CSIR	Council for Scientific and Industrial Research
DIC	Dissolved Inorganic Carbon
EKE	Eddy kinetic energy
ERT	Extremely Randomised Trees
fCO_2	Fugacity of carbon dioxide
FCO_2	Sea-air CO_2 flux
FFN	Feed-Forward Neural-Network
GBM	Gradient Boosting Machine
GLODAP	GLobal Ocean Data Analysis Project
HOTS	Hawaii Ocean Time Series
IAV	Interannual Variability
IQRIA	Interquartile Range calculated over interannual variability
K21E	K-means configuration: 21 clusters - column E from Figure 5
LDEO	Lamont Doherty Earth Observatory
MLD	Mixed-layer depth
OSSE	Observing system simulation experiment
pCO_2	Partial pressure of carbon dioxide
PgC yr^{-1}	10^{15} grams of carbon per year
Riav	Relative interannual variability
RMSE	Root-mean-square error
SOCAT	Surface Ocean CO_2 ATlas
SOCCOM	Southern Ocean Carbon and Climate Observations and Modeling
SOCOM	Surface Ocean CO_2 Mapping
SSS	Sea surface salinity
SST	Sea surface temperature
SVR	Support Vector Regression
TA	Total Alkalinity

S1 Description of clustering and regression features

Here we present the products and data processing steps associated with the various features (clustering and regression) used throughout our analysis. Specifically, we use:

- Sea surface temperature (SST), the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) daily, quarter degree product from the Group for High-Resolution Sea Surface Temperature (GHRSST), which combines satellite and in-situ data (Donlon et al. 2012). The SST anomaly is used as a derived feature in our analyses. The annual mean for each year is subtracted from the SST product, leaving the variability around the mean for that year. We include this metric as it is a measure of intraseasonal variability of SST.
- Sea ice fraction (ICE), the estimates provided from the OSTIA monthly product (Donlon et al. 2012).
- Sea surface salinity (SSS), the EN4 monthly product which performs an objective analysis of ship-based observations (Good et al. 2013).
- atmospheric pCO_2 (pCO_2^{atm}), a product derived from atmospheric mole fraction of CO_2 (xCO_2) measurements gathered in ObsPack v3 (Masarie et al. 2014). To generate a gridded pCO_2^{atm} monthly product from atmospheric xCO_2 sea surface and flask measurements, the atmospheric xCO_2 measurements were: (1) averaged along equal latitude (assuming that xCO_2 is well-mixed across longitudes), (2) linearly interpolated to fill latitudinal gaps, and (3) extrapolated longitudinally to create a global latitudinally-varying time-series of xCO_2. Finally, pCO_2^{atm} was calculated using the monthly gridded atmospheric xCO_2 and the monthly sea level pressure from the ERA-Interim 2 reanalysis product (Dee et al., 2011), using Equation 1 from Dickson et al. (2007).
- mixed layer depth (MLD), the Argo Mixed Layers monthly product generated from Argo density profiles (Holte et al., 2017). Precisely, we use the $\log_{10}$ transformation onto the Holte et al. (2017) MLD product and create a monthly climatology, thus imposing the assumption that there is no interannual variability.
- chlorophyll-a (Chl-a), the Globcolour monthly product (Maritorena et al. 2010). Two features were created for this variable: Chl-a and Chl-a′ (Table 1). First, the Chl-a feature was created by applying the $\log_{10}$ transformation. The Globcolour satellite product is only available from 1998. The climatology of Chl-a (1998-2016) was used to fill the period before 1998 and remaining cloud gaps. Low concentration random noise was inserted in high-latitude winter regions (areas and season for which there is not Chl-a climatology). A Chl-a feature-variable was generated as an anomaly product (Chl-a′ - also $\log_{10}$), which was calculated by subtracting the climatology (calculated using chlorophyll-a data from 1998-2016) from the satellite product.
- wind vectors (u and v) and speed (U_{10}, the 6-hourly ERA-interim version 2 product (Dee et al., 2011). Wind speed was calculated for each 6-hourly time step using the equation in Table 1 and was then averaged into monthly means.

- eddy kinetic energy (EKE), u and v surface current components (integrated for depth < 15 m) of the monthly Globcurrent product (Rio et al., 2014). Specifically, EKE was calculated using the equation in Table 1, where u' is calculated as $\overline{u} - u$ and similarly with v.
- surface ocean pCO_2 climatology, the Lamont-Doherty Earth Observatory (LDEO) product referenced to the year 2000 (Takahashi et al., 2009).

