https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

Detection and Tracking of Carbon Biomes via Integrated Machine Learning

Sweety Mohanty[1,2], Lavinia Patara[1], Daniyal Kazempour[2], and Peer Kröger[2]

[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[2]University of Kiel, Germany

Correspondence: Sweety Mohanty (smohanty@geomar.de)

Abstract.

In the framework of a changing climate, it is useful to devise methods capable of effectively assessing and monitoring the changing landscape of air-sea CO_2 fluxes. In this study, we developed an integrated machine learning tool to objectively classify and track marine carbon biomes under seasonally and interannually changing environmental conditions. The tool was applied to the monthly output of a global ocean biogeochemistry model at 0.25° resolution run under atmospheric forcing for the period 1958-2018. Carbon biomes are defined as regions having consistent relations between surface CO_2 fugacity (fCO_2) and its main drivers (temperature, dissolved inorganic carbon, alkalinity). We detected carbon biomes by using an agglomerative hierarchical clustering (HC) methodology applied to spatial target-driver relationships, whereby a novel adaptive approach to cut the HC dendrogram based on the compactness and similarity of the clusters was employed. Based only on the spatial variability of the target-driver relationships and with no prior knowledge on the cluster location, we were able to detect well-defined and geographically meaningful carbon biomes. A deep learning model was constructed to track the seasonal and interannual evolution of the carbon biomes, wherein a feed-forward neural network was trained to assign labels to detected biomes. We find that the area covered by the carbon biomes responds robustly to seasonal variations in environmental conditions. A seasonal alternation between different biomes is observed over the North Atlantic and Southern Ocean. Long-term trends in biome coverage over the 1958-2018 period, namely a 10% expansion of the subtropical biome in the North Atlantic and a 10% expansion of the subpolar biome in the Southern Ocean, are suggestive of long-term climate shifts. Our approach thus provides a framework that can facilitate the monitoring of the impacts of climate change on the ocean carbon cycle and the evaluation of carbon cycle projections across Earth System Models.

1 Introduction

By absorbing roughly 25% of human-induced carbon emissions (Friedlingstein et al., 2022), the global ocean is a critical component of the Earth's climate and has, until now, mitigated the effects of anthropogenic climate change. The ocean's ability to take up CO_2 depends on both physical processes (the "solubility pump") and biological processes (the "biological pump") (Sarmiento and Gruber, 2006). The biological soft-tissue pump is driven by the absorption of dissolved inorganic carbon (DIC) by photosynthetic primary producers in the sunlit ocean and by its release into the ocean interior through organic matter rem-

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

ineralization. The solubility pump is driven by various factors, notably ocean temperature, chemistry, and circulation. Sea surface temperature (SST) strongly affects CO_2 solubility, with colder waters capable of absorbing more CO_2 than warmer waters. The chemical composition of seawater also plays a role, with waters characterized by higher alkalinity capable of absorbing higher quantities of CO_2 for a given DIC concentration (Williams and Follows, 2011). Ocean circulation and mixing strongly influence air-sea CO_2 fluxes through their effect on the vertical exchanges of DIC and alkalinity between the ocean surface and its interior.

The processes that govern air-sea CO_2 fluxes display considerable spatial and temporal variability. SST imprints a strong north-south gradient of CO_2 solubility, with colder high-latitude waters exhibiting a higher CO_2 uptake than warmer tropical waters (Williams and Follows, 2011). Overlaid to the solubility-driven gradients, patterns of primary productivity and ocean circulation strongly affect the spatial variability of air-sea CO_2 fluxes. Subpolar regions of high primary productivity act overall as a strong CO_2 sink (Takahashi et al., 2009; Mikaloff Fletcher et al., 2007; DeVries et al., 2023). Exceptions are the subpolar latitudes of the Southern Ocean, the equatorial Pacific, and the eastern boundaries of the ocean basins, where wind-driven upwelling of high-DIC waters and the incomplete utilization of upwelled nutrients makes them prone to CO_2 outgassing (Takahashi et al., 2009; Mikaloff Fletcher et al., 2007). High-latitude regions are strongly influenced by sea ice cover, which seasonally hinders the air-sea exchanges of CO_2 and affects surface stratification and primary productivity. The above processes display a substantial seasonal evolution: SST peaks in summer, primary production is highest in spring and summer, and upwelling and vertical mixing are most intense in the cold and wind-swept winter months. Past studies have subdivided the ocean in regions where the air-sea CO_2 flux seasonality is more in phase with SST ("thermal" control), and others where it is more in phase with DIC ("non-thermal" control) (Takahashi et al., 2002; Prend et al., 2022). The oligotrophic subtropical gyres are thermally-driven, whereas polar and subpolar regions (characterized by strong biological production and DIC physical transport) are mostly non-thermally-driven (Takahashi et al., 2002; Prend et al., 2022). While the seasonal cycle is the strongest source of temporal variation, the natural variability of the climate system also introduces year-to-year changes in the processes governing the CO_2 uptake (Landschützer et al., 2016; Gruber et al., 2023). Prominent examples are the El Nino Southern Oscillation (ENSO), which modulates the strength of the upwelling in the equatorial Pacific (Feely et al., 2006), the Southern Annular Mode, which modulates the strength of the Southern Ocean upwelling and associated CO_2 outgassing (Lovenduski et al., 2007), and the North Atlantic Oscillation, which modulates the strength of the subpolar North Atlantic deep mixing and overturning circulation with implications for the carbon cycle (Pérez et al., 2013; Patara et al., 2011).

These widely varying environmental conditions have prompted past studies to objectively classify the global ocean in marine biogeochemical biomes. Marine biomes are characterized by coherent physical forcing and environmental conditions, which are representative of distinctive ecosystem structures (Longhurst, 1995; Sonnewald et al., 2020; Oliver et al., 2015). The classification in marine biomes has several applications, such as evaluating and comparing ocean biogeochemical models (Vichi et al., 2011; DeVries et al., 2023), producing air-sea CO_2 flux reconstructions based on sparse observational data (Landschützer et al., 2013), and efficiently interpreting increasingly large datasets produced by Earth System Models (Jones

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

and Ito, 2019; Couespel et al., 2024). Recently, biome classification went beyond ecosystem applications and explored carbon uptake structures (Fay and McKinley, 2014; Jones and Ito, 2019; Krasting et al., 2022; Couespel et al., 2024). For instance, Fay and McKinley (2014) defined marine biomes based on pre-defined limits of sea ice concentration, SST, mixed layer depth, and chlorophyll values. These biomes, capable of following dynamical ocean boundaries, have been extensively used to assess and compare air-sea CO_2 fluxes in the recent RECCAP-2 project (DeVries et al., 2023).

Against the backdrop of rapidly evolving machine learning (ML) methods, recent studies have contributed a set of tools for categorizing the global ocean into marine biomes (Landschützer et al., 2013; Jones and Ito, 2019; Sonnewald et al., 2020; Krasting et al., 2022; Mohanty et al., 2023a; Couespel et al., 2024). In their work, Couespel et al. (2024) build target-driver relationships between air-sea CO_2 flux and biogeochemical predictors over a time series and use Gaussian Mixture Models to cluster the identified temporal associations into carbon regimes. Jones and Ito (2019) also use Gaussian Mixture Models to segment the ocean surface based on the surface budget of dissolved inorganic carbon, whereas Landschützer et al. (2013) use self-organizing maps to cluster the non-linear relationships between CO_2 partial pressure and its drivers. Sonnewald et al. (2020) presents the Systematic Aggregated Eco-Province (SAGE) method for constructing eco-provinces, which integrates t-stochastic neighbor embedding (t-SNE) and DBSCAN clustering. The works by Krasting et al. (2022) shed light on Arctic ocean acidification, where water mass properties are segmented into four clusters using the SAGE method.

When combined with the ability to track the biomes in time, ML-based detection methods could potentially be used to monitor the time evolution of marine biomes under changing climate conditions. For instance, Reygondeau et al. (2020) implemented a regression-based ensemble approach to predict four biomes (subdivided into 56 biogeochemical provinces) in the future. To this end, they used a supervised method based on the location and properties of the 56 biogeochemical Longhurst provinces (Longhurst, 1995). However, due to the strong fluidity of ocean biomes in response to seasonal and interannual changes in environmental conditions, using tracking methods tied to specific locations is not ideal. The challenges in designing ML-based methods for tracking ecosystems over time are numerous. These include: i) the lack of high-quality annotated geoscientific datasets needed for training and validation steps, ii) building an intricate algorithm to capture the complex spatio-temporal variability within biomes, and iii) the requirement for considerable computational resources, time, or financial investment, depending on the scale of the available data. Nonetheless, tracking provinces over time can be used to assess and predict transformations in ecosystem functioning and carbon cycle dynamics. Biogeochemical provinces are dynamic entities whose spatial extent and position fluctuate in response to climate variations and are anticipated to be further influenced by forthcoming global climate change (Reygondeau et al., 2020; Couespel et al., 2024). Monitoring these changes over time will deliver early signs of ecosystem disruptions, such as ocean acidification (Krasting et al., 2022), allowing for timely intervention and protection measures.

