MOCCA

a Multi-sourced Ocean Carbonate Chemistry Analysis

or

a story about getting the right answer for the wrong reason

Friedrich Burger

June 16, 2024

Jupyter notebooks with the analyses as well as model weights can be found in the repository `https://github.com/friedrichs-repo/MOCCA/`.

Contents

1 Introduction

Deep learning has been heavily used in climate science in recent years (Reichstein et al., 2019), with applications ranging from forecasting, statistical down-scaling, pattern identification, process parameterization, emulation of physical models, to data interpolation. The success of deep learning in such data-driven applications is based on the versatility of neural networks Goodfellow et al. (2016). In theory, these can be used to learn any functional relationship between predictors and variables of interest, as reflected by the universal approximation theorem Hornik et al. (1989).

A particularly important application of deep learning is the interpolation of sparse ship-based measurements of the fugacity of carbon dioxide (fCO_2) in the surface ocean. A globally consistent and complete field of fCO_2 is necessary to estimate the air-sea CO_2 flux and thus necessary to estimate the fraction of anthropogenic carbon emissions that is taken up by the ocean. The global carbon budget (Friedlingstein et al., 2023), an annual assessment of global carbon emissions and sinks, currently estimates the oceanic sink from seven observation-based products. While based on very similar underlying fCO2 data and similar mostly satellite-based predictors, these seven products follow different approaches for interpolating the sparse fCO_2 data.

Four of these products are based on feed-forward neural networks: The CMEMS-LSCE-FFNNv2 (Chau et al., 2022) utilizes a 100-member neural network ensemble, bootstrapping from the months before and after a fCO_2 measurements and leaving the months with fCO_2 measurements for independent evaluation. The MPI-SOMFFN (Landschützer et al., 2016) builds on a two step procedure, where first different clusters of similar ocean conditions are determined using a self-organizing map approach and then neural networks are trained to predict fCO_2 in each cluster separately. Similarly, OS-ETHZ-GRaCER (Gregor and Gruber, 2021) provides an ensemble of varying cluster assignments with neural-network-based fCO_2 regression in each cluster. NIES-ML3 (Zeng et al., 2022) is based on three model estimates, from a random forest, a gradient boost machine, and a feed-forward neural network. The remaining three observation-based products build on multiple linear regressions

for A_T and C_T (fundamental variables to calculate fCO_2 and other carbonate system variables; JMA-MLR; Iida et al., 2021), extreme gradient boosting to predict the missfit between global ocean biogeochemical models and fCO_2 measurements (LDEO-HPD; Gloege et al., 2022), and a autoregressive multiple linear regression approach (Jena-MLS; Rödenbeck et al., 2022).

The largest uncertainty in these spatially and temporally interpolated fCO_2 fields roots in the sparsity and uneven distribution of the underlying fCO_2 measurements that are collected in the Surface Ocean CO_2 Atlas (Bakker et al., 2016). In particular, measurements are sparse in high-latitude regions and particularly in the Southern Ocean. One approach to soften this issue is applying neural-network based regression seprately in clusters with similar ocean-biogeochemical conditions (Landschützer et al., 2016; Gregor and Gruber, 2021), grouping data-sparse regions with others with similar conditions. The approach taken here, instead, attempts to tackle these limitations by not only using fCO_2 measurements for training, but also pH measurements from biogeochemical Argo floats (Johnson et al., 2017), which provide critical additional data in the Southern Ocean, and other biogeochemical data as provided by the Global Ocean Data Analysis Project (GLODAPv2; Olsen et al., 2016). Furthermore, the neural network in MOCCA is pretrained on CMIP6 Earth system models to foster inference of surface fCO_2 in sparsely sampled ocean regions with the prior knowledge about functional relationships from Earth system models. The use of multiple sources of data is enabled by a flexible model structure that allows to train on multiple target variables. Similarly to Iida et al. (2021), this flexible approach also aims at providing coherent estimates for all variables of the oceanic carbonate system. However, in the current version, the learning from multiple sources fails, with reductions in loss for the data from the Surface Ocean CO_2 Atlas, for example, coming at a price of increased loss on the other data.