In summary, all the above feature datasets were generated at the global scale, at a monthly frequency from 1982 to 2016 (except for climatological products) and onto a 1° × 1° resolution grid (specifically following the SOCAT v5 grid for consistency purposes). The gridded step was achieved using the Pandas and xarray packages in Python (McKinney, 2010; Hoyer and Hamman, 2017). Note that the regridding of the 4° × 5° pCO_2^{clim} also involved a moving average convolution wind of 5° × 5°) to smooth the data.

S2 Descriptions of regression methods

S2.1 Shuffled train-test experiment

An experiment was performed to assess the difference in the root-mean-square error when the train-test split was shuffled, vs using random years as the splitting criteria. The exact same training procedure was applied to the model as done in Section 2 of the main article. The train-test shuffled split (0.8: 0.2) uses a random subset of the data without preserving order in any way. Importantly this means that cruise tracks are split. The RMSE was calculated using the test split for each year and then averaged. The RMSE scores for Extremely Randomised Trees were 12.35 µatm and 18.62 µatm for the shuffled and year-grouped splits respectively. This effect would be larger in methods that are prone to overfitting. However, as a precaution, we recommend that this train-test split procedure is applied.

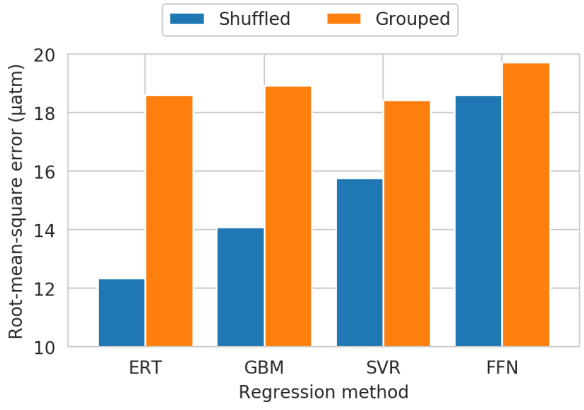

Figure S1: demonstrates that all machine learning methods used in this study suffer from overtraining; however, the FFN method is least prone to the effect, but has the lowest RMSE. The ERT is most prone, with GBM and SVR also suffering from the effect.

S2.2 K-means clustering

The basic principle of K-means clustering is as follows: 1) place N random centroids, where N is the number of clusters; 2) find the nearest centroid (usually with Euclidean distance) and assign samples to the respective clusters; 3) compute new centroids by calculating the mean position of all the samples belonging to a cluster; 4)

repeat steps 2 and 3 until there is no change in the membership of samples to clusters (Hastie et al. 2009). Mini-batch K-means applies the same principle, but data are split into batches to reduce the computational cost for large datasets with minimal loss of performance (Sculley 2010).