Building upon this motivation, in this study we build a new strategy capable of detecting and tracking over time carbon biomes, which are defined as regions of consistent relationships between surface CO_2 fugacity (a quantity closely related to

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

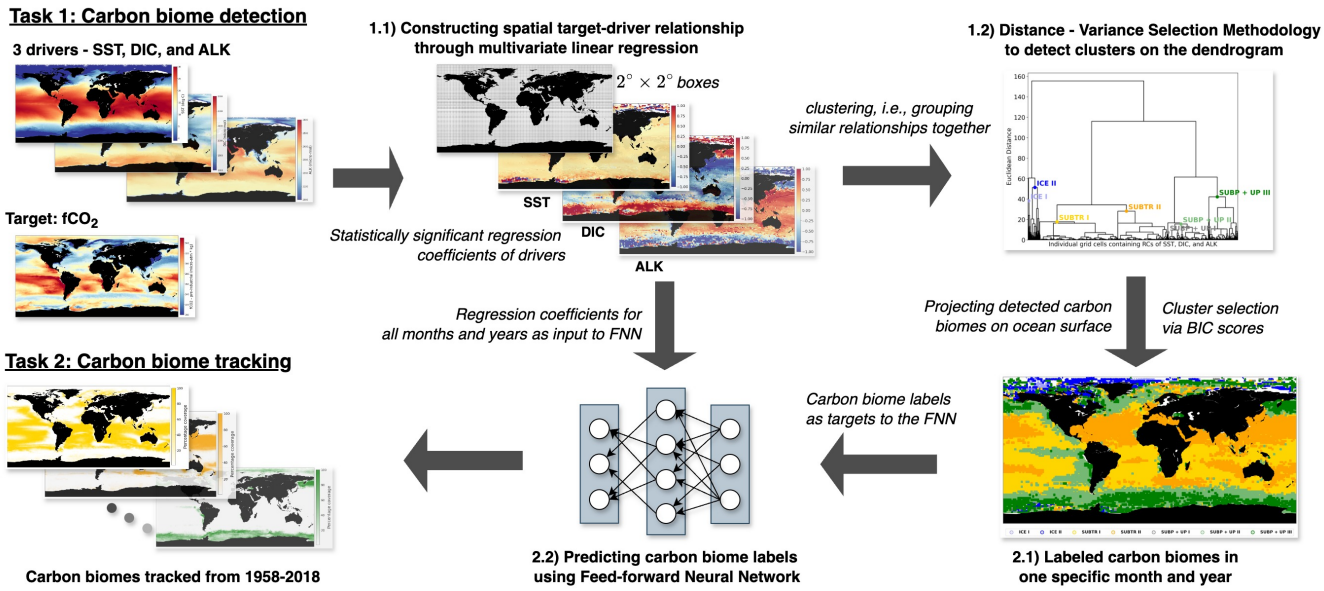

Figure 1. Schematic presentation of the step-by-step approach taken by our study to detect (Task 1) and track (Task 2) marine carbon biomes.

CO_2 partial pressure) and its drivers. Instead of applying clustering on the drivers directly, we built multiple localized target-driver relationships between CO_2 fugacity and its predominant drivers (i.e., surface temperature, dissolved inorganic carbon, and alkalinity). Then, we applied agglomerative hierarchical clustering to group similar target-driver connections and detect the carbon biomes through a distribution-aware technique (Mohanty et al., 2023a). Once the clusters are labeled as specific carbon biomes, we employ a simplistic version of a neural network to capture the connections between the labels and the target-driver relationships, enabling the tracking or the prediction of carbon biomes in time. Our approach thus provides a framework that will facilitate monitoring climate change's impacts on the ocean carbon cycle and evaluating carbon cycle projections across Earth System Models.

The manuscript is structured as follows: Section 2 describes the global ocean biogeochemistry model used for our analysis and the variables collected from it, the technique to build the target-driver relationships, the application of clustering method to detect the carbon biomes, and the building blocks of neural network to track the biomes over time. Section 3 elucidates the outcome of the target-driver analysis and the detected clusters, as well as the tracking of the carbon biomes over time. Section 4 discusses our main findings, highlights the study's limitations, and elaborates on potential research ideas for future studies.

2 Methodology

We define a carbon biome as a region characterized by common relationships between carbon uptake and its drivers. Specifically, we use sea surface fugacity of CO_2 (fCO_2) as the target variable and sea surface temperature (SST), surface dissolved

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

inorganic carbon (DIC), and surface alkalinity (ALK) as its drivers. These variables are obtained from a simulation with a global ocean biogeochemistry model (Sect. 2.1). In the face of the intricate and spatially heterogeneous relationship between fCO_2 and its drivers, we construct multiple localized linear relationships within discrete regions, each spanning a $2° \times 2°$ dimension, as explained in Sect. 2.2. Subsequently, an agglomerative hierarchical clustering methodology is employed, leveraging the collection of regional multivariate linear regression models. Notably, we employ a Distance-Variance Selection methodology (Mohanty et al., 2023a) tailored to the specifics of our task, thereby automating the detection of clusters on the dendrogram, as outlined in Sect. 2.3. We introduce the application of artificial neural networks in Sect. 2.4 to track the detected carbon clusters. Fig. 1 schematically visualizes the entire analytical pipeline, encapsulating the sequential processes.

2.1 Ocean Model Output

We use the monthly output of a global ocean biogeochemistry model composed of the ocean sea ice model NEMO-LIM2 (Madec, 2016) and the biogeochemistry model MOPS (Kriest and Oschlies, 2015; Chien et al., 2022). The model configuration (hereafter called ORCA025-MOPS) is discretized on a global grid having approximately 0.25° horizontal resolution (Barnier et al., 2007) and 46 vertical levels. MOPS simulates the lower trophic levels of the ecosystem and carbonate chemistry using nine biogeochemical tracers (phosphate, nitrate, phytoplankton, zooplankton, detritus, dissolved organic matter, oxygen, DIC, and alkalinity). Calcium carbonate dissolution and production, as well as its effects on alkalinity, are parameterized based on Schmittner et al. (2008). The carbonate chemistry and the air-sea CO_2 exchanges are based on Orr et al. (2017), with an approximate and non-iterative method to compute the carbonate chemistry equilibrium (Follows et al., 2006). This non-iterative solution has been selected for this high-resolution model as a trade-off between computational efficiency and output realism. ORCA025-MOPS shows realistic spatial patterns and seasonality of air-sea CO_2 fluxes, as recently assessed in the RECCAP-2 intercomparison project (DeVries et al., 2023). However, a few outliers off the Peru coast exhibit overestimated CO_2 outgassing and have been removed from the analysis.

The spin-up of ORCA025-MOPS is the following: a NEMO-MOPS configuration at 0.5 ° horizontal resolution (ORCA05-MOPS) was initialized from Levitus et al. (1998) for the temperature and salinity, from GLODAPv.2 (Lauvset et al., 2016; Key et al., 2015) for alkalinity and pre-industrial DIC, and from World Ocean Atlas 2013 (Garcia et al., 2019) for oxygen, nitrate, and phosphate. ORCA05-MOPS was run for three cycles of the atmospheric reanalysis data set JRA55-do forcing (Tsujino et al., 2018) from 1958-2018. Starting from a pre-industrial value of 284.32 ppm, the atmospheric CO_2 mixing ratio increased since 1850 following Meinshausen et al. (2017). Two distinct DIC tracers were used to separate between natural fCO_2 (run under pre-industrial atmospheric CO_2 mixing ratio) and contemporary fCO_2 (run under historical atmospheric CO_2 mixing ratio).

We extracted four metrics from the monthly ORCA025-MOPS output: SST, surface DIC, surface alkalinity (ALK), and sea surface fCO_2. fCO_2, which equals pCO_2 corrected for the non-ideal behaviour of the gas (Pfeil et al., 2013), determines the direction and magnitude of the air-sea CO_2 flux (Wanninkhof, 2014). We selected fCO_2 instead of air-sea CO_2 flux for our analysis since fCO_2 carries the imprints of temporal and spatial variability of carbon uptake and outgassing patterns without

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

being sensitive to uncertainties in gas exchange parameterizations. We furthermore decided to use only these three drivers (without including sea surface salinity) since they are known to drive most of the fCO_2 variability (Williams and Follows, 2011; Lauderdale et al., 2016). For both fCO_2 and DIC, we use their natural components rather than their contemporary components since we are not interested at this stage in including the anthropogenic carbon increase in the biome detection. The detected biomes are, however, substantially similar when using contemporary fCO_2 and DIC (Mohanty et al., 2023b).

2.2 Constructing the target-driver relationships

As stated in the previous section, the carbon biomes are detected based on the relationship that fCO_2 has with its three main drivers (SST, DIC, and ALK). We decided to build biomes on target-driver relationships rather than on the drivers themselves since our hypothesis is that this methodology is better capable of capturing the regionally specific carbon dynamics. To construct these local linear target-driver relationships, the global ocean is subdivided into boxes of $2° \times 2°$ dimension, and spatial relationships between the target variable, fCO_2, and its three drivers are computed over each box using multivariate linear regression (MLR), a supervised machine learning approach. We have chosen the grid size to be $2° \times 2°$ as target-driver relationships tend to be mostly linear on a smaller scale (Fig. A1). For each month from 1958 to 2018, the MLR inside each $2° \times 2°$ box is obtained according to Eq. (1):

$$fCO_2 = coef_{SST} \cdot SST + coef_{DIC} \cdot DIC + coef_{ALK} \cdot ALK + C \tag{1}$$

where $coef_{SST}, coef_{DIC}, coef_{ALK}$ are the regression coefficients of SST, DIC, and ALK, respectively, and C is the regression constant.

The regression coefficients (hereafter RCs) provide quantitative measures of the strength and direction of the relationships between the selected drivers (SST, DIC, ALK) and the target variable (fCO_2). A positive coefficient indicates a positive relationship (as the value of the independent driver increases, the value of the dependent target also tends to increase). In contrast, a negative coefficient indicates a negative relationship (as the value of the independent driver decreases, the value of the dependent target also tends to decrease). Larger coefficients suggest a stronger influence of the corresponding drivers in that particular $2° \times 2°$ grid box on fCO_2, making it more significant. In Fig. 2 a-c, an example of the RCs for January 2009 is shown, highlighting regions where similar relationships between the target and drivers exist.

