MOCCA

An attempt for

a Multi-sourced Ocean Carbonate Chemistry Analysis

Thesis for the Certificate of Advanced Studies in *Advanced Machine Learning* at University of Bern

Friedrich Anton Burger

August 1, 2024

Jupyter notebooks with the analyses as well as model parameters 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 downscaling, 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. This project tests another approach to tackle this issue; 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; Lauvset et al., 2024). Furthermore, the neural network is here pretrained on CMIP6 Earth system models to test whether this fosters 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), the approach taken here could, in theory, provide coherent estimates for all variables of the oceanic carbonate system. However, facing problems with sparseness of non-fCO_2 data and limited transferability of the functional relationships from climate models to real world data, the approach taken does not improve spatial interpolation of surface-ocean fCO_2.

2 Overall model architecture

The model architecture, here called *MOCCA* for brevity, is based on a three-step procedure (Figure 1):

1 In a first step (section 3), 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*

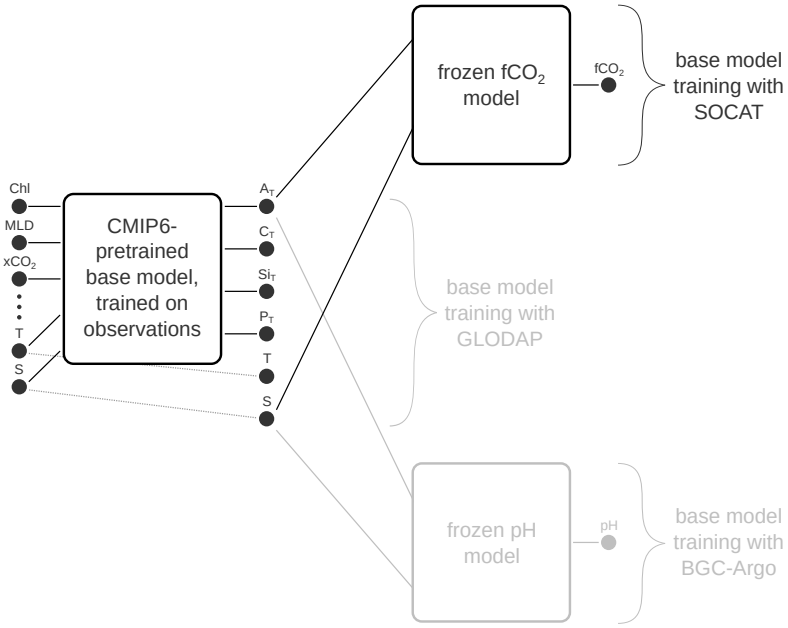

Figure 1: Scheme displaying the overall planned model architecture for MOCCA. Components in grey colors will eventually not be used due to insufficient data availability from GLODAP and BGC-Argo.

2.0 (Orr and Epitalon, 2015). Using the numerical carbonate chemistry package directly in the neural network training would consume too much time and is thus not feasible. This step is performed first to identify appropriate hyper parameters and training procedure in a situation where near-perfect learning is possible.

2 In a second step (section 4), 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.

3 In a third step (section 5), the CMIP6-pretrained base model from the previous step is trained with observational data for fCO_2 from the Surface Ocean CO_2 Atlas (SOCAT). To do so, the surrogate fCO_2 neural network (frozen to prevent parameter updates) is combined with the pretrained base model, and the loss is calculated between predicted fCO_2 and SOCAT data. The original idea was to also include carbonate chemistry drivers from GLODAP, and pH from BGC-Argo floats, with the loss is either directly calculated from the output of the base model (GLODAP), or from the output of the surrogate fCO_2 or pH models (for SOCAT and BGC-Argo respectively). However, it will be seen that inclusion of GLODAP and BGC-Argo data, as envisioned here, is not feasible.

,

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 μatm 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 μatm 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 μatm (fCO_2 model) and 0.00008 (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

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), si-

[1] For reference, fCO_2 across the generated samples varies between 0.001 μatm and 5108 μatm.

[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.