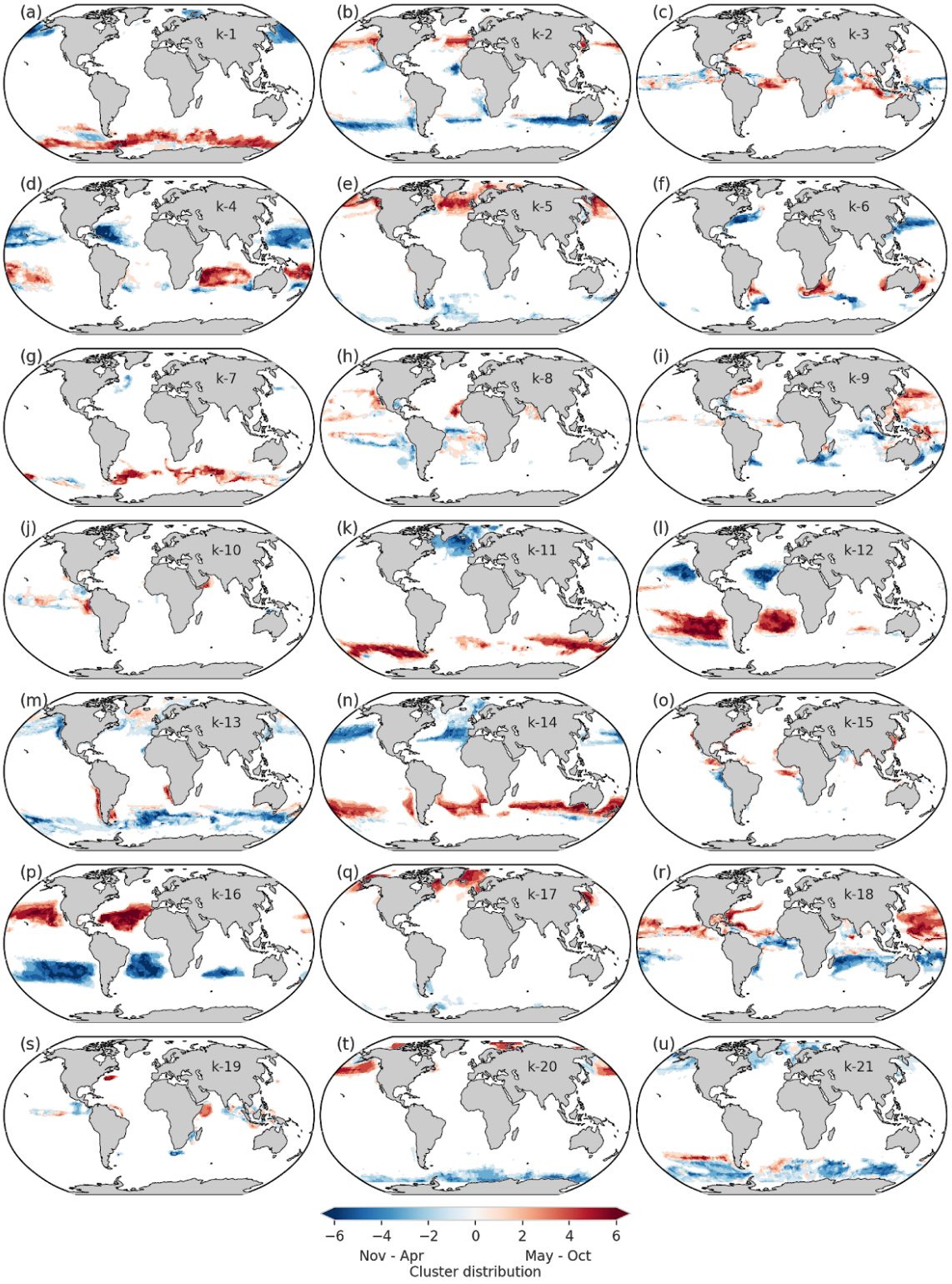

Figure S2: The membership of clusters (as a climatology) for the K21E configuration (Figure 5), where each panel represents a cluster with the number shown on the figure. The blue indicates that data is dominant in November to April period and red shows when data is dominant for the May to October period.

S2.3 Supervised regression methods

S2.3.1 Extremely Randomised Trees

Extremely Randomised Trees (ERT) is a derivative of Random Forest Regression (Breiman, 2001; Geurts et al. 2006). ERT fits multiple decision trees to a dataset, where a decision tree uses recursive partitioning of the target data based on splitting criteria in feature-variables. The use of multiple trees reduces the high variability that decision trees typically suffer from while theoretically maintaining low bias. However, Random Forest Regression is often prone to overfitting to the training data (Gregor et al. 2017). ERT reduces this by using the best of a selection of random cut-points when a decision tree is being trained. We use the Scikit-Learn implementation of ERT in Python (Pedregosa et al. 2012). The primary optimisation functions are the number of trees, the minimum number of observations at the terminal branches, and the number of random features in the subset of features from which each decision tree is trained.

S2.3.2 Gradient Boosting Machines

Gradient boosting machines (GBM) use multiple weak learners (typically decision trees) that are sequentially fitted to minimise the residuals of the previous fit (Friedman, 2001). This is known as additive learning, where the algorithm fixes what is learnt. While GBM's have been proven to be good at dealing with imbalanced datasets, it is more likely to overfit to the training data as the model has the potential for high complexity (Dietterich, 1995; Frery et al. 2017). Tuning the hyper-parameters to prevent overfitting is thus critical. As such, the following hyper-parameters were tuned in our study: number of trees (determined by improvement threshold), depth of trees (by adjusting the maximum depth of trees and minimum number of points per node) and learning rate. We use the XGBoost Python package which has a parallel implementation of GBM (Chen 2016).