We have chosen MLR over univariate linear regression because MLR allows for considering multiple factors simultaneously and can capture more complex relationships between predictors and the dependent variable than univariate regression, which considers only one predictor at a time. The RCs from MLR can differ in magnitude and even in sign from regression slopes computed using univariate regression (compare Fig. 2 and Fig. A1). This is not surprising since MLR attempts to optimize the R^2 score by fitting a hyperplane among the target and three drivers. As a result, RCs computed through MLR can be negative, even though the univariate target-driver relationship is positive (e.g., compare the fCO_2-ALK multilinear relationship in Fig. 2 c with the univariate relationship in Fig. A1 g. Finally, we have opted for MLR instead of other methods employing neural

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

networks or Random Forests (which build non-linear relationships between the target and drivers) since MLR is fast, and its output allows for a simpler and more explainable interpretation of the relationships between the drivers and the target, which are mostly linear over the $2^\circ \times 2^\circ$ boxes (Fig. A1).

2.3 Detection of carbon biomes using Adaptive Agglomerative Hierarchical Clustering

The RCs from the MLR (Sect. 2.2) serve as the foundation to detect the carbon biomes. Carbon biomes are detected based on the linear target-driver relationships obtained over $2^\circ \times 2^\circ$ boxes. We employ an unsupervised machine learning approach, namely the agglomerative hierarchical clustering (HC) technique (Müllner, 2011b) to construct a dendrogram based on the aggregation of the MLR coefficients (RC_{SST}, RC_{ALK}, RC_{DIC}). Before applying the HC technique, the RCs are first normalized (Fig. 2 d-f) to have a mean zero and standard deviation one. Normalizing RCs ensures that all variables contribute equally to the agglomeration process, regardless of their original scales. We select a hierarchical clustering method for two main reasons: 1) it prevents the necessity for a predetermined number of clusters, thus circumventing subjective bias, and 2) it simplifies the visual exploration of the resulting dendrogram, thereby aiding in the interpretation of pertinent clusters (carbon biomes) and underlying data distributions. Besides, hierarchical clustering enables the extraction of clusters with varying degrees of granularity when the dendrogram is cut at different levels (Lin et al., 2022).

The HC algorithm initiates by treating each normalized RCs of individual $2^\circ \times 2^\circ$ boxes as a distinct cluster. Subsequently, pairs of singleton clusters are iteratively merged until all clusters combine into one prominent cluster, encompassing all locally linear regression models. In conjunction with Euclidean Distance, Ward Linkage is employed to construct the dendrogram. The Euclidean Distance $ed(p,q)$ between two data points (i.e., Euclidean Distance between the RCs of two grid boxes) p and q is measured as shown in Eq. (2).

$$ed(p,q) = \sqrt{(q_{coef_{SST}} - p_{coef_{SST}})^2 + (q_{coef_{DIC}} - p_{coef_{DIC}})^2 + (q_{coef_{ALK}} - p_{coef_{ALK}})^2} \tag{2}$$

The Ward linkage is based on the Ward variance minimization algorithm (Müllner, 2011a) and it aims to lower the variance while combining two clusters, and the Euclidean Distance quantifies the distance between two clusters by measuring the increase in the sum of squares of individual clusters following their combination. The heights of the U-shaped links within the dendrogram signify this merging distance (i.e., the merging cost) in terms of the Euclidean distance. In other words, two branches merging at a higher height are more dissimilar (have a higher merging cost) than two branches merging at a lower height.

In a recent study (Mohanty et al., 2023a), we found that the conventional approach for identifying clusters within a dendrogram, entailing the selection of a specific distance value along the dendrogram's vertical axis, is not optimal for capturing the local statistical distributions, which vary substantially among branches. Because of unequal data distribution among the dendrogram branches, selecting a global cut at a lower distance would result in an excessive number of clusters on one branch, whereas picking a cut at a higher distance value would produce too few clusters on another branch (Fig. 3 a). To overcome this limitation, Mohanty et al. (2023a) devised a novel adaptive approach to provide local cuts to the dendrogram. The method

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

relies on the distance-variance selection technique and detects multiple local cuts on the dendrogram by considering both the compactness and similarity of the clusters. This algorithm operates based on two parameters, the change in distance ($\Delta Dist$) and the change in variance (ΔVar), that indicate the changes in distance and variance between different U-shaped links while traversing down the dendrogram from its root to its leaf nodes. This adaptive method serves the dual purpose of ensuring that the resultant clusters, representing carbon biomes, exhibit similarity and compactness, thus enhancing the robustness of cluster detection.

As we did not fix the number of provinces to be detected, we chose Bayesian Information Criterion (BIC) scores to select a meaningful partitioning (see evaluation in Appendix B). BIC scores are a statistical measure that augments the identification of an optimal number of clusters that effectively capture the underlying structure, balancing goodness of fit and model complexity. As detailed in Appendix B, we conducted 16 experiments with different parameter choices for $\Delta Dist$ and ΔVar. Based on our evaluation, we opted for $\Delta Dist$=25.0 and ΔVar=0.1, as we can already distinguish distinct clusters with the lowest difference in the distance with the fixed difference in variance.

2.4 Tracking carbon uptake provinces using Feed-forward Neural Networks

After having recognized the carbon biomes on the dendrogram, our intent now is to monitor their dynamics over time, revealing any evolving patterns within them. Since each biome is defined by the regression coefficients (RCs) of the drivers, obtaining localized RCs for subsequent months becomes mandatory. However, conducting adaptive clustering again in another month poses the challenge of connecting and matching two sets of identified biomes. Moreover, performing one-to-one matching based solely on RCs without location information is hard. Hence, we shifted our focus to training neural networks to monitor and track carbon biomes over time. Based on the data from the initial month of the temporal sequence, the clustering process detects distinct carbon biomes. These clusters are identified and labeled, forming the foundation for subsequent tracking. The RCs of subsequent months are fed to the Feed-forward neural networks (FNNs) as a classification problem to categorize them into distinct carbon uptake provinces.

We chose a feed-forward neural network for predicting the carbon biomes for two main reasons - non-linearity detection and making the deep learning model scalable. 1) Non-linearity detection: As the association between coefficients of SST, DIC, and ALK and the cluster labels is complex and cannot be defined by a linear relationship, FNNs can capture the underlying intricate patterns effectively. FNNs can learn elaborate interactions among input features(coefficients of SST, DIC, and ALK), which is crucial when predicting carbon biomes. For example, the impact of SST will vary from one spatial location to the other depending on the concentration of DIC or ALK, and a neural network is able to learn such relationships from the data. The FNNs' ability to model non-linear connections is promising when dealing with multidimensional environmental variables. 2) Scalability during model construction: FNNs are versatile and allow flexibility in designing the model architecture, i.e., the number of layers, neurons per layer, activation functions, etc. FNNs can be scaled up by adding more layers or neurons to accommodate larger datasets and to learn more complex relationships between drivers and their targets. This adaptability

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

enables the network to adapt to the sophistication of the data and biome labels, thus improving prediction accuracy.

Our tracking step involves training FNNs using the labeled carbon biomes obtained from the initial month as target labels. These neural networks are trained using this labeled dataset to learn the underlying patterns and relationships within the clusters. The input to the FNNs is the regression coefficients of SST, DIC, and ALK of January 2009. The aim is to impart the network with the ability to discern and predict the connection of data (regression coefficients of the drivers) to the identified clusters based on their respective characteristics. It focuses mainly on the features that define carbon biomes. Once trained, these neural networks are deployed to predict or assign cluster labels to the data points observed in the subsequent months of the temporal series. With the training process, the model learned to associate different RCs with different carbon regimes. Then, we use the same trained model to predict and track carbon provinces from January 1958 to December 2018.

For the tracking task, our FNN model comprised multiple dense layers with rectified linear unit (ReLU) (Agarap, 2018) activation functions interspersed with dropout layers for regularization. The model architecture consisted of an input layer followed by several hidden layers, each containing 64, 128, 256, 128, and 64 neurons, respectively. The choice of ReLU activation functions in the hidden layers facilitates the learning of non-linear relationships within the data. Furthermore, L2 regularization with a regularization parameter of 0.01 was applied to the kernel weights of each dense layer to mitigate overfitting. The output layer, comprising 7 neurons, utilized the softmax activation function (Goodfellow et al., 2016) to deliver a probability distribution over the 7 classes of carbon biomes. The model was trained using the ADAM (Adaptive Moment Estimation)(Kingma and Ba, 2017) optimizer with a learning rate of 0.001 and optimized for categorical cross-entropy loss. Additionally, performance metrics such as accuracy, precision, and recall were monitored during training to assess the model's predictive capabilities. The training process was conducted over 50 epochs with a batch size of 32, and the model's performance was evaluated on a validation dataset. Overall, the implemented FNN architecture with appropriate regularization and optimization techniques aimed to effectively capture the underlying patterns in the data and achieve robust classification performance. To prevent overfitting, early stopping with patience of 5 epochs was employed as a regularization technique. This led to finishing our training within 30 epochs.