2 Overall model architecture

MOCCA is based on a three-step procedure (Figure 2). In a first step, surrogate neural network models are trained to fit the functional relationship between a set of variables needed to solve the oceanic carbonate system (*carbonate chemistry drivers*) and two specific carbonate chemistry variables, pH and fugacity of CO_2 (Section 3). The functional relationship is learned from the numerical carbonate chemistry package *mocsy 2.0* (Orr and Epitalon, 2015). This step is performed first to identify appropriate hyper parameters and training procedure in a situation where near-perfect learning is possible.

In a second step, a neural network is trained on model output from CMIP6 Earth system models to learn the statistical relationships between a set of predictors and the carbonate chemistry drivers. This model will serve as a prior estimate to improve inference for fCO_2 and pH in undersampled regions where only few fCO_2 and pH measurements are available.

In a third step, the CMIP6-pretrained model from the previous step is trained with observational data for fCO_2 from the Surface Ocean CO_2 Atlas

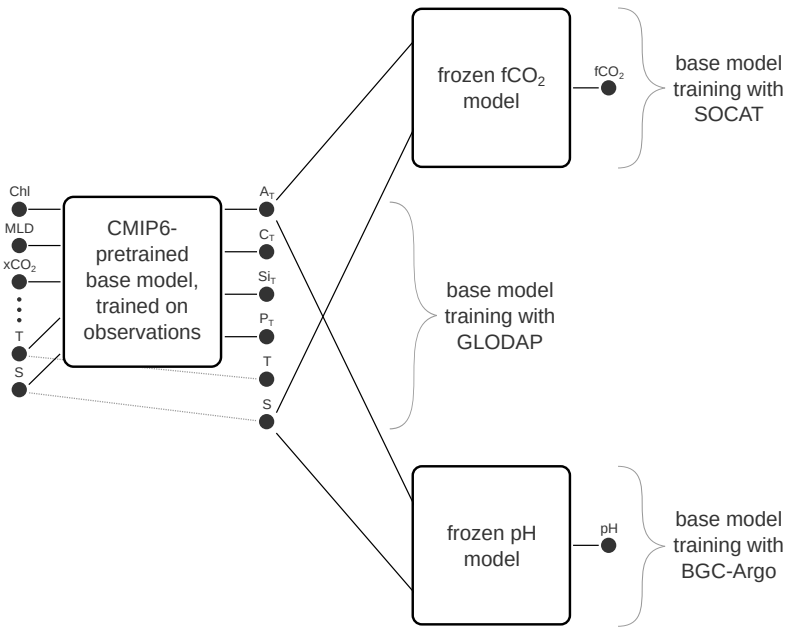

Figure 1: Scheme displaying the overall model architecture of MOCCA.

(SOCAT), carbonate chemistry drivers from GLODAP, and pH from BGC-Argo floats. To do so, the loss is either directly calculated from the output of the base model (GLODAP), or the fCO_2 or pH models are appended to the base model (for SOCAT and BGC-Argo respectively). In the latter case, the layers of the fCO_2 or pH models are frozen to prevent further parameter updates.