S2.3.3 Feed-Forward Neural-Network

Feed-Forward Neural-Networks (FFN) is the most commonly used non-linear approach in Rödenbeck et al. (2015). We use the Multi-layer perceptron function in Scikit-Learn – a.k.a. FFN. The principle is that a network with random weights is generated (similar to the coefficients in linear regression). Samples are passed forward through the network to estimate target values. The discrepancy between the estimates and the targets is back-propagated through the weights until the targets are met with sufficient accuracy. The primary tuning parameter in the FFN is the architecture of the network (number of hidden layers and weights per layer). We follow the same procedure in determining the number of weights as Landschützer et al. (2013) where the size of the network can be up to $\frac{n}{30}$ where n is the number of samples in the training subset (Amari et al. 1997). We also tune the learning rate (α) – an α that is too small could result that the model gets stuck in a local minimum.

S2.3.4 Support Vector Regression

Support Vector Regression (SVR) applied with a Gaussian kernel is analogous to an FFN (Drucker et al. 1997; Romero and Toppo, 2007). The difference is that the SVR estimates the complexity of the problem from the data

using robust statistics giving the number of support vectors – the subset points that determine the hyperplane on which estimates lie. The theory of SVR is described in Gregor et al. (2017). While the performance of SVR is often good, it does not scale well to large datasets. However, the two-step, cluster–regression approach reduces the size of the problem drastically, making it possible to use SVR at monthly by one-degree resolution. We use SVR implementation in the Scikit-Learn package (Pedregosa et al. 2012). We standardise the features before implementing SVR with $\frac{x-\bar{x}}{\sigma}$, where $\bar{x}$ is the average of x and σ the standard deviation of x. There are two hyperparameters that we tune: C – controls the total allowable error (relative to the size of the margins), and γ – radius of the Gaussian.

S3 Results and Discussion

S3.1 Explanation of the increase in RMSE around the year 2002

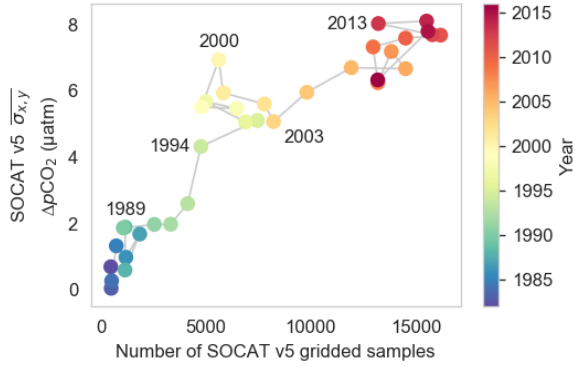

Figure S3: The number of SOCAT v5 monthly gridded data per year plotted against the standard deviation for that year, with the years shown by colour. An increase in standard deviation against the number of samples is observed around the year 2000 coinciding with the increase in RMSE for all methods during the same period.

S3.2 Ensemble member evaluation

In this section, we assess the evaluation of the ensemble members (Figure S4) and we also compare three different combinations of ensemble members (Table S1). There is a stronger bias in the members belonging to the BIO23 clustering configuration (Figure S4e-h), but including these methods results in a lower bias and RMSE score (Table S1).

Table S1: Bias and root-mean-squared error (RMSE) for three different ensemble member configurations. These configurations were tried in light of the fact that the BIO23 configurations have large biases. However, the inclusion of BIO23 regressions (with the exception of ERT) improves the overall bias and RMSE.

Ensemble	Bias (µatm)	RMSE (µatm)
K21E: SVR + FFN + GBM	-0.38	18.02
K21E: SVR + FFN + GBM + ERT	-0.30	17.97
ML6 (as in Section 3.2)	0.21	17.31