We relied on metrics such as accuracy, precision, and recall to assess the performance of our FNN model in the context of multiclass prediction tasks. Accuracy measures the correctness of the model's predictions across all biome labels. It is calculated as the ratio of the number of correctly predicted cases to the total number of cases in the dataset. For a particular cluster label, precision is the measure of correctly predicted instances over all the instances that are predicted as that specific class or biome label. Furthermore, for a respective cluster label, recall is the proportion of correctly classified instances over all the instances of that specific class or biome label. The different parts of the FNN are explained in the Appendix D.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

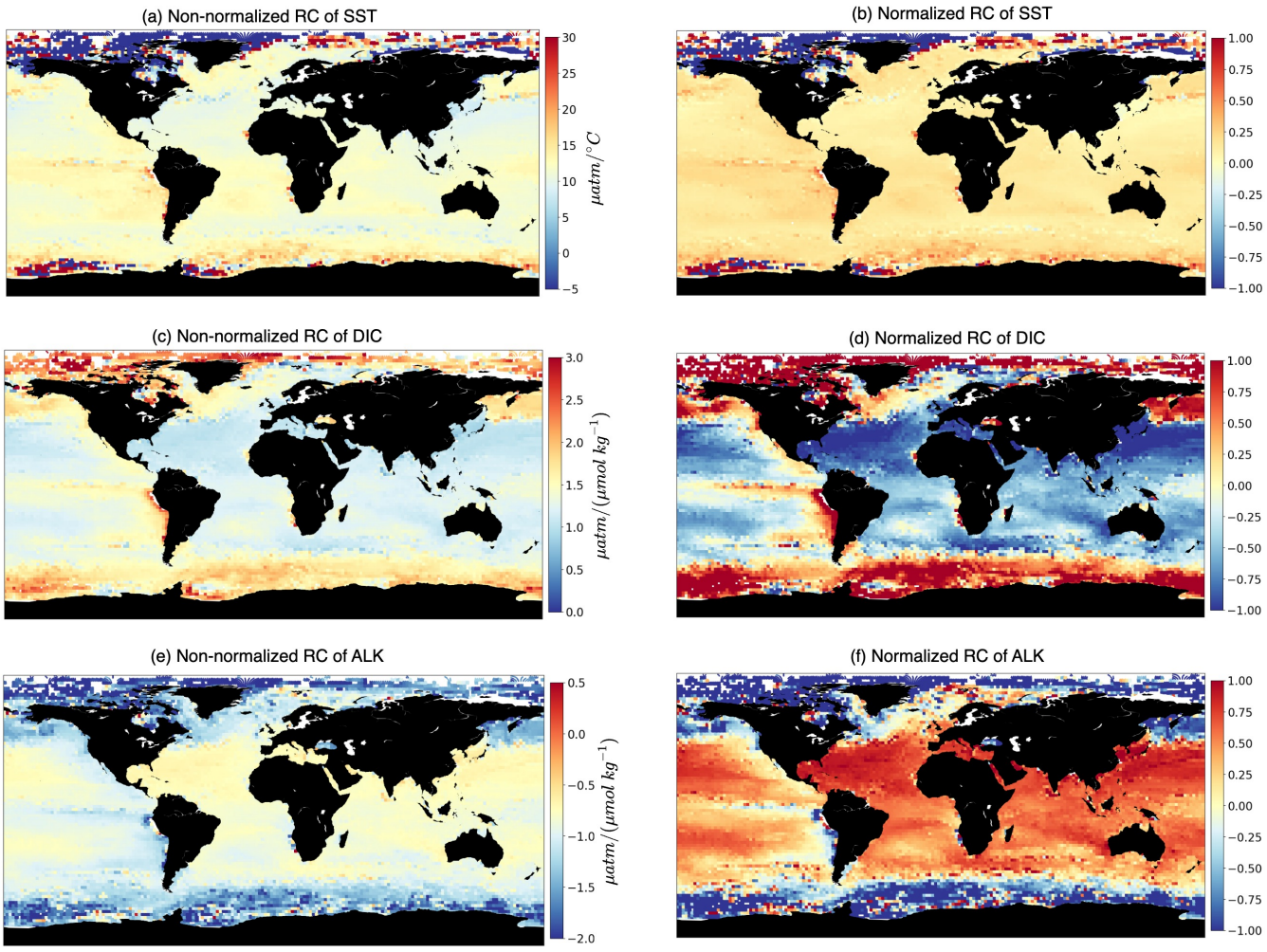

Figure 2. Spatial multivariate linear regression coefficients. The maps show non-normalized (left column) and normalized (right column) regression coefficients (RCs) computed over $2^\circ \times 2^\circ$ boxes of fCO_2 with respect to its three drivers, i.e. (a,b) SST, (c,d) surface DIC, and (e,f) surface ALK.

3 Results

3.1 Detection of carbon biomes

The relationship between fCO_2 and its three drivers (hereafter called RC_{SST}, RC_{DIC}, RC_{ALK}) exhibits large spatial variations (Fig. 2), indicative of different dynamics acting over different regions of the global ocean. We find an expected positive relationship of fCO_2 with respect to SST over large swaths of the global ocean (Fig. 2 a) since higher SST reduces the seawater CO_2 solubility and thus enhances its fugacity at the ocean surface. Over most of the global ocean RC_{SST} is typically

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

included between 10 and 16 $\mu atm°C^{-1}$. When pCO_2 is affected only by temperature, Takahashi et al. (1993) determined a relative variation in pCO_2 of 0.0423 $°C^{-1}$, equivalent to 16.9 $\mu atm°C^{-1}$ for a pCO_2 value of 400 μatm. The deviation of our simulated RC_{SST} from this expected value indicates the influence of non-thermal processes on fCO_2. Also, we are considering RCs from multilinear regressions which, as mentioned in Sect. 2.2, may differ from the univariate perspective. At polar latitudes, RC_{SST} values are much higher, are mostly negative, and are characterized by higher spatial variability. In this respect, the following points should be considered: i) SST in polar regions deviates only slightly from the freezing temperature (Fig. A1), which leads to high RC_{SST} even in the presence of moderate variations in fCO_2; ii) the rather counter-intuitive negative RC_{SST} values can be understood by considering that when an increase in surface temperature melts sea ice, the ensuing air-sea CO_2 exchanges and phytoplankton growth lead to a reduction of fCO_2; iii) leads and fractures in sea ice may generate a strong spatial variability in fCO_2 due to the array of processes (air-sea CO_2 exchanges, biological productivity) which are set in motion when a previously sea ice-covered region is exposed to the atmosphere. RC_{DIC} is positive basically everywhere, indicating an expected positive dependence of fCO_2 on DIC. RC_{ALK} is instead mostly negative since increases in ALK have a buffering effect on fCO_2. RC_{DIC} and RC_{ALK} therefore tend to have opposite and specular effects on fCO_2 (Fig. 4 c).

The RCs resulting from the MLR have been normalized before being fed to the hierarchical clustering (HC) algorithm (Sect. 2.3) so that, over the globe, RCs for each driver have mean zero and standard deviation one (Fig. 2, right column). The normalized RCs (hereafter RC^{norm}) can thus be understood as anomalies with respect to the global mean. As an example, negative values of RC^{norm}_{DIC} over the subtropical gyres indicate a dependence of fCO_2 on DIC that is lower than the global average. The HC algorithm run on RC^{norm} values of January 2009 produces the dendrogram shown in Fig. 3 a. As explained in Sect. 2.3, the resulting clusters depend on the choice of where to cut the dendrogram. Instead of selecting a fixed height for the dendrogram cut, we used the distance-variance selection methodology (Mohanty et al., 2023a) to define local cuts based on the underlying data distribution. This procedure yields a total of seven carbon biomes, each possessing analogous characteristics in terms of target-driver relationships (Fig. 4 a-c). The dendrogram detects three main branches, exhibiting distinct combinations of RC^{norm} values (Fig. 4 d) and specific geographical locations (Fig. 3 b).

The leftmost branch on the dendrogram detaches itself from the other branches at an elevated height, indicating that it is very dissimilar from the other two branches. This branch is located in all months at polar latitudes, especially in the Arctic (blue colors in Fig. 3 b, Fig. 5 a, b and Fig. C1), and is hereafter called the ICE branch. The ICE branch distinguishes itself by having strongly negative RC^{norm}_{SST} and RC^{norm}_{ALK} and strongly positive RC^{norm}_{DIC} (Fig. 4). The ICE branch is thus characterized by large spatial gradients in fCO_2 and its drivers, negative dependence from SST, and a strong positive dependence on DIC. Our analysis thus suggests that the ICE branch is mostly driven by non-thermal processes but is characterized by a distinctive dependence of fCO_2 on SST due to the interaction with sea ice (as explained above). Based on our parameter choice, two ICE sub-branches are identified as biomes (ICE I and ICE II). The two ICE biomes are characterized by different magnitudes (but the same sign) of the RC^{norm} values. Specifically, ICE I corresponds to fewer boxes and is characterized by the most extreme values of the RC^{norm}, whereas ICE II covers larger parts of the polar ocean and has somewhat lower magnitudes of RC^{norm}

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

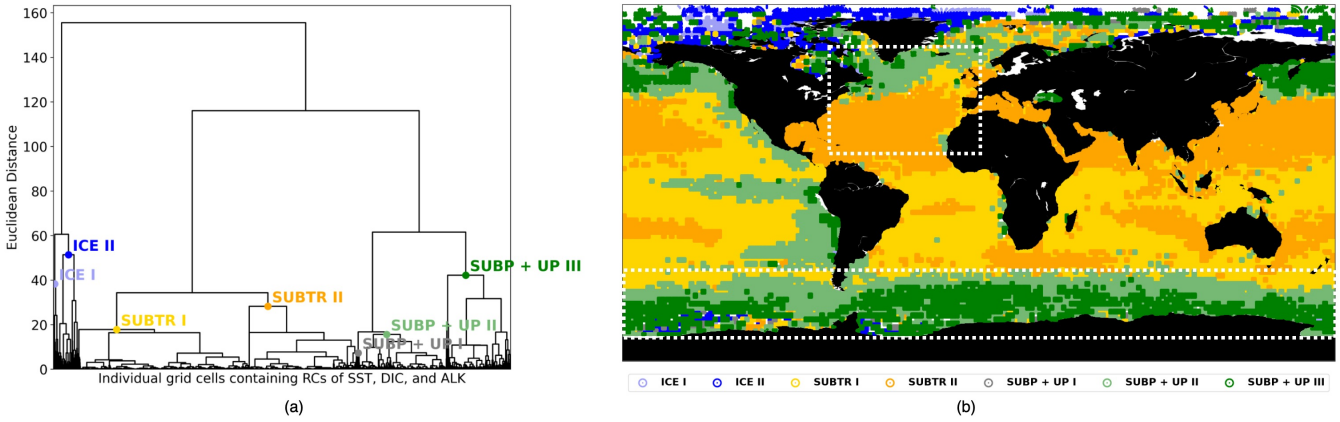

Figure 3. Carbon biomes in January 2009 detected through Hierarchical Clustering (HC) (a) Dendrogram resulting from the HC, with local cuts based on the distance-variance selection methodology. The text indicates the names of the detected seven clusters (i.e., the carbon biomes). (b) Geographical location of the detected clusters. The white boxes illustrate the basins - the North Atlantic and the Southern Ocean - which will be analyzed in Sect. 3.2.

values.