3 Surrogate models for mocsy fCO2 and pH

Jupyter Notebooks

train_fco2_model.ipynb
train_ph_model.ipynb

In a first step, surrogate multi-layer perceptron models were trained to replace numerical solution of the oceanic carbonate system. To do so, samples for total alkalinity (A_T), dissolved inorganic carbon (C_T), temperature (T), salinity (S), total silicate (Si_T), and total phosphate (P_T) were randomly generated from uniform distributions (Table 1).
The training data size was set to 5 000 000 samples and the validation data size was set to 1 000 000 samples. Mocsy 2.0 (Orr et al., 2015a) was then used to calculate fCO_2 and pH for these samples. After normalizing per feature with the means and standard deviations given in Table 1, multilayer perceptron models with three identical hidden layers were trained with mocsy fCO_2 and pH as labels.
For fCO_2, model complexity was iteratively increased until a desired maximum deviation of less than $1\,\mu$atm was reached (the measurement uncertainty for pCO_2 as reported by Orr et al, 2015b)[1]. During training, learning rate was decreased from 10^{-3} to 10^{-5} following an exponential learning rate schedule over 10 000 epochs. The decay in learning rate was chosen to shift from an initial identification of an optimal region in the parameter space to finding an optimal set of parameters for which the mean squared error over the training and validation sets converges to a similar and low value.
Hidden layer size was increased from an initial 64 hidden layer units (8833 trainable parameters), 80 units (13601 parameters), 96 units (19393 parameters), 128 units (34049 parameters), to 160 units (52801 parameters). The largest model[2] hit a maximum deviation of $1.08\,\mu$atm over the validation set (Figure 2). The same model architecture was then also used to train the pH model, resulting in a maximum deviation over the validation set of 0.0044.
With root mean squared errors (RMSE) of $0.026\,\mu$atm (fCO_2 model) and 0.00008

[1] For reference, fCO_2 across the generated samples varies between $0.001\,\mu$atm and $5108\,\mu$atm.

[2] To give some context about the model complexity: The number of paramters of this model is comparable to that of a Taylor expansion of a function with six arguments to 15th order. Assuming an efficient use of the MLP parameters, a good fit to the numerical solution from mocsy is thus expected.

	minimum	maximum
A_T	$1000\,\mu\text{mol}\,\text{kg}^{-1}$	$3000\,\mu\text{mol}\,\text{kg}^{-1}$
C_T	$1000\,\mu\text{mol}\,\text{kg}^{-1}$	A_T
T	-2 °C	35 °C
S	10 PSU	50 PSU
Si_T	$0\,\mu\text{mol}\,\text{kg}^{-1}$	$134\,\mu\text{mol}\,\text{kg}^{-1}$
P_T	$0\,\mu\text{mol}\,\text{kg}^{-1}$	$4\,\mu\text{mol}\,\text{kg}^{-1}$

Table 1: Minima and maxima of the uniform distributions used to generate samples for the fCO_2 and pH surrogate models. The range for A_T was chosen such that it easily encompasses open ocean variations in A_T. That for C_T is limited to values lower A_T since larger C_T do not occur in the ocean. The ranges for T and S are chosen according to Lueker et al., 2000, whose parameterizations for K_1 and K_2 are used in mocsy. Finally, the maximum values for Si_T and P_T were chosen to be the global maxima found in the monthly climatologies for Si_T and P_T from World Ocean Atlas 2023. These uniform distributions have means $(max + min)/2$ and standard deviations $(max - min)/\sqrt{12}$, expcept for C_T where mean and standard deviation are given by $min + (max - min)/4$ and $(max - min) \cdot \sqrt{7/144}$, respectively. These means and standard deviations are used for feature normalization.

(pH model), these surrogate models provide a precision that is comparable to numerical carbonate chemistry packages: Orr et al., 2015b report a desired numerical uncertainty of 0.1 μatm and 0.0003, respectively.
The accurracy of the fCO_2 and pH neural network models is similar for the six million random samples for A_T, C_T, T, S, Si_T, and P_T from the CMIP6 models (section 4). Specifically, RMSE is 0.029 μatm and 0.00006, respectively, and the maximum deviations are 0.24 μatm and 0.0006, respectively.

4 CMIP6-pretrained base model

Jupyter Notebooks

CMIP6_data_preparation.ipynb
train_CMIP6_base_model.ipynb
under ./single_CMIP6_model_experiments:
CMIP6_data_preparation_1model.ipynb
train_CMIP6_base_model_1model.ipynb
train_CMIP6_base_model_1model_with_cld_pr.ipynb
train_CMIP6_base_model_1model_large_model.ipynb