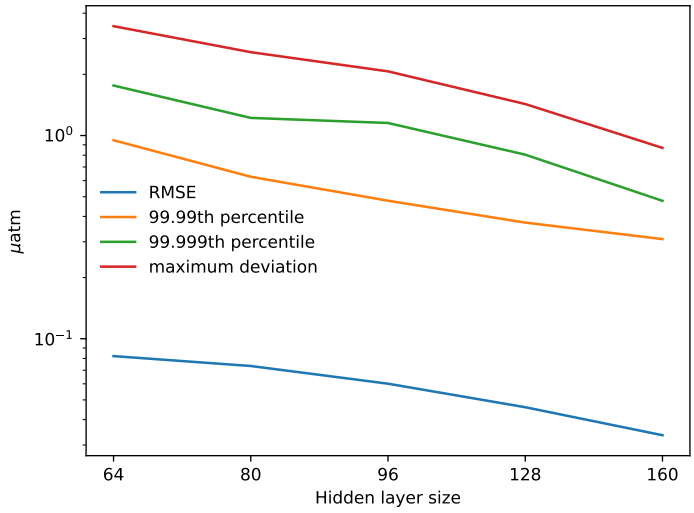

Figure 2: The evolution of root mean squared error (RMSE), the 99.99th and 99.999th percentiles of deviation and the maximum deviation between the surrogate model and mocsy fCO_2 over the validation data set with 1000 000 randomly generated samples.

nus 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 owing to the large size of the training data set relative to the model complexity and to the consistency of the climate model data, 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 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

[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 (also see the discussion on the fraction of variance unexplained in section 5).

[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.

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 Model tuning on observational data

Jupyter Notebooks

gridding_the_GLODAP_data.ipynb
gridding_the_bgc-argo_data.ipynb
consistency_of_GLODAP_with_BGC-Argo_and_SOCAT.ipynb
train_base_model_on_SOCAT.ipynb
training_on_SOCAT_varying_complexity_and_dropout.ipynb
under ./single_CMIP6_model_experiments:
train_base_model_1model_on_MPI_ESM.ipynb

5.1 Observational data preparation

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 intended to be used as label data are listed in Table 3. The predictor 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 resolution. 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 ungridded measurements. These data are here gridded first (see jupyter notebooks listed above). From GLODAP, two different types of gridded data are derived; one with all four concentrations (A_T, C_T, Si_T, and P_T) available, and one only

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`

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

Given the small amount of data available from GLODAP and BGC-Argo, only few data are used to calculate the grid cell values. For example for GLODAP-4, 60% of grid cells are just representing one measurement, 88% of cells represent the average of maximally two measurements. This is in contrast to SOCAT, where the percentage of grid cells based on one or maximally two measurements is 3% and 6%, respectively. It is therefore questionable whether these derived gridded fields adequately represent the mean conditions in the grid cells, which extend over an area of about a 100 km×100 km and over a month. To test the adequacy of representation, the three data-sets are cross-compared in overlapping grid cells (Figure 3). As a baseline, pH is calculated from GLODAP-4 data and compared to the pH measurements provided by GLODAP (Figure 3a), providing information about the effect of measurement uncertainty and uncertainty in the carbonate chemistry calculation. The root mean squared error of 0.006 is smaller than the estimated measurement uncertainty of pH of 0.01 Lauvset et al. (2024). As such, almost all variance in the

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/`

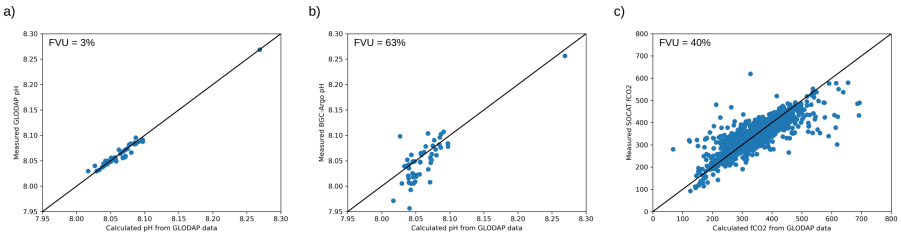

Figure 3: Cross-comparison of data from the gridded GLODAP-4, BGC-Argo, and SOCAT fields where overlapping. (a,b) Comparison of pH calculated from GLODAP-4 variables to measured pH from GLODAP (a) and from BGC-Argo (b). Both panels only show data for grid cells where data from GLODAP and BGC-Argo is available. (c) Comparison of fCO_2 calculated from GLODAP-4 to measured fCO_2 from SOCAT. The fraction of variance unexplained (FVU) is given in each panel. FVU is the ratio of the mean squared error (MSE) to the variance of the explained variable. An FVU of 0% indicates a perfect representation, while an FVU of 100% indicates that no variations can be explained, corresponding to a model that always predicts the mean (in which case MSE equals variance).

calculated pH can be explained by the pH measurements (ratio of mean squared error to variance in calculated pH (the fraction of unexplained variance) is only 3 %). In contrast, BGC-Argo derived pH has much less explanatory power (Figure 3b), with 63 % of the variance in calculated GLODAP pH unexplained. With the only difference between panels a) and b) being the additional uncertainty from the poor representation of grid values from few measurements, this step seems to crucially reduce the quality of the data. Similarly, large discrepancies are also found when comparing GLODAP-calculated fCO_2 with SOCAT fCO_2 measurements (Figure 3c), with a root mean squared error of 34 μatm. Overall, I here conclude that the quality of the gridded fields derived from GLODAP and BGC-Argo is not sufficient for them to be used as additional sources to train a neural network alongside the more robust SOCAT data.