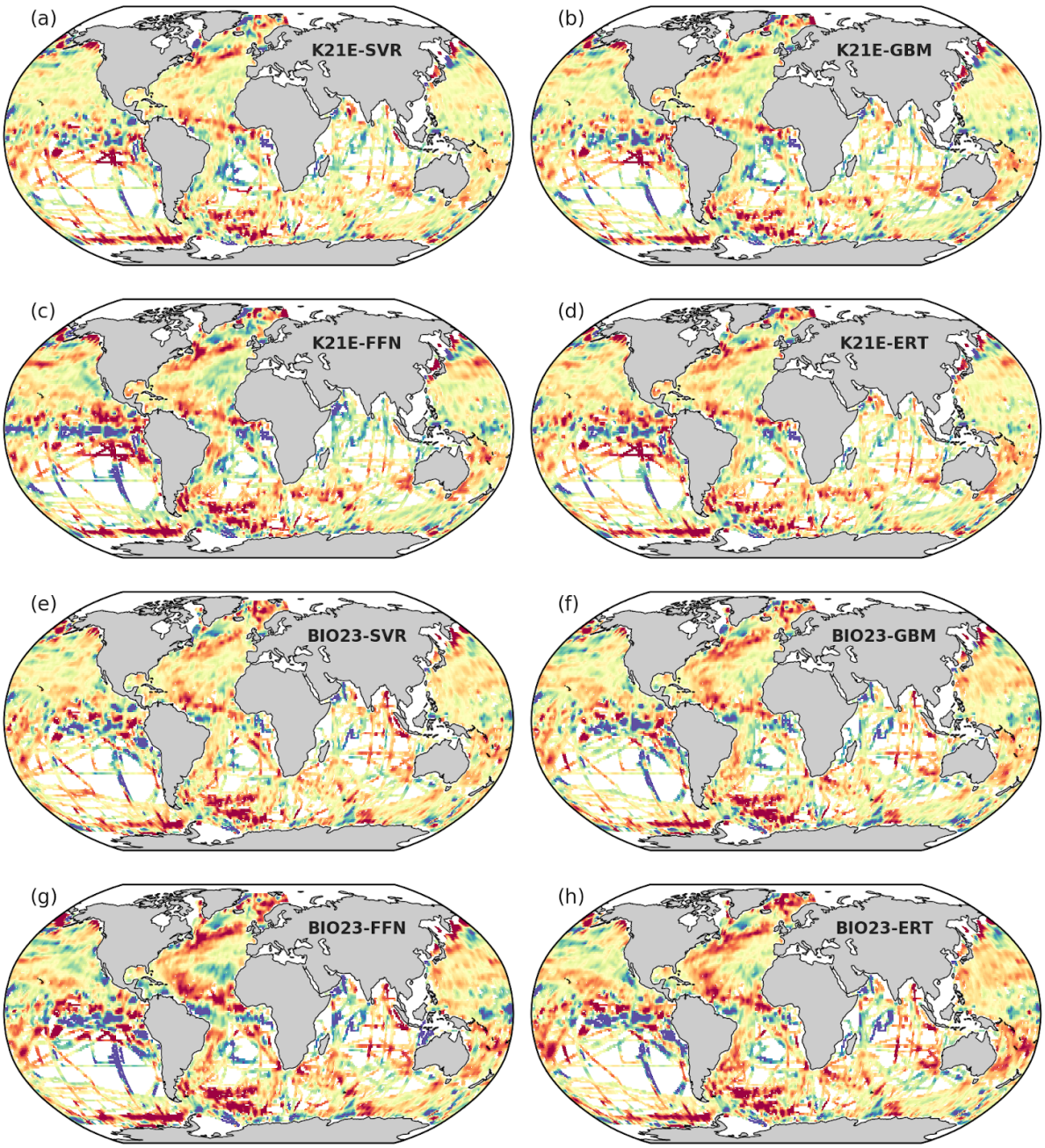

Figure S4: The biases from the robust test-estimates for the four regression methods in the K21E cluster (a-d) and similarly the four regression methods in the BIO23 cluster (e-h). See the main text for details about the clusters. The abbreviations for the regression methods are Support Vector Regression (SVR), Gradient Boosting Machine (GBM), Feed-Forward Neural-Network (FFN), and Extremely Randomised Trees (ERT). A convolution has been applied to make it easier to see the regional nature of the biases and RMSE. This is a partner figure to Figure 7a shows the bias for every ensemble member.