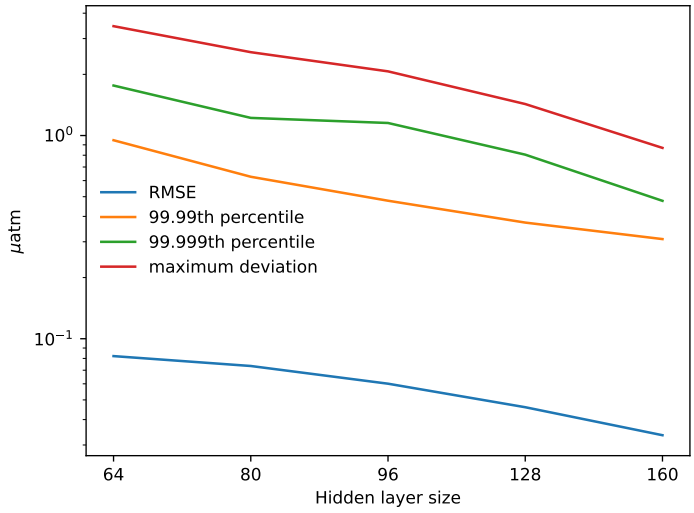

Figure 2: The evolution of root mean squared error (RMSE), the 99.99th and 99.999th percentiles of deviation and the maximum deviation between the surrogate model and mocsy fCO_2 over the validation data set with 1000 000 randomly generated samples.

As a first step, we pre-train a neural network on the statistical relationships between predictor variables and sea surface A_T, C_T, Si_T, and P_T. The predictor variables chosen here are sea surface temperature (SST), sea surface salinity (SSS), mixed layer depth (MLD), sea surface height (SSH), chlorophyll-a concentration (chl), sea ice concentration (ice), 10 m easterly wind (u), 10 m northerly wind (v), dry volume molar ratio of CO_2 in the atmosphere (CO_2), sinus and cosinus of the month of year ($\sin(m/12 \cdot 2\pi)$ and $cos(m/12 \cdot 2\pi)$), as well as latitude and longitude encoded using spherical coordinates following Gade (2010). The representations for month of year and location are used to avoid discontinuities in the months between December and January and in longitude at the dateline (±180 °E).

The neural network is trained on 6 million samples (5 million for training and 1 million for validation) that were evenly drawn from three CMIP6 Earth system models: the UKESM1-0-LL, the MPI-ESM1-2-LR, and the CMCC-ESM2. Despite a high-resolution version of the MPI model, these were the only ones to provide all predictors on monthly-mean resolution. Data from multiple Earth system models were used to ensure that the model learns general relationships that are not specific to a certain Earth system model with respective biases. The model data is taken from the period 1993-2022 (*historical* simulation until 2014, followed by *SSP2-4.5*). The predictors MLD and chl were log-transformed prior to training to foster learning. The log transformation was applied only where the resulting distribution was more Gaussian (where the test statistic of a Kolmogorov-Smirnov test, i.e. the maximum difference between the empirical distribution function of the data and the distribution function of a standard normal distribution, was smaller after transformation). Finally the predictors for the CMIP6 pre-trained base model are nromalized to zero mean and unit variance and the 4 labels are normalized as specified in Table 1 (to match the normalization used for the fCO_2 and pH surrogate models).

For the training, the same architecture that was already used to train the surrogate models (despite not applying a final ELU non-linearity on the output layer, such that the last hidden layer is linearly projected) is used, again with 10 000 epochs and a batch size of 1000. Root mean squared errors of 6.1 μmol kg^{-1} (A_T), 6.3 μmol kg^{-1} (C_T), 0.8 μmol kg^{-1} (Si_T), and 0.03 μmol kg^{-1} (P_T) are obtained (on the validation set and after backtransforming to unnormalized labels). As such, RMSEs are more than a magnitude lower than the standard deviations of A_T, C_T, Si_T, and P_T in the CMIP6 model data samples, given by 111.7, 99.1, 26.3, and 0.5 μmol kg^{-1}, respectively[3]. Likely due to the large size of the training data set relative to the model complexity, overfitting appears not to be a problem. The validation loss continuously decreases, ending up 11 % larger than the training loss[4].

A part of this error can be explained by the fact that the neural network

[3]RMSE equals the standard deviation for a model that always predicts the mean of a variable. Hence, a lower RMSE implies skill in predicting spatial and temporal variations in the variable.

