

# MOCCA

Multi-sourced Oceanic Carbonate Chemistry Assimilation

Friedrich Burger

February 2024

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Model architecture</b>	<b>1</b>
2.1	Overview . . . . .	1
2.2	Surrogate models for mocsy fCO <sub>2</sub> and pH . . . . .	1
2.3	CMIP6-pretrained DIC and Alk model . . . . .	2
2.4	SOCAT-, GLODAP-, and BGC-Argo-based model tuning . . . . .	2
<b>3</b>	<b>Model performance</b>	<b>2</b>

## 1 Introduction

neural networks...Goodfellow et al. (2016)

## 2 Model architecture

### 2.1 Overview

### 2.2 Surrogate models for mocsy fCO<sub>2</sub> and pH

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<sub>2</sub> 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<sub>2</sub> and pH as labels.

For  $\text{fCO}_2$ , model complexity was iteratively increased until a desired maximum deviation of less than  $1\text{ }\mu\text{atm}$  was reached (the measurement uncertainty for  $\text{pCO}_2$  as reported by Orr et al, 2015b)<sup>1</sup>. During training, batch size was iteratively increased from 1000 to 500 000 and learning rate was iteratively decreased from  $10^{-3}$  to  $10^{-5}$  over up to 10 000 epochs. The increase in batch size and decrease in learning rate was implemented 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<sup>2</sup> hit a maximum deviation of  $0.87\text{ }\mu\text{atm}$  over the validation set (Figure 1). The same model architecture was then also used to train the pH model, resulting in a maximum deviation over the validation set of 0.0023.

With root mean squared deviations of  $0.034\text{ }\mu\text{atm}$  ( $\text{fCO}_2$  model) and 0.00006 (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  $0.1\text{ }\mu\text{atm}$  and 0.0003, respectively.

## 2.3 CMIP6-pretrained DIC and Alk model

## 2.4 SOCAT-, GLODAP-, and BGC-Argo-based model tuning

# 3 Model performance

## References

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

<sup>1</sup>For reference,  $\text{fCO}_2$  across the generated samples varies between  $8\text{ }\mu\text{atm}$  and  $4317\text{ }\mu\text{atm}$ .

<sup>2</sup>To give some context about the model complexity: The number of parameters 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
$A_T$	$1700 \mu\text{mol kg}^{-1}$	$2700 \mu\text{mol kg}^{-1}$
$C_T$	$1700 \mu\text{mol kg}^{-1}$	$A_T$
$T$	$2^\circ\text{C}$	$35^\circ\text{C}$
$S$	$19 \text{ PSU}$	$43 \text{ PSU}$
$\text{Si}_T$	$0 \mu\text{mol kg}^{-1}$	$134 \mu\text{mol kg}^{-1}$
$\text{P}_T$	$0 \mu\text{mol kg}^{-1}$	$4 \mu\text{mol kg}^{-1}$

Table 1: Minima and maxima of the uniform distributions used to generate samples for the  $\text{fCO}_2$  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  $\text{Si}_T$  and  $\text{P}_T$  were chosen to be the global maxima found in the monthly climatologies for  $\text{Si}_T$  and  $\text{P}_T$  from World Ocean Atlas 2023. These uniform distributions have means  $(\text{max} + \text{min})/2$  and standard deviations  $(\text{max} - \text{min})/\sqrt{12}$ , except for  $C_T$  where mean and standard deviation are given by  $\text{min} + (\text{max} - \text{min})/4$  and  $(\text{max} - \text{min}) \cdot \sqrt{7/144}$ , respectively. These means and standard deviations are used for feature normalization.

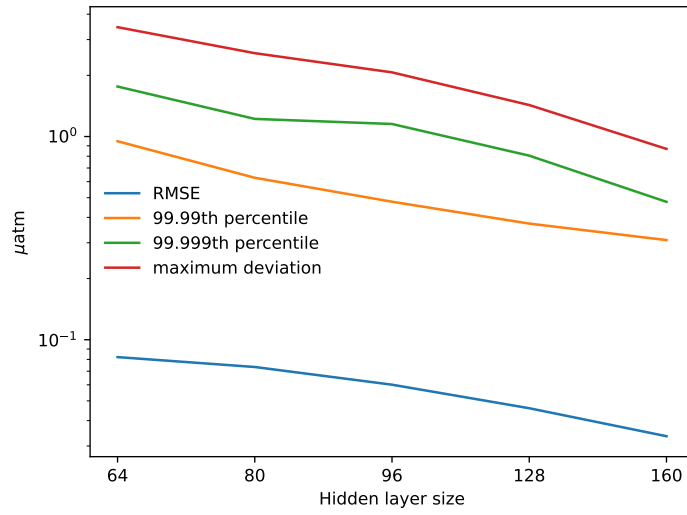


Figure 1: The evolution of root mean squared error (RMSE), the 99.99th and 99.999th percentiles of deviation and the maximum deviation between the surrogate model and mocsy  $\text{fCO}_2$  over the validation data set with 1000 000 randomly generated samples.