S3.3 Explanation for high RMSE in Taylor diagrams vs. annually averaged RMSE

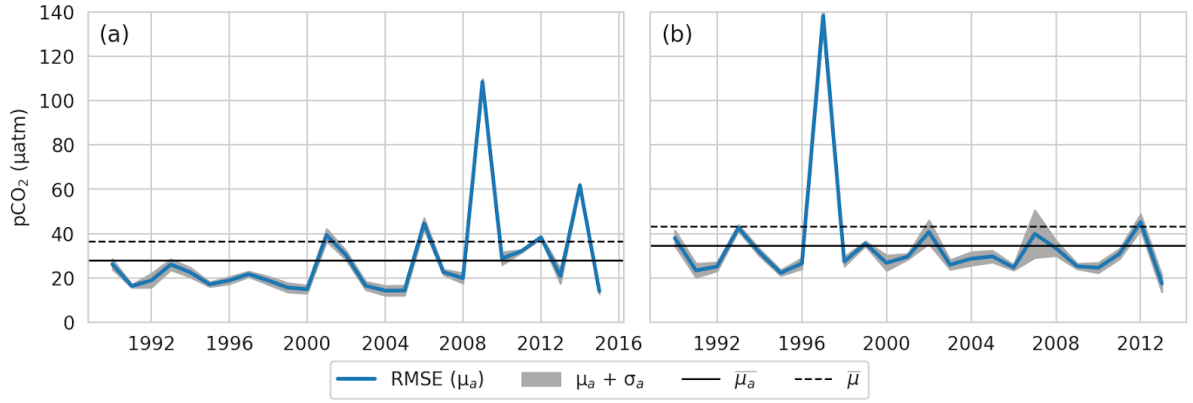

Figure S5: The annually calculated RMSE scores for LDEO (a) and GLODAP v2 (b) averaged for all gap-filling methods (blue line), with the grey filled area showing the standard deviation between methods. The solid black line shows the average of the annually calculated RMSE as shown in Table 5, while the dashed black line shows the RMSE without the annual weighting.

S3.4 Relative area and contributions of oceanic regions to FCO_2

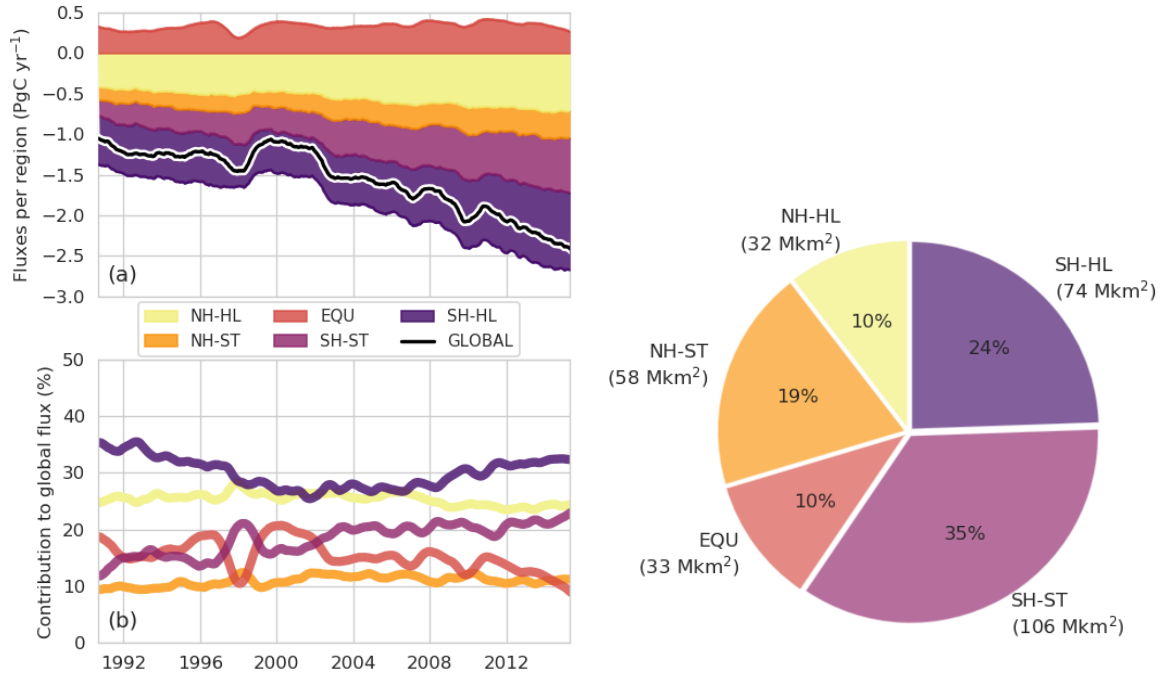

Figure S6: (a) A stacked area plot showing the magnitude of sea-air CO_2 fluxes for each region with a 12-month rolling mean and (b) shows the relative contribution of these regions to the total flux. (c) A pie chart showing the relative area of each ocean region in Figure 2.