[4]Technically speaking, the training data loss is calculated for all batches in an epoch separately and then averaged. It is thus calculated slightly different than the validation loss.

mapping is necessarily imperfect since the three models imply different statistical relationships between predictors and labels. Creating 6 million samples only from the UKESM1-0-LL model, the root mean squared errors become significantly smaller, now being 4.0 μmol kg^{-1} for A_T, 4.2 μmol kg^{-1} for C_T, 0.5 μmol kg^{-1} for Si_T, and 0.02 μmol kg^{-1} for P_T. The remaining errors should be mainly due to insufficient information in the predictor variables to fully explain the variations in the four concentrations. Adding further predictors may enhance the skill of a model. However, an experiment with total cloud cover and precipitaion as additional atmospheric predictors resulted only in marginal improvements of errors. In another sensitivity test, the training using UKESM1-0-LL model data was also repeated using a much wider model architecture (256 units per hidden layer, 150 % increase in number of parameters). In the first half of the training, validation loss steadily declines, becoming 24 % smaller than the validation loss for the default model architecture with 160 units per hidden layer. This highlights a potential for increasing model performance with a larger neural network to a certain extent. In the second half of the training, however, overfitting results in a steady increase in validation loss. As such, a fully-connected neural network of this size requires either more training data to converge without overfitting or some regularization technique. Given that there is an order of magnitude less data availble from observations, the smaller network architecture appears to be a good choice for the next step, where the model is fine-tuned with observational data following a transfer learning protocol.

5 SOCAT-, GLODAP-, and BGC-Argo-based model tuning

Jupyter Notebooks

gridding_the_GLODAP_data.ipynb
gridding_the_bgc-argo_data.ipynb
train_base_model_on_observations.ipynb

In this step, the CMIP6-pretrained base model shall be trained on observational data. The data products used for each predictor are listed in Table 2 and those used for the label data are listed in Table 3. These data are generally on much higher-resolution than the regular 1°-latitude × 1°-longitude grid used here. As such the data for SST, SSS, MLD, chl, ice, SSH, u, and v were first binned to this coarser reslution. For CO_2 a global and annual-average representative value was used. The Surface Ocean CO_2 Atlas (SOCAT) provides average fCO_2 measurements on the desired 1°×1° grid. The data from GLODAP and BGC-Argo, however is only available as individual un-gridded measurements. These data were gridded here in the jupyter notebooks listed above. From GLODAP, two different types of gridded data were derived; one with all four concentrations (A_T, C_T, Si_T, and P_T) available, and one only

with A_T and C_T (the more important quantities) available. As a result, four data categories are used here for training and validation: SOCAT, BGC-Argo, GLODAP-4, and GLODAP-2. Removing grid values where not all predicors provide data, these data categories encompass 324296, 9449, 10629, and 3959 gridded values, respectively.

The 14 predictors (incl. representations of month of year and geographical location as outlined in the last section) were then normalized with the means and standard deviations from the CMIP6 samples. This was done to improve 'out-of-the-box' CMIP6-pretrained model performance on the observational data. Following the scheme in Fig. 2, a modular MLP was then defined, with the base model with pretrained weights and biases from the CMIP6 data training, the frozen surrogate fCO2 model mapping the output from the base model to fCO_2 if a respective flag is provided in the MLP model call, and the frozen surrogate pH model to map the base model output to pH given the respective flag is provided in the model call.

As a first experiment, the model was used to predict fCO_2, pH, A_T, C_T, Si_T, and P_T over the validation sets of the respective data categories SOCAT, BGC-Argo, GLODAP-4, and GLODAP-2. In this zero-shot inference scenario, the model archieved the correct order of magnitude. However, RMSEs of 49.5μatm, 0.038, 126.6μmol kg^{-1}, 114.5μmol kg^{-1}, 15.0μmol kg^{-1}, and 0.25μmol kg^{-1}, respectively, are similar to the sample standard deviations, highlighting limited predictive power.

In a next step, the base-model should be finetuned on observational data. This approach is motivated by the relatively low amount of observational data compared to model complexity and by the biased distribution of obesrvational data, while climate model data is available everywhere in the world ocean such that a model pretrained on climate model data may be less spatially biased. The idea was to roughly follow the protocoll by Géron (2019), who proposes to first retrain the final layer of the base model (a linear projection layer in this case), and then to subsequently unfreeze layers from the output side and fine-tune them with a small learning rate, until model performance on the validation data does not improve anymore.