5.2 Test case: tuning a CMIP6 pre-trained model on data from another model

After having established to only use SOCAT observational data, we now proceed with fine-tuning the CMIP6-pretrained base model on the SOCAT data. The fine-tuning approach is motivated by the relatively low amount of observational data compared to neural network complexity and by the biased distribution of observational data. In contrast, climate model data is abundant and available everywhere in the world ocean. As such, a neural network pretrained on climate model data may be less spatially biased. The fine-tuning roughly follow the protocol by Géron (2019), who proposes to first retrain the final layer of the model (the last linear projection layer of the base model in this case), and then to subsequently unfreeze more layers, proceeding from the output side to the input side of the network, and to fine-tune them with a small learning rate, until model performance on the validation data does not improve anymore.

The procedure is here first tested in a controlled environment: The pretrained model based on data from the UKESM model only (introduced in section 4) is here tuned with fCO_2 labels and the set of 14 predictors (incl. representations of month of year and geographical location as outlined in section 4) from the MPI model. To archieve a comparable situation to real data, the data is constrained to where also SOCAT data is available, resuling in 305636 samples. The 14 predictors are normalized with the means and standard deviations from the UKESM samples. This is done to improve 'out-of-the-box' UKESM-pretrained model performance on the MPI-ESM data. Following the scheme in Fig. 1, a modular MLP is then defined, with the base model with pretrained weights and biases from UKESM data, and 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.

First, the *zero-shot* regression performance is tested. The mean squared error exceeds variance on the validation set (RMSE of 53 μatm), indicating near-zero predictive skill for fCO_2 based on the pre-training on another climate model. Retraining the last projection layer of the base model drastically increases model skill, resulting in a RMSE of 17 μatm and a fraction of variance unexplained (FVU) of only 16%. Retraining the second-last layer with a low learning rate (decaying from 10^{-4} to 10^{-5}), and subsequently finetuning the whole base model with a learning rate of 10^{-5}, RMSE drops to 5.7 μatm and FVU to 1.7 %.

However, as already indicated by the poor *zero-shot* performance, there is no advantage from tuning a base model that was pre-trained with spatially uniformly distributed data from another climate model, at least when testing performance on the necessarily also spatially biased validation set: When training a neural network from scratch on the MPI model, RMSE is 5.2 μatm and FVU to 1.4 %. As such, almost all variations in fCO_2 in the validation set can

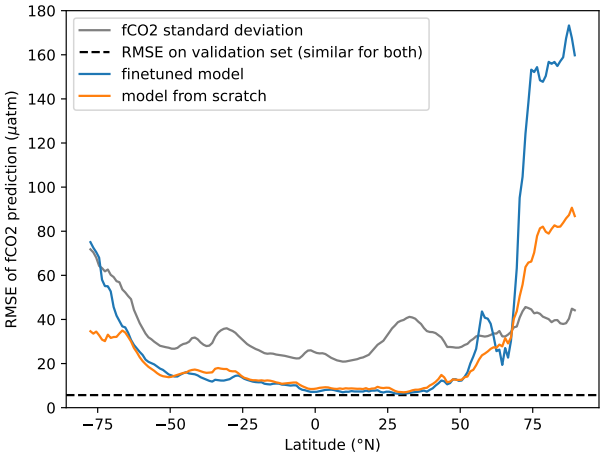

Figure 4: Zonal mean of the root mean squared error (RMSE) in each grid cell over the full 30 year period, calculated between simulated fCO_2 and fCO_2 predicted by the fine-tuned neural network and the neural network trained from scratch on data from the MPI model. The grey dashed line indicates the RMSE on the spatially biased validation set (similar for both models). The grey line shows the zonal mean temporal standard deviation of simulated fCO2. The ratio between RMSE and standard deviation (the square root of FVU) indicates how much temporal variability can be resolved by predicted fCO_2.

be resolved based on the 14 predictors, irrespective of fine-tuning or training from scratch. It can thus also be concluded that the predictors contain sufficient information to resolve almost all variations in fCO_2, and that, despite the small sample size compared to model complexity, overfitting is not problematic for the case of climate model data.