The rightmost branch on the dendrogram is geographically located at subpolar latitudes as well as in upwelling regions, i.e., Equatorial Pacific and coastal upwelling areas (green colors in Fig. 3 b, Fig. 5 c-f and Fig. C1). This branch, which we call SUBP+UP, distinguishes itself from the other two branches by having positive values of RC^{norm}_{SST} and RC^{norm}_{DIC} and negative values of RC^{norm}_{ALK} (Fig. 4 d). RC^{norm}_{DIC} always has a larger magnitude than RC_{SST}, suggesting increased importance of non-thermal processes in driving fCO_2 variability in those regions. The selected parameter settings detect three SUBP+UP biomes (SUBP+UP I, SUBP+UP II, SUBP+UP III), each characterized by different flavours of the RC^{norm} combinations. SUBP+UP III is characterized by the highest RC^{norm}_{DIC} values of this branch, suggestive of a stronger dependence of fCO_2 on non-thermal processes, and is predominantly located in high-latitude subpolar areas (Fig. 5 c, d). SUBP+UP II is characterized by small values of all three RC^{norm} values (i.e., they are closer to the global mean) and often marks the transition towards mid-latitude and subtropical regimes (5 e, f). SUBP+UP I has a strongly positive RC^{norm}_{SST} and is found at the edge with the ICE branch (Fig. 3 b). However, since it contained only 0.55% of the total sample size, it was discarded from the tracking process (see Sect. 3.2).

The middle branch in the dendrogram is geographically located in the Tropics, in the subtropical gyres as well as on large parts of the North Atlantic (orange colors in Fig. 3 b, Fig. 5 g-j and Fig. C1). This branch, which we call SUBTR, distinguishes itself from the other two branches by having positive values of RC^{norm}_{SST} and RC^{norm}_{ALK} and negative values of RC^{norm}_{DIC} (Fig. 4 d). Negative (positive) values of RC^{norm}_{DIC} (RC^{norm}_{ALK}) are indicative of regions where the dependence of fCO_2 on DIC and ALK is lower than the global average. This combination of RC values suggests the enhanced importance of thermal processes

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

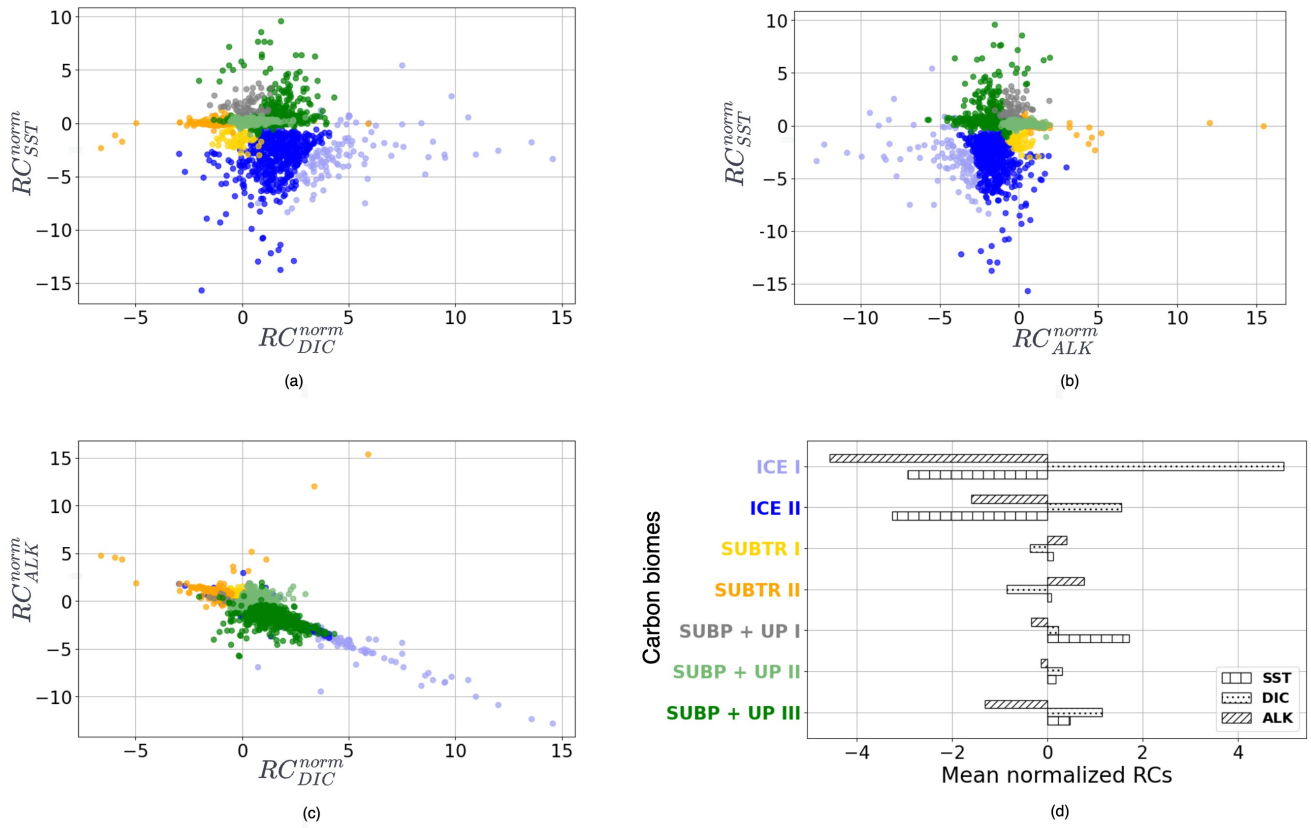

Figure 4. Distribution of normalized regression coefficients (RC^{norm}) in the carbon biomes detected for January 2009. (a) RC^{norm}_{SST} against RC^{norm}_{DIC}, (b) RC^{norm}_{SST} against RC^{norm}_{ALK}, (c) RC^{norm}_{ALK} against RC^{norm}_{DIC}. (d) RC^{norm}_{SST}, RC^{norm}_{DIC}, and RC^{norm}_{ALK} averaged over the carbon biomes. Colors indicate the detected carbon biomes: ICE biomes in blue shading, SUBP+UP biomes in green shading, and SUBTR biomes in orange shading. SUBP+UP I was not tracked (see main text) and is therefore shown in grey.

in driving fCO_2 variability in the SUBTR branch. The HC algorithm detects two SUBTR biomes (SUBTR I and SUBTR II), characterized by different flavours of RC^{norm} combinations. In SUBTR II, RC^{norm}_{DIC} is particularly low and is located in the central parts of the subtropical gyres (Fig. 5 g, h). SUBTR II is, therefore, identified as the biome most strongly driven by thermal processes. SUBTR I is instead characterized by more moderate values of the RC^{norm}s and is found over large swaths of the global ocean, including subtropical, middle, and even subpolar latitudes (Fig. 5 i, j).

Finally, it should be noted that the RC^{norm} values are more compact over the SUBTR biomes and SUBP+UP II biomes than over the remaining biomes. This can be visualized in Fig. 4 a-c, where a large variance around the mean is found for the ICE and SUBP+UP I biomes. This different statistical distribution of the data within the different dendrogram branches is one of the reasons for which, for this data set, local cuts to the dendrogram work better than a fixed global cut.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

3.2 Tracking of carbon biomes

Temporal tracking of carbon biomes was performed using a forward-feed neural network (FNN) model. The input to the FNN model is the RC^{norm} values obtained through MLR, and the target labels are the seven carbon biomes detected on the dendrogram (Fig. 1). As explained in Sect. 2.4 and further discussed in Sect. 4, we selected January 2009 as the month on which to run the HC and produce the biome labels and temporally predicted all other months during the 1958-2018 period using FNN. 80% of the RC^{norm} values were used for training and validation, and the remaining twenty percent were used to test the learned model. During training, the validation accuracy reached 94.81%, precision was 95.22%, and recall stood at 94.06%. The good results from our FNN model on the training data were also reflected in the test data, where we obtained a test accuracy of 94.83%, precision of 95.39%, and recall of 94.5%. We have chosen the biomes predicted by the FNN with the highest probability. More information on FNN evaluation can be found in Appendix D, including the confusion matrix on the test set, highlighting the above-mentioned values of test accuracy, precision, and recall (Fig. D1 e). The biome SUBP+UP I contained only 0.55% of the total sample size and was excluded by the FNN model while learning the association between the biome label and the RCs.

The tracking of carbon biomes allows us to address the following questions: Q1) Does the geographical location of the biomes vary significantly from year to year? Q2) Are there seasonal fluctuations of the biome coverage in the course of a year? Q3) Are there year-to-year changes and long-term trends in the biome coverage in response to climate variability and change? It should be stressed that both the detection and the tracking are solely based on the RC^{norm}_{SST}, RC^{norm}_{DIC}, and RC^{norm}_{ALK} values, without any prior knowledge of the geographical locations of the biomes. The temporal tracking is, therefore, a purely location-agnostic process that depends only on the carbon dynamics specific to that region.

Fig. 5 shows the percentage coverage of the 5 main biomes (SUBTR I, SUBTR II, SUBP+UP II, SUBP+UP III, ICE III) computed over the 1958-2018 period, distinguishing between January and July. All biomes are characterized by core regions where the coverage reaches 100%, peripheral regions where the biome is found only in some years, and external regions where the biome is never detected. From Fig. 5 and Fig. 6 c-f, we see that the biome coverage is overall consistent from year to year, indicating that the trained FNN is able to predict over time regions with similar RC patterns consistently. Only a couple of years (1969 and 2010) were found to be inconsistent with the overall pattern.

