

# **Feed-Forward Neural Network of Surface Ocean CO<sub>2</sub>**

Dec 18, 2025

Stamatis Frangoulis  
Kristin Patton

San Diego State University  
MATH 540

### Abstract

The alarming rise of  $CO_2$  in the atmosphere is a major driver of global warming. The excessive  $CO_2$  is stored in carbon sinks such as bodies of water that absorb the  $CO_2$  from the atmosphere. Disproportionate levels of  $CO_2$  in oceanic carbon sinks can lead to ocean acidification—a process that alters seawater chemistry and harms marine life. The model presented in this paper leverages the worldwide  $CO_2$  database along with previous literature on  $CO_2$  modeling using neural networks to model  $CO_2$  levels. A feed-forward neural network (FFNN) is implemented to detect variability in surface-ocean  $CO_2$  and provide estimates for regions with limited measurements.

## 1 Introduction

The rapid increase of  $CO_2$  in the atmosphere is a direct factor to global warming and other detrimental consequences for the environment [7]. Measuring these increases and variations in the atmosphere is vital in monitoring the effects of the climate change.  $CO_2$  is naturally present in the atmosphere and is stored in various environmental carbon sinks; these sinks include large bodies of water, forests, and soils [1]. For oceans, Surface  $CO_2$  measures how much carbon dioxide is absorbed by the ocean from the atmosphere. This process is called ocean acidification because, when carbon dioxide dissolves in water, it forms carbonic acid, which releases hydrogen ions. The pH of the water reflects the concentration of these hydrogen ions—that is, how acidic the water is. Abnormally high acidity harms the ocean's biology by eating away at the minerals necessary for sea life

to thrive and can cause harmful algae blooms. Due to an increase in the burning of fossil fuels and other detrimental practices, carbon sinks are absorbing more  $CO_2$ , making the ocean more acidic. Since the 1960s, the ocean has absorbed 25% of the total anthropogenic  $CO_2$  [6]. Since this phenomenon has only been identified in recent years, the sparse measurements from earlier decades provided an incomplete and inaccurate representation of it, capturing only a small fraction of the overall picture. In this paper, we propose a method for modeling the amount of carbon dioxide absorbed by the ocean that builds upon existing literature and employs a feed-forward neural network (FFNN).

## 2 Description of Data

### 2.1 History of $fCO_2$ Data measurement

Researchers from all over the world upload measurements of surface ocean  $CO_2$  to the largest database of surface ocean  $CO_2$  called SOCAT (Surface Ocean  $fCO_2$  Atlas) [3]. This resource compiles observations from more than 100 international contributors within the marine carbon research community. The dataset consists of measurements of the fugacity of carbon dioxide ( $fCO_2$ ) which is how much  $CO_2$  can be absorbed from the atmosphere.  $fCO_2$  is measured in units of microatmospheres ( $\mu atm$ ). The data consists of 41.4 million observations dating as far back as 1957 [3]. The measurements were mostly collected by scientific teams at specific locations using gliders, floats or other oceanographic sampling platforms. Some of the data also consists of more precise measurements collected by sensors placed on ships that perform global voyages. This data is specifically meant for quantifying the ocean's carbon sinks as well as ocean acidification [3].

### 2.2 Features

The data consists of the following features:

- $fCO_2$  in  $\mu atm$ (microatmosphere)
- Latitude
- Longitude
- Sea surface temperature(SST),
- Sea Surface Temperature Anomaly (SST-AVE)
- Sea Surface Salinity(SSS)
- Mixed Layer Depth (MLD)
- Chlorophyll Concentration(CHL)
- Year
- Month

These features are deemed essential oceanographic measurements and have a strong correlation with the concentration of  $fCO_2$ .

### 2.3 Preprocessing

The data is downloaded from SOCAT directly as a csv file and is filtered for problematic and non existing values. Then, it is separated into measurements from 1990 to 2013 and measurements from 2014. The first part

is used to train the model(see Appendix A.1) while the second part is used to test and validate the model by comparing the output of the model to the actual observations(see Appendix A.2). The data is then normalized by the mean and standard deviation

$$v = \frac{v - \bar{v}}{s}$$

and put into tensor format in order to be processed through the model.