Since training neural networks here on climate model data, one can also analyse RMSE on the whole grid - allowing to see how the trained neural networks perform in data-sparse regions (Figure 4). The comparison of the two models on the complete global data again shows no advantage of the pretrained model over that trained from scratch. On the contrary, the latter model performs better in data-sparse high-latitude regions. Furthermore, it consistently shows a much lower RMSE compared to temporal standard deviation of fCO_2 except in the Arctic Ocean (north of 70 °N), indicating capability of resolving spatial and temporal variations in fCO_2 outside well-sampled regions.

Finally, it is interesting to see what kind of solutions for A_T, C_T, Si_T, and P_T are predicted after training on fCO_2. The mapping $(A_T, C_T, Si_T, P_T) \rightarrow fCO_2$ is no bijection, different combinations of input variables can result in

the same fCO_2, e.g. through an infinitesimal change in the input variables that satisfies $\Sigma_i \partial \mathrm{fCO}/\partial x_i \cdot \delta x_i = 0$. As a result, the neural network finds solutions for A_T, C_T, Si_T, and P_T that predict a correct fCO_2, but are far off realistic values. Specifically, the trained network predicts too large values for all four variables (see notebook train_base_model_1model_on_MPI_ESM). As a result, the predicted solution for the four variables can not be expected to yield a realistic result for other carbonate chemistry variables such as pH. This limitation could be potentially overcome by finding a way to include training data for pH (BGC-Argo) and A_T, C_T, Si_T, and P_T (GLODAP).

5.3 Training on SOCAT data

The fine tuning procedure is now applied to real observational data (notebook train_base_model_on_SOCAT.ipynb). The total sample size, i.e. grid cells where data from all predictors as well as SOCAT fCO_2 label data is available, is given by 311440. The 14 predictors from real world data are normalized with the means and standard deviations from the CMIP6 samples, again to improve *out-of-the-box* CMIP6-pretrained model performance on the observational data. The modular MLP model is defined as introduced in subsection 5.2, but with weights and biases from the base model pretrained on all three CMIP6 models.

As a first experiment, the CMIP6-pretrained model is used to predict fCO_2 without prior training on the observational data. In this zero-shot inference scenario, the model predicts fCO_2 with a RMSE of 35 μatm on the validation set. As such it can not resolve most of the variance in the fCO_2 validation set (FVU of 73 %), highlighting limited predictive power without training on observational data.

In a next step, the base-model should be finetuned on observational data, as it was done on the MPI Earth system model in subsection 5.2. As for the case of transfer learning with Earth system model data, retraining the last projection layer of the base model clearly increases model skill, resulting in a RMSE of 26 μatm and a fraction of variance unexplained (FVU) of only 39%. Retraining the second-last layer with a low learning rate (decaying from 10^{-4} to 10^{-5}), RMSE becomes 19 μatm with a FVU of 20 %.

As for the Earth system model data case, finetuning a CMIP6-pretrained model is not improving loss on the validation set compared to training a model from scratch on SOCAT data only. In the latter case, RMSE after training is 18 μatm. As such, fine-tuning a CMIP-pretrained model does not improve accuracy on unseen data compared to training a model from scratch. In contrast to the Earth system model data only case, however, overfitting is observed: the validation loss approaches 16 μatm early in the training, followed by a degradation of performance on the validation set.

With no advantage of using the pretrained model proposing to train a model

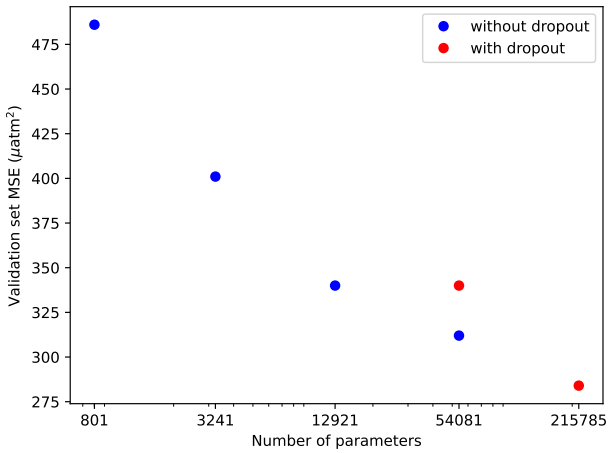

Figure 5: Validation loss (MSE) for training on observed fCO_2 from SOCAT as a function of model complexity (number of parameters). Blue dots indicate networks without dropout layers and red dots indicate networks with dropout.

from scratch, the question of an optimal architecture for such a model, providing sufficient complexity for a global regression of fCO_2 while avoiding overfitting, remains. In the following, three smaller networks, and two with regularization achieved by dropout layers (Srivastava et al., 2012) are tested (notebook training_on_SOCAT_varying_complexity_and_dropout.ipynb). To simplify training, these neural networks predict fCO_2 directly, given that there is no advantage of predicting A_T, C_T, Si_T, and P_T in the absence of GLODAP and BGC-Argo training data. Furthermore, the predictor data is now directly normalized based on the training data set, since there is no use of normalizing the predictors to CMIP6 means and standard deviations when not starting from a CMIP6-pretrained model. This should improve learning during training.