When alternatingly training on all four data categories, however, I faced several difficulties. First, the loss on the different data categories is on different orders of magnitude, such that the training will prioritize reducing the loss in one variable over reducing the loss in another variable - with, for example, pH varying much less then fCO2, resulting in much smaller MSEs for pH than for fCO2 and less training success for pH. This problem can be, in general tackled by weighting the MSE loss accoringly for the differnet data categories. More fundamentally, however is the problem that improved loss in one variable caused decreased performance in other variables. As a test, I trained the base model only on the SOCAT data set, expecting that the training would optimize the base model in a way that lead to reasonable estimates for A_T, C_T, Si_T, and P_T (the output neurons that are mapped to fCO_2). Yet, model optimization resulted in unrealistically large values for these variables, in a way that still resulted in plausible values for fCO_2 and a decline in loss. Mathematically,

Predictor	Data product
Sea surface temperature	ESA SST CCI and C3S[1]
Sea ice fraction	ESA SST CCI and C3S[1]
Sea surface salinity	CMEMS ARMOR3D L4[2]
Mixed layer depth	CMEMS ARMOR3D L4[2]
Sea surface height	CMEMS L4 Sea Surface Heights[3]
Chlorophyll-a concentration	Copernicus-GlobColour[4]
CO_2 mole fraction	Manua Loa Hawaii in-situ data[5]
Eastward near-surface wind	ECMWF Reanalysis v5 (ERA5)[6]
Northward near-surface wind	ECMWF Reanalysis v5 (ERA5)[6]

Table 2: List of predictors used to predict A_T, C_T, Si_T, and P_T and the data product used for each predictor. [1]`https://doi.org/10.48670/moi-00169`; only available until Oct. 2022, [2]`https://doi.org/10.48670/moi-00052`, [3]`https://doi.org/10.48670/moi-00148`, [4]`https://doi.org/10.48670/moi-00281`; only available from Sept. 1997 on, [5]`https://scrippsco2.ucsd.edu/data/atmospheric_co2/primary_mlo_co2_record.html`, [6]`https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means`

there seem to be different ways to optimize A_T, C_T, Si_T, and P_T that lead to plausible fCO_2 and pH downstream. The problem may be enhanced by potential biases in the different data categories - adjusting to a bias in one data category can degrade performance in other categories. This finding suggests that simultaneous training on all four data sets may not synergistically improve the predictive power of the model but rather result in trade offs, reducing model performance and hampering interpretability of the results.

References

D. C. E. Bakker, B. Pfeil, C. S. Landa, N. Metzl, K. M. O'Brien, A. Olsen, K. Smith, C. Cosca, S. Harasawa, S. D. Jones, S. Nakaoka, Y. Nojiri, U. Schuster, T. Steinhoff, C. Sweeney, T. Takahashi, B. Tilbrook, C. Wada, R. Wanninkhof, S. R. Alin, C. F. Balestrini, L. Barbero, N. R. Bates, A. A. Bianchi, F. Bonou, J. Boutin, Y. Bozec, E. F. Burger, W.-J. Cai, R. D. Castle, L. Chen, M. Chierici, K. Currie, W. Evans, C. Featherstone, R. A. Feely, A. Fransson, C. Goyet, N. Greenwood, L. Gregor, S. Hankin, N. J. Hardman-Mountford, J. Harlay, J. Hauck, M. Hoppema, M. P. Humphreys, C. W. Hunt, B. Huss, J. S. P. Ibánhez, T. Johannessen, R. Keeling, V. Kitidis,

Label	Data product
fCO2	SOCATv2023[1]
A_T, C_T, Si_T, P_T	GLODAPv2 2023[2]
pH	BGC-Argo[3]

Table 3: List of label data used to train the base model. [1]`https://socat.info/index.php/2023/06/20/v2023-release/`, [2]`https://glodap.info/index.php/merged-and-adjusted-data-product-v2-2023/`, [3]`https://www.seanoe.org/data/00311/42182/`