Distinctive features characterize the different biomes in terms of seasonality. SUBTR II (Fig. 5 g-h), the most thermally-driven biome, seasonally dominates in the winter months in the subtropical gyres (i.e., in the northern hemisphere in January and in the southern hemisphere in July). This is possibly related to the fact that in winter when there is little biological productivity, fCO_2 is mostly driven by SST. SUBTR I (Fig. 5 i-j), covering large swaths of the global ocean, has a tendency for higher coverage in the summer months. This is potentially because SUBTR I has a less pronounced thermal control than SUBTR II and, therefore, dominates in those locations and months for which non-thermal processes also play some role. SUBP+UP III has its greatest extent in the winter months of the subpolar and high latitudes when strong non-thermal controls (e.g., convec-

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

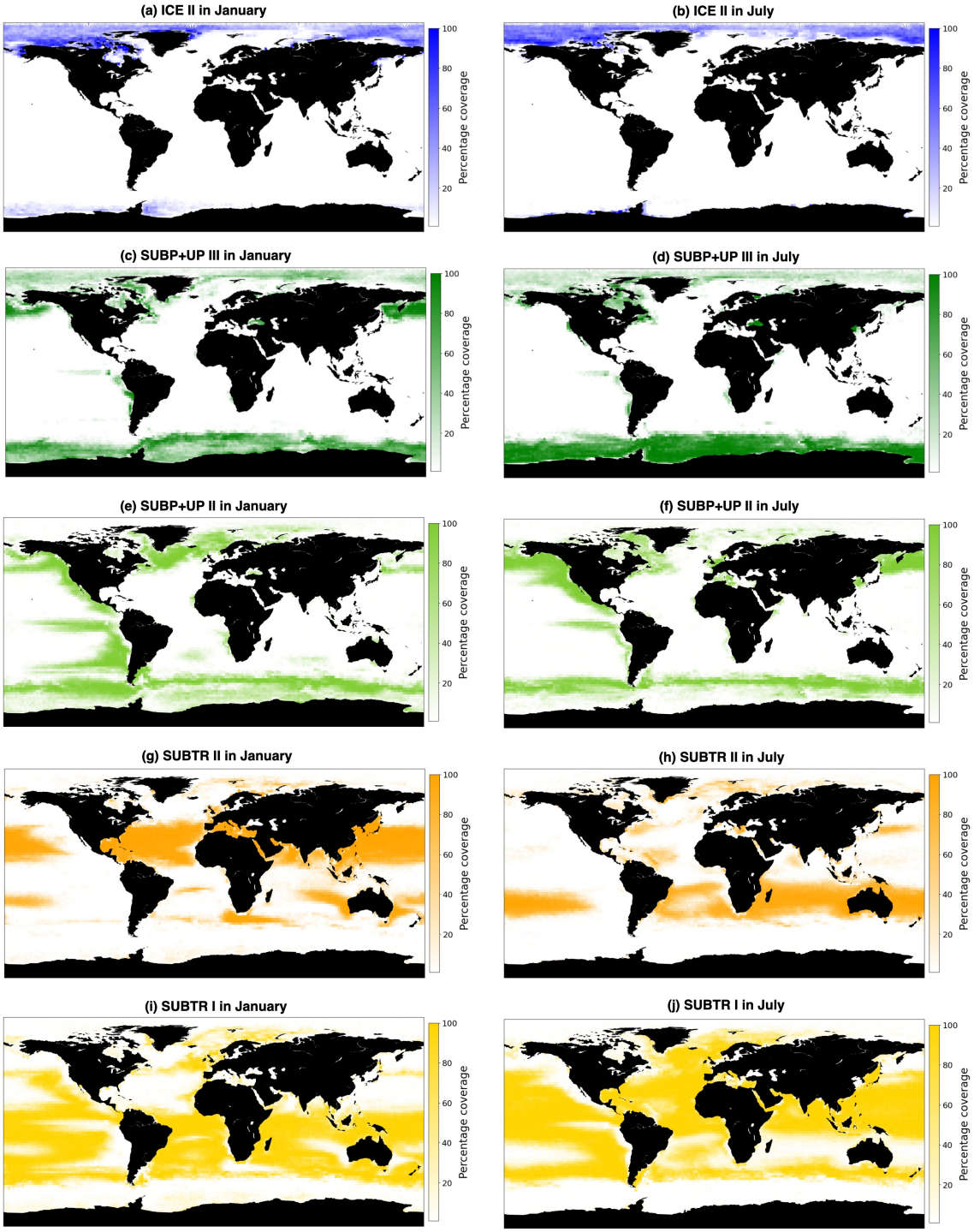

Figure 5. Percentage coverage of carbon biomes over the years 1958 to 2018, where a value of 100% indicates that the biome is present in each simulation year in the 2x2 box and a value of 0% that it is never present in the 2x2 box. Shown are the main 5 carbon biomes: a,b) ICE II, c,d) SUBP+UP III, e,f) SUBP+UP II, g,h) SUBTR II, and i,j) SUBTR I in January (left column) and July (right column).

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

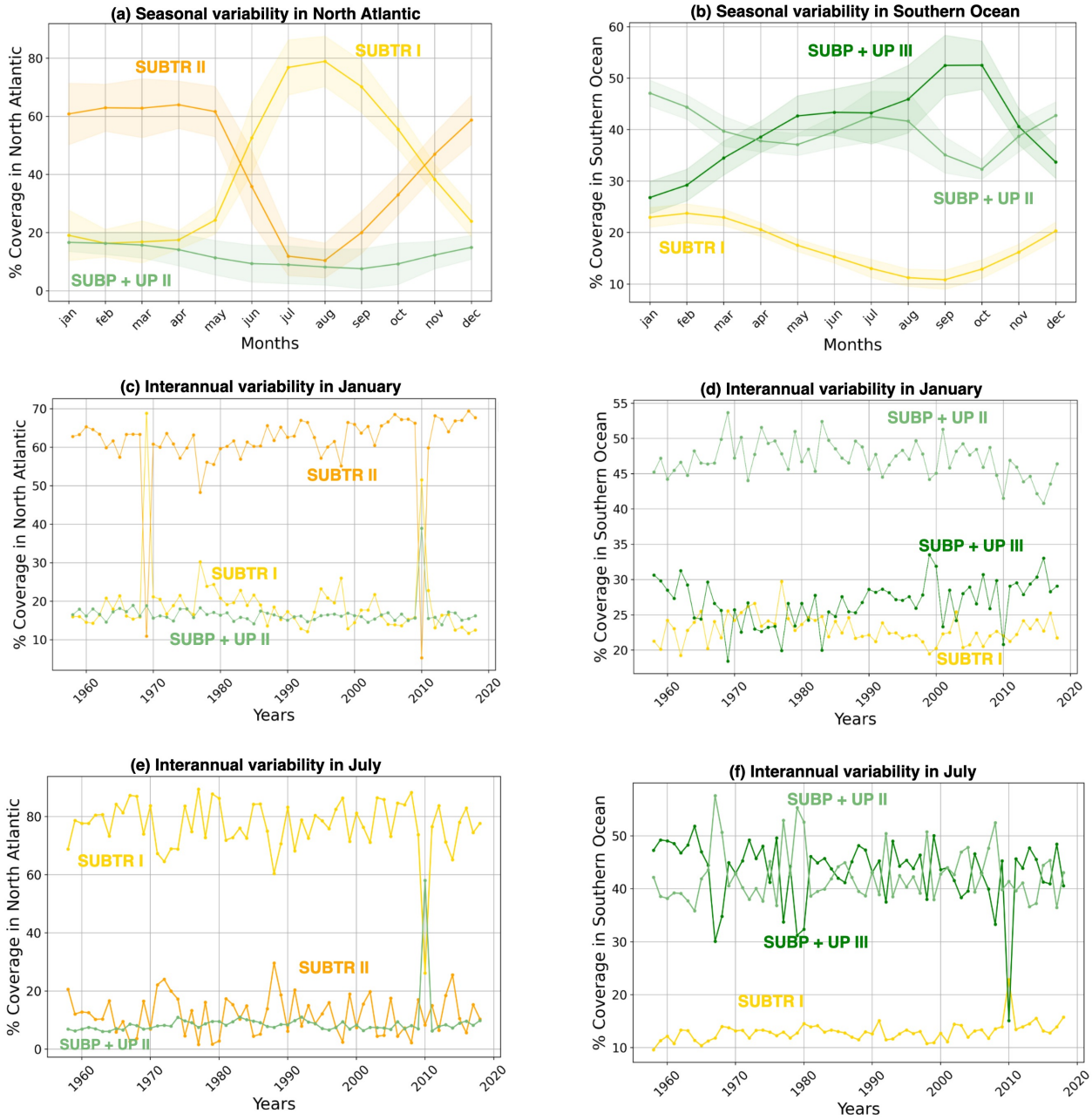

Figure 6. Seasonal and interannual variations of carbon biomes coverage. Shown is the percentage coverage of the three main biomes over the North Atlantic (left column) and Southern Ocean (right column). The limits of the basins are shown in Fig. 3 b. The percentage coverage is computed as the weighted area average of the carbon biome divided by the total area of the basin so that a value of 100% would be achieved when the biome covers the whole basin. (a,b) Mean seasonal cycle of the carbon biome coverage, with shading indicating the standard deviation across the 61 simulation years. (c-f) Weighted area average of the percentage coverage of carbon biomes over 1958-2018 in the months of c,d) January and e,f) July.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