S3.5 Relative importance of feature-variables for Gradient Boosting Machines

Gradient boosting machines (GBM) are able to estimate the relative importance of a feature-variable by the iterative nature and stacking of decision trees – available in the XGBoost package. This approach is useful when

applying the gradient boosting machines to regions that do not change with time, such as the CO_2 biomes used in our study, (Figure S7).

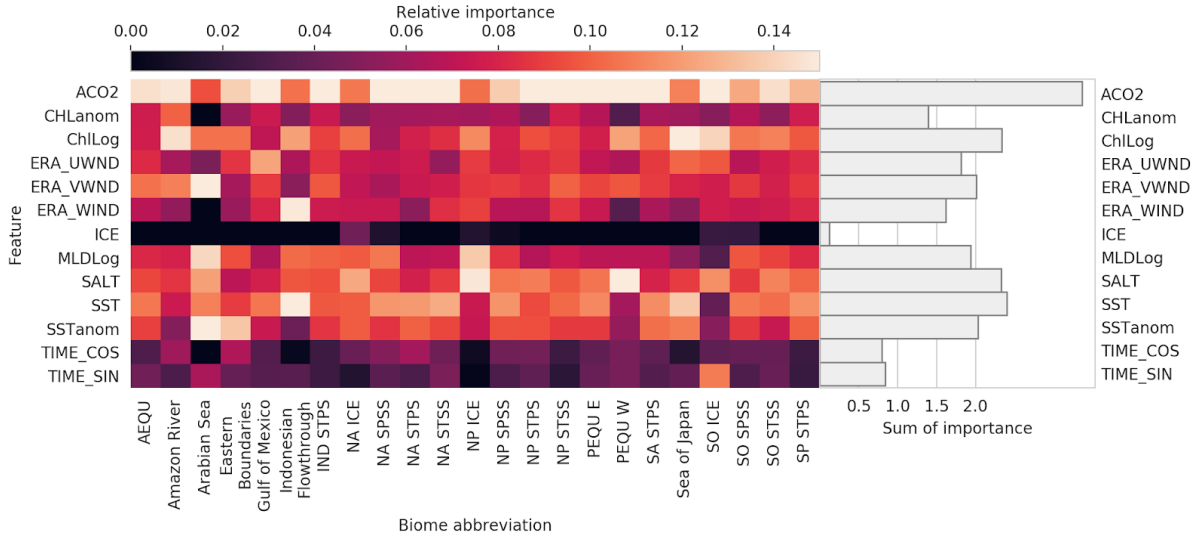

Figure S7: The feature-importances for a gradient boosting machine (GBM) run with the two-step cluster-regression, where the modified Fay and McKinley (2014) biomes (Figure 2) were used as clusters (x-axis). Atmospheric pCO_2 was omitted from the figure as the variable dominates the importance and thus skews the colour-map. The sum of each column is one. The figure on the right shows the sum of the rows, indicating the total importance of feature-variables.

References

Amari, S., Murata, N., Finke, M. and Yang, H. H.: Asymptotic Statistical Theory of Overtraining, , 8(5), 985–996, 1997.

Breiman, L.: Random forests, Mach. Learn., 45(1), 5–32, doi:10.1023/A:1010933404324, 2001.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, pp. 785–794, ACM Press, New York, New York, USA., 2016.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., Vitart, F., Hólm, E. V, Kållberg, P. and Thépaut, J. N.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828, 2011.

Dickson, A. G., Sabine, C. L. and Christian, J. R., Eds.: Guide to Best Practices for Ocean CO2 Measurements., 2007.

Dietterich, T. G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, Mach. Learn., 40(2), 139–157, doi:10.1023/A:1007607513941, 2000.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E. and Wimmer, W.: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system, Remote Sens. Environ., 116, 140–158, doi:10.1016/j.rse.2010.10.017, 2012.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V. N.: Support vector regression machines, Adv. Neural Inf. Process. Syst. 9, 1, 155–161, doi:10.1.1.10.4845, 1997.

Fay, A. R. and McKinley, G. A.: Global open-ocean biomes: Mean and temporal variability, Earth Syst. Sci. Data, 6(2), 273–284, doi:10.5194/essd-6-273-2014, 2014.

Frery, J., Habrard, A., Sebban, M., Caelen, O. and He-Guelton, L.: Efficient Top Rank Optimization with Gradient Boosting for Supervised Anomaly Detection, Lect. Notes Comput. Sci., 10534 LNAI, 20–35, doi:10.1007/978-3-319-71249-9_2,

2017.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine, Ann. Stat., 29(5), 1189–1232, doi:10.1214/aos/1013203451, 2001.