The three smaller networks are chosen such that the number of parameters increases by a factor of 4 between each, from 801 parameters (three hidden layers of size 16) to 3241 parameters (hidden layer size 36) to 12921 parameters (hidden layer size of 76) to 54081 parameters (the default neural network used before; hidden layer size of 160). The methodology for dropout follows Srivastava et al. (2014), using dropout layers between the three hidden layers, a dropout rate of 20 %, and a high initial learning rate of 10^{-2}. Hidden layer size is chosen as 160 (default) and 324 (215785 parameters).

With increasing model complexity, loss on the validation set generally decreases(Figure 5). For the smaller neural networks overfitting does not occur, as indicated by a steady decline of validation loss during training and by com-

parable loss on the training and validation sets. With overfitting present for the default network size (54081 parameters), dropout offers a means of further increasing model complexity while avoiding a loss of generalizability. As discussed in Srivastava et al. (2014), networks with dropout require larger hidden layers for the same network performance. As such, it comes as no surprise that a model without dropout at 54081 parameters performs better on the validation set than one with dropout (Figure 5). However, the large model with dropout (215785 parameters) performs superior, and, at the same time, shows no sign of overfitting (validation and training loss comparable and a steady decrease of validation loss throughout training). It's RMSE on the validation set after training is 16.8 μatm (FVU=16.6 %).

6 Discussion

The performance of the large dropout model is qualitatively comparable to the literature. Gregor and Gruber (2021) report a RMSE of their neural network-based regression model, OceanSODA, of 12 μatm in the open ocean and 28 μatm in coastal regions, qualitatively comparable to the RMSE of 16.8 μatm found here over both coastal and open ocean domains.

The best-performing model in this project, the large dropout model, is not making use of the additional sources of data for model training that were tested here, namely A_T, C_T, Si_T, and P_T from GLODAP and pH from BGC-Argo. Inclusion of these data was here found to be problematic given large representation uncertainty when binning these sparse data on a regular grid. Overcoming these difficulties and including these data still offers high potential for increasing the skill of neural-network based interpolation of surface ocean carbonate chemistry data, in particular in the Southern Ocean where BGC-Argo data is comparably abundant.

In contrast to the initial hypothesis, model performance was not improved when training a neural network first on CMIP-6 Earth system before training on observed SOCAT fCO_2 data. The poor results with transfer learning were here associated with a lack of transferable information between CMIP6 models and observations, as indicated by the low *zero-shot* skill of the pretrained model before fine-tuning. Emulating SOCAT observations by an independent CMIP6 model, it was also shown that transfer learning approach did not improve skill in regions with no SOCAT data to train on.

The latter experiment highlights the usefulness of climate model data as test beds for neural networks, with climate models providing gap-less around the globe and thus a *ground-truth* to test neural network-based interpolations. However, climate model data is generally different from real world data, as it is not including measurement and representation errors, and as the it is derived from equations that represent a simplified version of Earth system dynamics. In

this project, a neural network trained on climate model data showed a smaller loss compared to when trained on observations, with less overfitting, and potentially superior generalizability to ocean regions with few observations (Figure 4). Such discrepancies should be kept in mind when using climate model data to draw conclusions about neural network performance in real-world cases.

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, 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`.

S. K. Lauvset, N. Lange, T. Tanhua, H. C. Bittig, A. Olsen, A. Kozyr, M. Álvarez, K. Azetsu-Scott, P. J. Brown, B. R. Carter, L. Cotrim da Cunha, M. Hoppema, M. P. Humphreys, M. Ishii, E. Jeansson, A. Murata, J. D. Müller, F. F. Pérez, C. Schirnick, R. Steinfeldt, T. Suzuki, A. Ulfsbo, A. Velo, R. J. Woosley, and R. M. Key. The annual update glodapv2.2023: the global interior ocean biogeochemical data product. *Earth System Science Data*, 16(4):2047–2072, 2024. doi: 10.5194/essd-16-2047-2024. URL `https://essd.copernicus.org/articles/16/2047/2024/`.

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/`.

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. doi: doi.org/10.48550/arXiv.1207.0580.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, jan 2014. ISSN 1532-4435.

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`.