### 3 The Model

Surface ocean  $fCO_2$  data can be used in machine learning to predict levels of carbon dioxide around the world and in future time periods. Specifically, feed-forward neural networks are popular in training models using supervised learning. A neural network consists of the input layer, hidden pooling layers, and the output layer. Activation layers such as ReLU or sigmoidal based functions send the input layer to the hidden layers. Pooling layers can be maximum pooling or average pooling where we take the maximums or the averages, respectively. By utilizing data from a recent 10 year period as technological advancement and emphasis on climate change have driven an increase

in  $fCO_2$  measurements and gathering data from a specific location, a model can be compared to the last 2 years of data to validate accuracy.

#### 3.1 Feed-Forward Neural Networks

Feed Forward Neural Networks (FFNN) or Deep feedforward neural networks are essential multilayer perceptrons. Similarly to most models in machine learning, it attempts to approximate a function  $f$ . Using activation functions, it maps an input  $x$  to a specific value  $y$  or, in a different context, a category  $y$  [4]. The reason they are called feed forward is because information flows only in one direction through the model and is not fed back into the model. Each neuron is adjusted by applying the training data through the following sigmoid function.

$$y_h = \frac{1}{1 + \exp(-b + w^T x)}$$

With  $w$  as the weight,  $b$  the bias and  $x$  the training data. This function is used at each neuron of the hidden layers as well as the output layer where the function translates the value into a  $ffCO_2$  value[5]. The weights are updated by minimizing

the cost function.

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Where  $\hat{y}$  is the outputs of the model and  $y$  is the actual values. The updates are performed using gradient descent through an ADAM (Adaptive Moment Estimation) optimizer.

$$w' = w - a \nabla L$$

where  $a$  is the learning rate. The updated weights are applied in back-propagation.

### 3.2 Model

For our model, we implemented a simple FFNN using pytorch (see Appendix A.3). The first column of neurons consists of the input size, the number of features given in the dataset. The output of the first neurons is passed through the first hidden

layer, which consists of 64 neurons. The output is then passed through a sigmoid activation function to add nonlinearity. The second layer consists of 32 neurons, which are then also passed through a sigmoid function. The final output is then given by the output layer. Lastly, the forward function ensures that the data is fed through the network.

The model is optimized using Mean Squared Error(MSE) as the criterion and using ADAM(Adaptive Moment Estimation) as the optimizer function. ADAM uses a momentum algorithm along with an RMSProp in order to perform gradient descent on the loss function and optimize the model. The model was run for 2000 epochs, where the loss dropped from 0.5 to about 0.209.

### 3.3 Model Structure

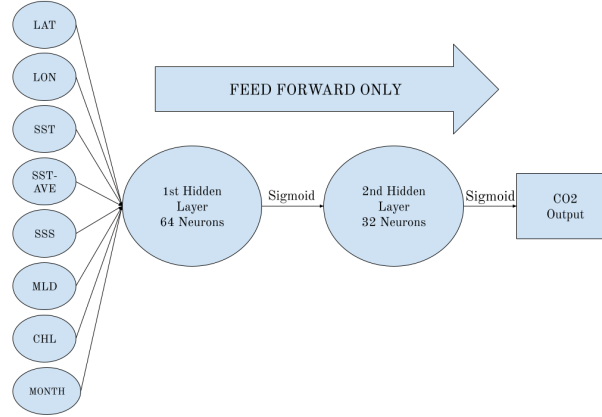


Figure 1: FFNN Structure

## 4 Results

### 4.1 Obtaining Results

In order to test and validate our model, we used values from 2014 as we trained the model with values up to 2013. We differentiated the data into  $fCO_2$  measurements and its features. We then passed through the features for a specific month through our model and compared the outputs with the actual observations. Initially, the model performed differently for various areas. Coastal areas and heavy marine traffic zones affect  $fCO_2$  mea-

surements significantly. These effects led to outlier measurements and values that are challenging to predict. However, after increasing the number of epochs from 50 to 2000, the model is able to incorporate these differences and output accurate results.

### 4.2 Results for Coronation Island (January 2014)

These results of a secluded region can be seen in the area around Coronation Island close to Antarctica where our model achieves a RMSE(Root Mean Squared Error) of 18.4.