A. Körtzinger, A. Kozyr, E. Krasakopoulou, A. Kuwata, P. Landschützer, S. K. Lauvset, N. Lefèvre, C. Lo Monaco, A. Manke, J. T. Mathis, L. Merlivat, F. J. Millero, P. M. S. Monteiro, D. R. Munro, A. Murata, T. Newberger, A. M. Omar, T. Ono, K. Paterson, D. Pearce, D. Pierrot, L. L. Robbins, S. Saito, J. Salisbury, R. Schlitzer, B. Schneider, R. Schweitzer, R. Sieger, I. Skjelvan, K. F. Sullivan, S. C. Sutherland, A. J. Sutton, K. Tadokoro, M. Telszewski, M. Tuma, S. M. A. C. van Heuven, D. Vandemark, B. Ward, A. J. Watson, and S. Xu. A multi-decade record of high-quality fco_2 data in version 3 of the surface ocean co_2 atlas (socat). *Earth System Science Data*, 8(2):383–413, 2016. doi: 10.5194/essd-8-383-2016. URL `https://essd.copernicus.org/articles/8/383/2016/`.

T. T. T. Chau, M. Gehlen, and F. Chevallier. A seamless ensemble-based reconstruction of surface ocean pco_2 and air–sea co_2 fluxes over the global coastal and open oceans. *Biogeosciences*, 19(4):1087–1109, 2022. doi: 10.5194/bg-19-1087-2022. URL `https://bg.copernicus.org/articles/19/1087/2022/`.

P. Friedlingstein, M. O'Sullivan, M. W. Jones, R. M. Andrew, D. C. E. Bakker, J. Hauck, P. Landschützer, C. Le Quéré, I. T. Luijkx, G. P. Peters, W. Peters, J. Pongratz, C. Schwingshackl, S. Sitch, J. G. Canadell, P. Ciais, R. B. Jackson, S. R. Alin, P. Anthoni, L. Barbero, N. R. Bates, M. Becker, N. Bellouin, B. Decharme, L. Bopp, I. B. M. Brasika, P. Cadule, M. A. Chamberlain, N. Chandra, T.-T.-T. Chau, F. Chevallier, L. P. Chini, M. Cronin, X. Dou, K. Enyo, W. Evans, S. Falk, R. A. Feely, L. Feng, D. J. Ford, T. Gasser, J. Ghattas, T. Gkritzalis, G. Grassi, L. Gregor, N. Gruber, O. Gürses, I. Harris, M. Hefner, J. Heinke, R. A. Houghton, G. C. Hurtt, Y. Iida, T. Ilyina, A. R. Jacobson, A. Jain, T. Jarníková, A. Jersild, F. Jiang, Z. Jin, F. Joos, E. Kato, R. F. Keeling, D. Kennedy, K. Klein Goldewijk, J. Knauer, J. I. Korsbakken, A. Körtzinger, X. Lan, N. Lefèvre, H. Li, J. Liu, Z. Liu, L. Ma, G. Marland, N. Mayot, P. C. McGuire, G. A. McKinley, G. Meyer, E. J. Morgan, D. R. Munro, S.-I. Nakaoka, Y. Niwa, K. M. O'Brien, A. Olsen, A. M. Omar, T. Ono, M. Paulsen, D. Pierrot, K. Pocock, B. Poulter, C. M. Powis, G. Rehder, L. Resplandy, E. Robertson, C. Rödenbeck, T. M. Rosan,

J. Schwinger, R. Séférian, T. L. Smallman, S. M. Smith, R. Sospedra-Alfonso, Q. Sun, A. J. Sutton, C. Sweeney, S. Takao, P. P. Tans, H. Tian, B. Tilbrook, H. Tsujino, F. Tubiello, G. R. van der Werf, E. van Ooijen, R. Wanninkhof, M. Watanabe, C. Wimart-Rousseau, D. Yang, X. Yang, W. Yuan, X. Yue, S. Zaehle, J. Zeng, and B. Zheng. Global carbon budget 2023. *Earth System Science Data*, 15(12):5301–5369, 2023. doi: 10.5194/essd-15-5301-2023. URL `https://essd.copernicus.org/articles/15/5301/2023/`.