Geurts, P., Ernst, D. and Wehenkel, L.: Extremely randomized trees, Mach. Learn., 63(1), 3–42, doi:10.1007/s10994-006-6226-1, 2006.

Good, S. A., Martin, M. J. and Rayner, N. A.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, J. Geophys. Res. Ocean., 118(12), 6704–6716, doi:10.1002/2013JC009067, 2013.

Gregor, L., Kok, S. and Monteiro, P. M. S.: Empirical methods for the estimation of Southern Ocean CO2: support vector and random forest regression, Biogeosciences, 14(23), 5551–5569, doi:10.5194/bg-14-5551-2017, 2017.

Hastie, T., Tibshirani, R. and Friedman, J. H.: The Elements of Statistical Learning: Data mining, Inference, and Prediction, Second Edi., Springer., 2009.

Holte, J., Talley, L. D., Gilson, J. and Roemmich, D.: An Argo mixed layer climatology and database, Geophys. Res. Lett., 44(11), 5618–5626, doi:10.1002/2017GL073426, 2017.

Hoyer, S. and Hamman, J. J.: xarray: N-D labeled Arrays and Datasets in Python, J. Open Res. Softw., 5, 1–6, doi:10.5334/jors.148, 2017.

Landschützer, P., Gruber, N., Bakker, D. C. E., Schuster, U., Nakaoka, S., Payne, M. R., Sasse, T. P. and Zeng, J.: A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink, Biogeosciences, 10(11), 7793–7815, doi:10.5194/bg-10-7793-2013, 2013.

Maritorena, S., Fanton D'andon, O. H., Mangin, A. and Siegel, D. A.: Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues, Remote Sens. Environ., 114, 1791–1804, doi:10.1016/j.rse.2010.04.002, 2010.

Masarie, K. A., Peters, W., Jacobson, A. R. and Tans, P. P.: ObsPack: A framework for the preparation, delivery, and attribution of atmospheric greenhouse gas measurements, Earth Syst. Sci. Data, 6(2), 375–384, doi:10.5194/essd-6-375-2014, 2014.

Mckinney, W.: Data Structures for Statistical Computing in Python. [online] Available from: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf (Accessed 16 February 2019), 2010.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, C., Thirion, B., Grisel, O., Blondel, M., Prettenhoffer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. and Cournapeau, D.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830, doi:10.1007/s13398-014-0173-7.2, 2011.

Rio, M.-H., Mulet, S. and Picot, N.: Beyond GOCE for the ocean circulation estimate: Synergetic use of altimetry, gravimetry, and in situ data provides new insight into geostrophic and Ekman currents, Geophys. Res. Lett., 41(24), 8918–8925, doi:10.1002/2014GL061773, 2014.

Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., Landschützer, P., Metzl, N., Nakaoka, S., Olsen, A., Park, G.-H., Peylin, P., Rodgers, K. B., Sasse, T. P., Schuster, U., Shutler, J. D., Valsala, V., Wanninkhof, R. and Zeng, J.: Data-based estimates of the ocean carbon sink variability – first results of the Surface Ocean pCO2 Mapping intercomparison (SOCOM), Biogeosciences, 12(23), 7251–7278, doi:10.5194/bg-12-7251-2015, 2015.

Romero, E. and Toppo, D.: Comparing support vector machines and feedforward neural networks with similar hidden-layer weights, IEEE Trans. Neural Networks, 18(3), 959–963, doi:10.1109/TNN.2007.891656, 2007.

Sculley, D. and D.: Web-scale k-means clustering, in Proceedings of the 19th international conference on World wide web - WWW '10, p. 1177, ACM Press, New York, New York, USA., 2010.

Takahashi, T. T., Sutherland, S. C., Wanninkhof, R. H., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G. E., Chavez, F. P., Sabine, C. L., Watson, A. J., Bakker, D. C. E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T. S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R. G. J., Wong, C. S., Delille, B., Bates, N. R. and de Baar, H. J. W.: Climatological mean and decadal change in surface ocean pCO2, and net sea–air CO2 flux over the global oceans, Deep. Res. Part II Top. Stud. Oceanogr., 56(8–10), 554–577, doi:10.1016/j.dsr2.2008.12.009, 2009.