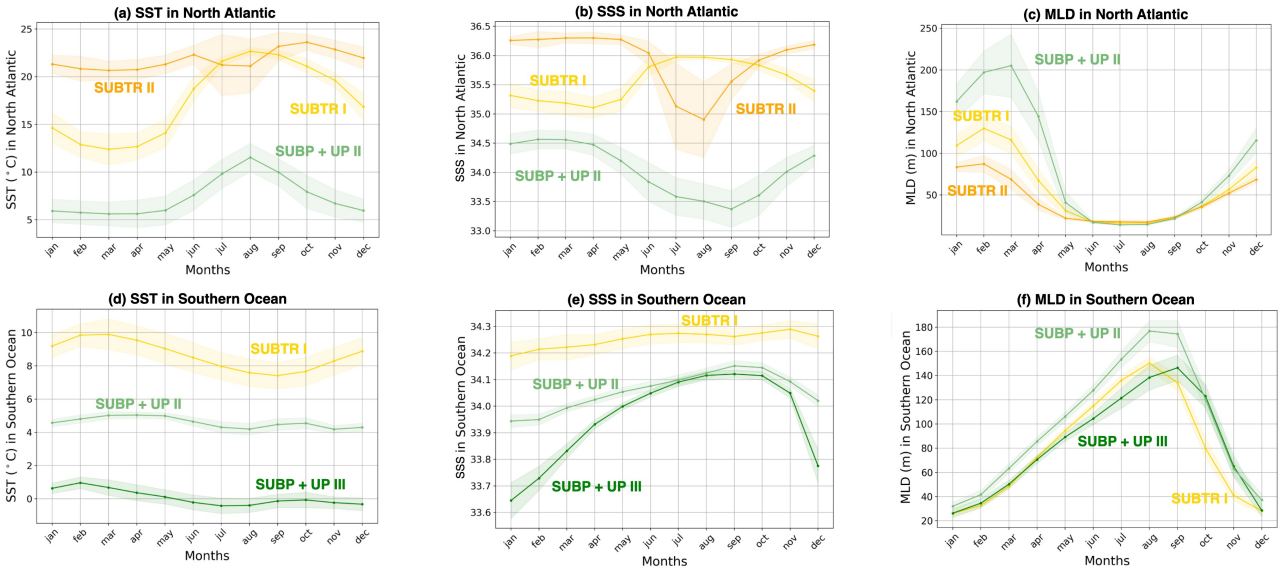

Figure 7. Seasonality of environmental parameters over biomes in the North Atlantic and Southern Ocean. Mean seasonal cycle over 1958-2018 of (a,d) SST, (b,e) sea surface salinity (SSS), and (c,f) mixed layer depth (MLD), averaged within the three main carbon biomes in the North Atlantic (top row) and Southern Ocean (bottom row). The limits of the basins are shown in Fig. 3 b. Shading indicates the standard deviation over the 61 simulation years.

tion and upwelling) drive fCO_2. In the summer months, SUBP+UP II, with its somewhat more nuanced non-thermal control, occupies those subpolar areas occupied by SUBP+UP III in winter. An exception is the North Atlantic, as will described later on. ICE II is found almost exclusively in the Arctic Ocean, with little seasonal variations.

The biome coverage and underlying environmental properties are further explored on two basins of climatic relevance (contours in 3 b): 1) the North Atlantic between 75°W-0°, and 10°N-70°N and 2) the Southern Ocean between 45°S and 77°S. Fig. 6 a, b shows the mean seasonal evolution of the percentage coverage of the three main biomes for each basin, with the standard deviation computed over the 61 simulation years shown as shading. A 100% coverage for a biome thus would indicate that the biome covers the whole basin, whereas a 0% coverage indicates that the biome never occurs in the basin. To gain a better perspective of the environmental conditions affecting each biome, we further analyzed the mean seasonal cycle of SST, sea surface salinity (SSS), and mixed layer depth (MLD) over each of the three main biomes detected in the two basins (Fig. 7). In analyzing these seasonal cycles of environmental parameters, it should be remembered that the area on which the averages are computed changes with time, which can complicate the interpretation. However, we believe this analysis is still useful as it allows us to better characterize the biomes in terms of environmental properties.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

The North Atlantic (NATL) exhibits a sharp divide between the subtropical and subpolar gyres, separated by the North Atlantic Current (Fig. 5). SUBTR I and SUBTR II seasonally compete, with SUBTR I dominating in summer (around 80% of the whole NATL) and SUBTR II in winter (around 60% of the whole NATL) (Fig. 6 a). SUBTR II shows overall the highest SST and SSS and the lowest winter MLD (Fig. 7 a-c), consistently with the strongly stratified subtropical gyre conditions. A dip in SST and SSS in July-August, with elevated uncertainty levels, is connected with the very low coverage of SUBTR II in these months (Fig. 5 h), an aspect which amplifies the importance of limited coastal and high-latitude regions. SUBP+UP II, the third most represented biome in the NATL, gains its highest coverage in winter (around 20% of the whole NATL). SUBP+UP II is mostly found in the subpolar NATL (Fig. 5 e,f), where it seasonally competes with SUBTR I, which is instead most present in summer. Interestingly, in the eastern parts of the subpolar North Atlantic, SUBTR I dominates in both winter and summer (Fig. 5 i,j), possibly because of the influence of the North Atlantic Current bringing subtropical waters to those areas. In other words, from a carbon perspective, the eastern subpolar NATL is more similar to mid-latitude and subtropical domains than the western subpolar NATL. As expected, SUBP+UP II has lower SST (with a peak and summer) and deepest winter MLD with respect to the SUBTR biomes (Fig. 7 a-c). The dip in SSS in the summer months is likely related to the fact that in summer SUBP+UP II mostly occupies sea ice melting regions (Fig. 5 f).

The Southern Ocean (SO) exhibits a seasonal competition of SUBTR I, SUBP+UP II, and SUBP+UP III (Fig. 6 b). SUBP+UP III is the coldest biome (Fig. 7 d), and it occupies the subpolar regions dominated by wind-driven upwelling and moderately deep mixed layers (Fig. 7 f). It is most extended in winter when it covers more than 50% of the whole SO. The enhanced SSS in the winter season (Fig. 7 e) is consistent with the increased wind-driven upwelling of relatively salty Circumpolar Deep Water in the winter months. In the summer months, SUBP+UP III recedes, and SUBP+UP II gains more relevance (Fig. 6 b). SUBP+UP II is overall warmer than SUBP+UP III and is found in mid-latitude regions characterized by deep winter MLD (Sallée et al., 2021). In all months, SUBP+UP II typically follows the path of the Antarctic Circumpolar Current as well as water mass formation areas to the north of it (Fig. 5 e, f). SUBTR I, with its higher SST and SSS, is a more thermally-driven biome with the highest coverage in the summer months (Fig. 6 b).

When performing the carbon biome tracking over the whole 1958-2018 period, we find that some biomes have expanded, whereas some have shrunk. In the NATL, SUBTR I and SUBTR II, which compete on seasonal time scales, show anticorrelated interannual variability over the 1958-2018 period. In winter, SUBTR I and SUBTR II also show diverging trends since the 1970s, with SUBTR II expanding by about 10% at the expense of SUBTR I. This might be related to a concomitant increase in SST over the North Atlantic subtropical gyre (Bulgin et al., 2020), which might have enhanced the thermal control of SST on fCO_2. In the Southern Ocean, SUBP+UP II and SUBP+UP III show competing variabilities on interannual time scales, and - since the late 1960s - an overall 10% increase in SUBP+UP III in summer, at the expenses of SUBP+UP II. This might be related to the concomitant increase in Southern Hemisphere westerly winds (Swart et al., 2015), which has created more favorable conditions for DIC upwelling and therefore enhanced the non-thermal control on fCO_2 (Gruber et al., 2023).

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

4 Discussion and Conclusions

In the framework of a rapidly evolving climate and ocean carbon cycle, the aim of this study was to develop a machine learning tool to detect ocean carbon biomes and track them under seasonally and interannually varying environmental conditions. We defined carbon biomes as regions having consistent relations between surface CO_2 fugacity (fCO_2) and its main drivers (temperature, dissolved inorganic carbon, and alkalinity). With a combination of localized multilinear regression (MLR) models, agglomerative hierarchical clustering (HC), and a forward-feed neural network, we were able to successfully detect and track ocean carbon biomes both seasonally and over the 1958-2018 period. The key features and novelties in our study are: i) employing target-driver relationships as an input to the HC algorithm, instead of directly environmental parameters, ii) using a distribution-aware clustering method to group these relationships, ensuring that each group is compact and cohesive, with similar internal relationships, and iii) building a tool to track the evolution of the detected clusters or carbon regimes over time. This methodology enabled us to detect well-defined carbon biomes (representative of subtropical, upwelling, subpolar, and sea ice-covered regimes) whose geographical location and spatial extent responded meaningfully to seasonal and interannually varying environmental conditions. It is to be stressed that the detection and tracking of the carbon biomes were done entirely without providing any information about the geographical location of the biomes. The fact that the detection uncovered biomes with meaningful geographical characteristics is purely the result of the different ways in which the fCO_2 reacts to its drivers, which in turn is intimately intertwined with the underlying ocean dynamics.

There are a number of considerations that need to be made regarding our methodology and some aspects where we see room for improvement in future studies. In order to construct labels needed to train the FNN, we needed to select a specific month and year (in this case, we selected January 2009). The question arises as to whether our results are sensitive to the choice of this selection. The regression coefficients (RCs) computed with MLR show a smooth transition from one month to another and are relatively invariant from one year to the next. Building upon this stability, we found that the carbon biomes detected for other random years are comparable to that based on the January 2009 baseline (not shown). However, we acknowledge that the month/year selection introduces a subjectivity that would be preferable to avoid. The hypothesis that the selection of different months and years may lead to different biome segmentations cannot be excluded, and we suggest that future work should investigate the volatility or stability of the chosen reference. Another option would have been to perform clustering on a monthly basis to detect the carbon biomes. However, this option comes with additional caveats. Firstly, employing HC with Ward linkage over an extensive spatial domain enclosing more than ten thousand grid boxes is intrinsically time-intensive. Secondly, the absence of an established technique to successfully match clusters between successive months beyond visual inspection delivers a big challenge. Given the broad temporal scope of our analysis traversing 732 months (61 years × 12 months), manual tracking via visual analysis would have been impractical and subjective. Therefore, we adopted a practical strategy where we labeled the seven detected clusters for one specific year and month and employed a neural network with predictive capacities to learn the intricate associations between the RCs and the assigned cluster labels. With the deep learning model successfully having learned the subtle relationships between these variables, it is able to predict cluster labels based on

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

input regression coefficients, along with corresponding probabilities. Subsequently, the carbon biomes predicted by the FNN with the highest probability are selected as the most probable clusters.