Table 1: Actual vs Predicted  $fCO_2$  with Normalized Error for Coronation Island

Actual $fCO_2$	Predicted $fCO_2$	Normalized Error
360.521	326.791199	0.093558
344.778	338.894867	0.017064
342.605	338.093903	0.013167
322.948	336.967102	0.043410
317.573	334.045624	0.051870
317.421	331.994659	0.045913
328.849	332.375549	0.010724
310.074	331.097107	0.067800

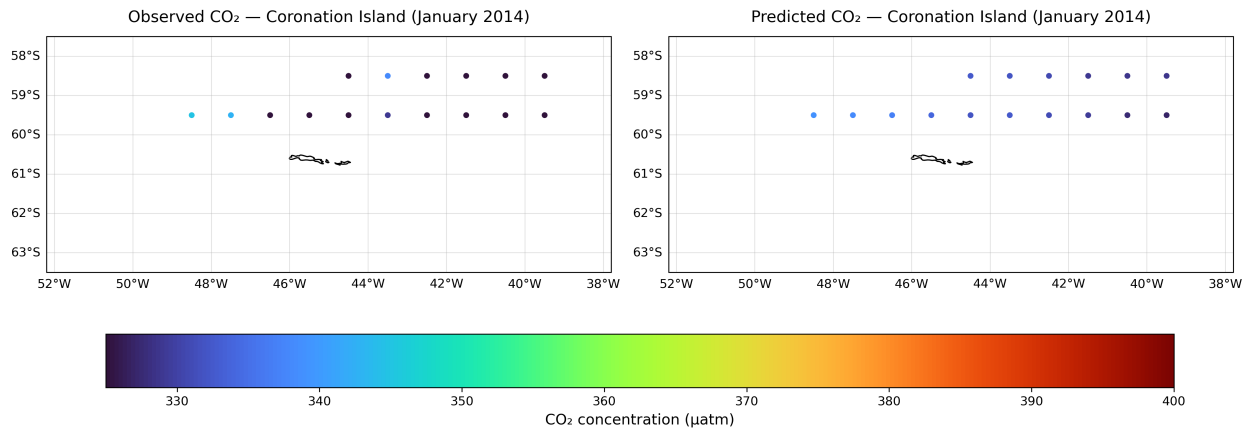


Figure 2: Plot of Actual and Predicted Values for Coronation Island (Jan 2014)

#### 4.3 Results for Irish Coastal Region (February 2014)

Similarly, the model predicts  $fCO_2$  values along the Irish coast with com-

parable accuracy. Despite the region being relatively busy and influenced by human activity, the model still performs well achieving an RMSE of 9.1.

Actual $fCO_2$	Predicted $fCO_2$	Normalized Error
366.831	360.823578	0.016377
366.649	362.838440	0.010393
363.646	367.819427	0.011477
366.176	374.728119	0.023355
365.000	376.743073	0.032173
363.327	377.782745	0.039787
363.280	376.799988	0.037216
363.512	374.858490	0.031214