K. Gade. A non-singular horizontal position representation. *THE JOURNAL OF NAVIGATION*, 63:395–417, 2010. ISSN 0021-9991. doi: doi:10.1017/S0373463309990415.

L. Gloege, M. Yan, T. Zheng, and G. A. McKinley. Improved quantification of ocean carbon uptake by using machine learning to merge global models and pco2 data. *Journal of Advances in Modeling Earth Systems*, 14(2):e2021MS002620, 2022. doi: https://doi.org/10.1029/2021MS002620. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002620`. e2021MS002620 2021MS002620.

I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. `http://www.deeplearningbook.org`.

L. Gregor and N. Gruber. Oceansoda-ethz: a global gridded data set of the surface ocean carbonate system for seasonal to decadal studies of ocean acidification. *Earth System Science Data*, 13(2):777–808, 2021. doi: 10.5194/essd-13-777-2021. URL `https://essd.copernicus.org/articles/13/777/2021/`.

A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Sebastopol, CA, 2nd edition, 2019. ISBN 978-1-492-03264-9.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL `https://www.sciencedirect.com/science/article/pii/0893608089900208`.

Y. Iida, Y. Takatani, A. Kojima, and M. Ishii. Global trends of ocean co2 sink and ocean acidification: an observation-based reconstruction of surface ocean inorganic carbon variables. *Journal of Oceanography*, 77(2):323–358, Apr 2021. ISSN 1573-868X. doi: 10.1007/s10872-020-00571-5. URL `https://doi.org/10.1007/s10872-020-00571-5`.

K. S. Johnson, J. N. Plant, L. J. Coletti, H. W. Jannasch, C. M. Sakamoto, S. C. Riser, D. D. Swift, N. L. Williams, E. Boss, N. Haëntjens, L. D. Talley, and J. L. Sarmiento. Biogeochemical sensor performance in the soccom profiling float array. *Journal of Geophysical Research: Oceans*, 122(8):6416–6436, 2017. doi: https://doi.org/10.1002/2017JC012838. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JC012838`.

P. Landschützer, N. Gruber, and D. C. E. Bakker. Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles*, 30(10): 1396–1417, 2016. doi: https://doi.org/10.1002/2015GB005359. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015GB005359.

A. Olsen, R. M. Key, S. van Heuven, S. K. Lauvset, A. Velo, X. Lin, C. Schirnick, A. Kozyr, T. Tanhua, M. Hoppema, S. Jutterström, R. Steinfeldt, E. Jeansson, M. Ishii, F. F. Pérez, and T. Suzuki. The global ocean data analysis project version 2 (glodapv2) – an internally consistent data product for the world ocean, 2016. URL https://essd.copernicus.org/articles/8/297/2016/.

J. C. Orr and J.-M. Epitalon. Improved routines to model the ocean carbonate system: mocsy 2.0. *Geoscientific Model Development*, 8(3):485–499, 2015. doi: 10.5194/gmd-8-485-2015. URL https://gmd.copernicus.org/articles/8/485/2015/.

M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, Feb 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0912-1. URL https://doi.org/10.1038/s41586-019-0912-1.

C. Rödenbeck, T. DeVries, J. Hauck, C. Le Quéré, and R. F. Keeling. Data-based estimates of interannual sea–air co_2 flux variations 1957–2020 and their relation to environmental drivers. *Biogeosciences*, 19(10):2627–2652, 2022. doi: 10.5194/bg-19-2627-2022. URL https://bg.copernicus.org/articles/19/2627/2022/.

J. Zeng, Y. Iida, T. Matsunaga, and T. Shirai. Surface ocean co2 concentration and air-sea flux estimate by machine learning with modelled variable trends. *Frontiers in Marine Science*, 9, 2022. ISSN 2296-7745. doi: 10.3389/fmars.2022.989233. URL https://www.frontiersin.org/articles/10.3389/fmars.2022.989233.