Table 2: Actual vs. Predicted  $fCO_2$  and Normalized Error

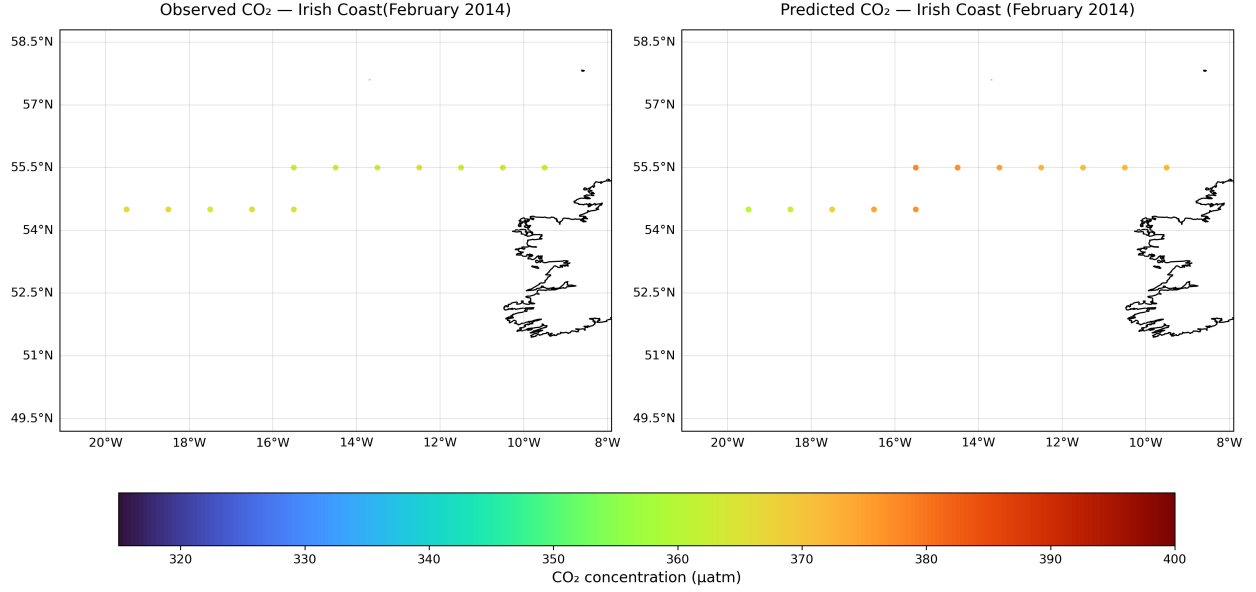


Figure 3: Plot of Actual and Predicted Values for Irish Coast (Feb 2014)

## 5 Discussion

The model agrees with the assumption that there are increased levels of  $fCO_2$  near coastal areas that contain more human activity [2]. This trend is visible in the observational data and is further corroborated by the model, which predicts a pronounced increase in  $fCO_2$  levels near the Irish coast as observed in the sample prediction. Subsequently the model highlights that remote ocean regions experience less variability in

their  $fCO_2$  values due to the lack of freshwater influx and human factors agreeing with [8]. Ultimately, the model cannot be considered accurate enough to predict a precise increase in  $fCO_2$  due to global warming, but it can detect local patterns and variations. Overall, the model succeeds in learning meaningful patterns in the  $fCO_2$  data and achieves relatively low RMSE scores. However, there are significant limitations due to the lack of available data in certain regions and restrictions in computational re-



sources. Future research should be conducted with a larger dataset and more precise measurements to better train the model. Additionally, combining greater computational capacity along with methods from more recent machine-learning literature will enable to capture a broader range of variability in the dataset, including more complex factors such as the presence of ports or nearby large  $fCO_2$  sources.

## References

- [1] John A. Taylor and Jon Lloyd. “Sources and Sinks of Atmospheric  $CO_2$ ”. In: *Australian Journal of Botany* 40.5 (1992), pp. 407–418. DOI: 10.1071/BT9920407. URL: <https://connectsci.au/bt/article-abstract/40/5/407/26326/Sources-and-Sinks-of-Atmospheric-CO2>.
- [2] Md. Abdus-Salam and Toshikuni Noguchi. “Impact of Human Activities on Carbon Dioxide ( $CO_2$ ) Emissions: A Statistical Analysis”. In: *The Environmentalist* 25.1 (2005), pp. 19–30. DOI: 10.1007/s10669-005-3093-4. URL: <https://link.springer.com/article/10.1007/S10669-005-3093-4>.
- [3] B. Pfeil et al. “A uniform, quality controlled Surface Ocean  $CO_2$  Atlas (SOCAT)”. In: *Earth System Science Data* 5.1 (2013), pp. 125–143. DOI: 10.5194/essd-5-125-2013. URL: <https://essd.copernicus.org/articles/5/125/2013/>.
- [4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, MA: MIT Press, 2016. URL: <http://www.deeplearningbook.org>.
- [5] J. Zeng et al. “Technical note: Evaluation of three machine learning models for surface ocean  $CO_2$  mapping”. In: *Ocean Science* 13.2 (2017), pp. 303–313. DOI: 10.5194/os-13-303-2017. URL: <https://os.copernicus.org/articles/13/303/2017/>.
- [6] Nicolas Gruber et al. “Trends and variability in the ocean carbon sink”. In: *Nature Reviews Earth & Environment* 4 (2023), pp. 119–134. DOI: 10.1038/s43017-022-00381-x. URL: <https://www.nature.com/articles/s43017-022-00381-x>.
- [7] Jessica Mkitarian. “Latest Ocean Carbon Data Atlas Shows a Significant Decline in Ocean  $CO_2$  Measurements”. In: *National Oceanic and Atmospheric Administration* (2023). URL: <https://globalocean.noaa.gov/latest-ocean-carbon-data-atlas-shows-a-significant-decline-in-ocean-co2-measurements/>.
- [8] Lucie A. C. M. Knor et al. “Drivers of  $CO_2$ –carbonate System Variability in the Coastal Ocean South of Honolulu, Hawai’i”. In: *Frontiers in Marine Science* 11 (2024). DOI: 10.3389/fmars.2024.1335438. URL: <https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2024.1335438/full>.

## A Appendix

Model and dataset used:

### A.1 Training Data

`train_data.csv`

### A.2 Test Data

`test_data.csv`

### A.3 Model

`model.ypnb`