Another aspect to consider regards the selection of feed-forward neural networks (FNN), which come with several challenges that can hinder their effectiveness in certain applications. One primary concern is the complexity of FNN architectures, which may lead to overfitting and difficulty in interpretation due to the large number of parameters involved. Further, FNNs are often criticized for their black-box personality (Irrgang et al., 2021), as they lack clarity in the decision-making approach, making it demanding to comprehend how inputs are connected to outputs. While convolutional neural networks (CNNs) (LeCun et al., 2015) and long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks are powerful alternatives to FNNs in certain contexts, they may not be well-suited for all tasks. CNNs are outstanding at pulling spatial features from data such as images, audio, videos, etc. (Goodfellow et al., 2016), and LSTM networks are adept at apprehending temporal dependencies in sequential data. However, these methods may not have been suitable for our study. CNNs and LSTMs would have required not just one (as we utilized in our FNN model) but rather hundreds of annotated maps of global oceans depicting ocean biomes. Considering that we are operating with 61x12=732 maps of RCs, either clustering or manually labeling half of them with respective biomes would have been required to train these models, while the remaining maps would have been utilized for biome tracking. Consequently, both clustering and manually determining the same biomes across multiple months would have been particularly time-consuming, compounded by the additional training time needed for these models. Essentially, we would have needed to specify the presence and locations of carbon biomes across multiple months even before initiating the tracking process.

The carbon biomes found in this study are geographically analogous to those found in past classification studies that focused on the ocean carbon cycle. Jones and Ito (2019) ran Gaussian Mixture Modeling (GMM) directly on different physical variables used to express the surface carbon budget, and distinguished between four classes of biomes: areas of strong vertical mixing and upwelling, areas of weak productivity, unproductive subtropical gyres and areas of coastal upwelling. Couespel et al. (2024) built instead temporal target-driver relationships between air-sea CO_2 flux and biogeochemical predictors over a time series and used GMM to cluster carbon regimes echoing inter-annual shifts. Landschützer et al. (2016) used self-organizing maps to cluster non-linear relationships between CO_2 partial pressure and relevant environmental parameters. Even though based on different classification methods and input parameters, the biomes found in these past studies bear a resemblance to those found here. It can, therefore, be concluded that different carbon properties share a similar dependence on the underlying dynamical context (e.g., oligotrophic subtropical gyres vs. productive upwelling regions), which in our case translates into distinct spatial target-driver relationships between fCO_2 and its drivers. Also, similarly to Jones and Ito (2019) and Couespel et al. (2024), we find that the same biome can occur in distinct geographical locations. This can be seen for e.g. the SUBP+UP biomes, which are found in both tropical and subpolar upwelling areas. The classification used here thus differs from that proposed by Fay and McKinley (2014), which involves splitting the ocean into four major ocean basins (Atlantic, Pacific, Indian, and Southern Ocean), followed by the application of criteria based on specific variable ranges (such as SST, MLD, chlorophyll concentration, and sea ice). The biomes found by Fay and McKinley (2014), which are widely used to evaluate and compare ocean

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

carbon cycle models (DeVries et al., 2023), thus have a clear geographical separation, which instead somewhat breaks down in the clustering method used here. This result suggests that geographically separated regions can be more closely connected regarding ocean carbon dynamics than their geographical location would suggest.

In this investigation, we find three main dynamical branches (sea ice-covered, subpolar+upwelling, and tropical+subtropical), each characterized by different combinations of spatial target-driver relationships and underlying environmental parameters. The three branches differ in how spatial changes in fCO_2 depend on spatial changes in SST and DIC. The subtropical branch has a weak dependence on DIC and stronger on SST, the subpolar+upwelling branch shows a strong dependence on DIC and a weaker dependence on SST, and the sea ice-covered regions have a strong dependence on both SST and DIC. We used the terms "thermal" and "non-thermal" controls of fCO_2 to give a semantic interpretation of the different regimes. We, however, acknowledge that this denomination is different with respect to past studies, which used these terms to separate regions where the seasonality of fCO_2 was in phase with SST (thermal control) or DIC (non-thermal control) (Takahashi et al., 2002; Prend et al., 2022). To the best of our knowledge, we are unaware of studies comparing how the connection between fCO_2 and its drivers differs spatially vs. seasonally. Owing to sparse observations, studying the spatial connection between fCO_2 and its drivers is challenging, even though an increasing amount of continuous observations, e.g., through saildrone (Sutton et al., 2021) and sailboat measurements (e.g., Behncke et al. (2024)), could soon change this. Future work could explore how the detection and tracking of carbon biomes would differ when using seasonal target-driver relationships instead of spatial target-driver relationships. For instance, we can speculate that seasonal target-driver relationships would probably not have yielded such large regression coefficients in the Arctic so the ICE biome would likely not have been so distant from the other biomes.

The rapidly changing climate conditions pose a significant threat to the ocean carbon cycle, and machine learning techniques are increasingly rising to the challenge of detecting ocean patterns, predicting changes, and making analysis processes more efficient (Irrgang et al., 2021; Couespel et al., 2024; Krasting et al., 2022). This study marks a step forward in the research field since it provides a robust tool for the temporal tracking of marine carbon biomes. We found that the biome coverage reacts consistently to the seasonality of environmental parameters, such as SST, mixing, and upwelling. We also found that the biome coverage can change over the years, possibly in connection with multi-decadal trends in wind and temperature. The possibility to detect and track meaningful carbon structures within the global ocean opens several opportunities. First of all, it provides a tool for narrowing down the massive volume of data produced by ocean biogeochemistry and Earth System Models and to focus on the evolution of coherent structures in the ocean instead of properties over every grid point. This approach, thus, could facilitate the evaluation and intercomparison of ocean biogeochemistry and Earth System Models in a compact and systematic fashion. When applied to future scenario runs, coherent detection and tracking of carbon biomes could yield an alternative prediction of the future carbon cycle evolution while at the same time providing a strong interpretation framework of the underlying carbon dynamics. Since the approach relies solely on methods unaware of geographical coordinates, it is best positioned to capture the fluidity of the biomes in response to e.g., changing sea ice and stratification patterns. Against

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

the backdrop of a rapidly changing climate as well as machine learning techniques, the approach presented here - combining a novel adaptive clustering technique and a robust tracking algorithm - is thus well suited to address these challenges.

Code availability. Our analysis was conducted in Python. Information on how to 1) extract and process ocean model outputs (NETCDF files), and 2) rerun the experiments of detection and tracking can be found here:
https://github.com/swemoh/Detection-and-Tracking-of-Carbon-Biomes

Data availability. A subset of the ORCA025-MOPS output used for this study will be made freely available on the GEOMAR OPeNDAP Server after acceptance.

Appendix A: Set up for Localized Target-driver Relationship

Our goal in our study was to collect the localized target-driver relationships instead of a global association. The Fig. A1 a illustrates $2^\circ \times 2^\circ$ boxes on the ocean surface, inside which we build the localized linear relationships between the target and the drivers in individual months over 61 years. The scatter plots b-d show the distribution of different drivers with respect to fCO_2 in the grid box highlighted in cyan from the Arctic basin. Here, RC_{SST} = -102.12 $\mu atm\ ^\circ C^{-1}$, RC_{DIC} = 0.689 $\mu atm\ \mu mol\ kg^{-1}$, and RC_{ALK} = -0.485 $\mu atm\ \mu mol\ kg^{-1}$. Similarly, the scatter plots e-g present the distribution of three drivers with respect to fCO_2 in the grid box highlighted in red from the North Atlantic basin. Here, RC_{SST} = 0.081 $\mu atm\ ^\circ C^{-1}$, RC_{DIC} = 1.047 $\mu atm\ \mu mol\ kg^{-1}$, and RC_{ALK} = -0.728 $\mu atm\ \mu mol\ kg^{-1}$. The scatter plot g shows that values of ALK keep increasing with respect to fCO_2 and the detected RC_{ALK} is negative. This occurs as we conduct multivariate linear regression, not univariate, and the final outcome of multivariate linear regression is influenced by the presence of all variables under consideration. Now, the scatter plots h-j highlight respective target-driver distribution in Equatorial Pacific, with RC_{SST} = 12.9 $\mu atm\ ^\circ C^{-1}$, RC_{DIC} = 1.349 $\mu atm\ \mu mol\ kg^{-1}$, and RC_{ALK} = -0.93 $\mu atm\ \mu mol\ kg^{-1}$. Finally, the plots k-m show the respective target-driver distribution in the Southern Ocean, with RC_SST = 9.22 $\mu atm\ ^\circ C^{-1}$, RC_DIC = 1.317 $\mu atm\ \mu mol\ kg^{-1}$, and RC_ALK = -0.845 $\mu atm\ \mu mol\ kg^{-1}$.

It should be noted that the uniformity in the quantity of data points across all boxes is not guaranteed. Firstly, the disparity arises from certain boxes encompassing land and sea areas. Additionally, the structure of the ocean model grid does not conform strictly to a consistent 0.25° regular grid. To address this disproportion, we employed statistical measures, specifically p-values, to test the significance of the regression coefficients. The subsequent analysis confines itself to those boxes with p-values less than 0.04. After this filtering, 99.64% of $2^\circ \times 2^\circ$ grid boxes were retained, ensuring a focus on statistically meaningful associations within the data. Following this step, our analysis yields a collection of diverse local/spatial linear relationships between fCO_2 and its associated drivers across the entire expanse of the ocean surface.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

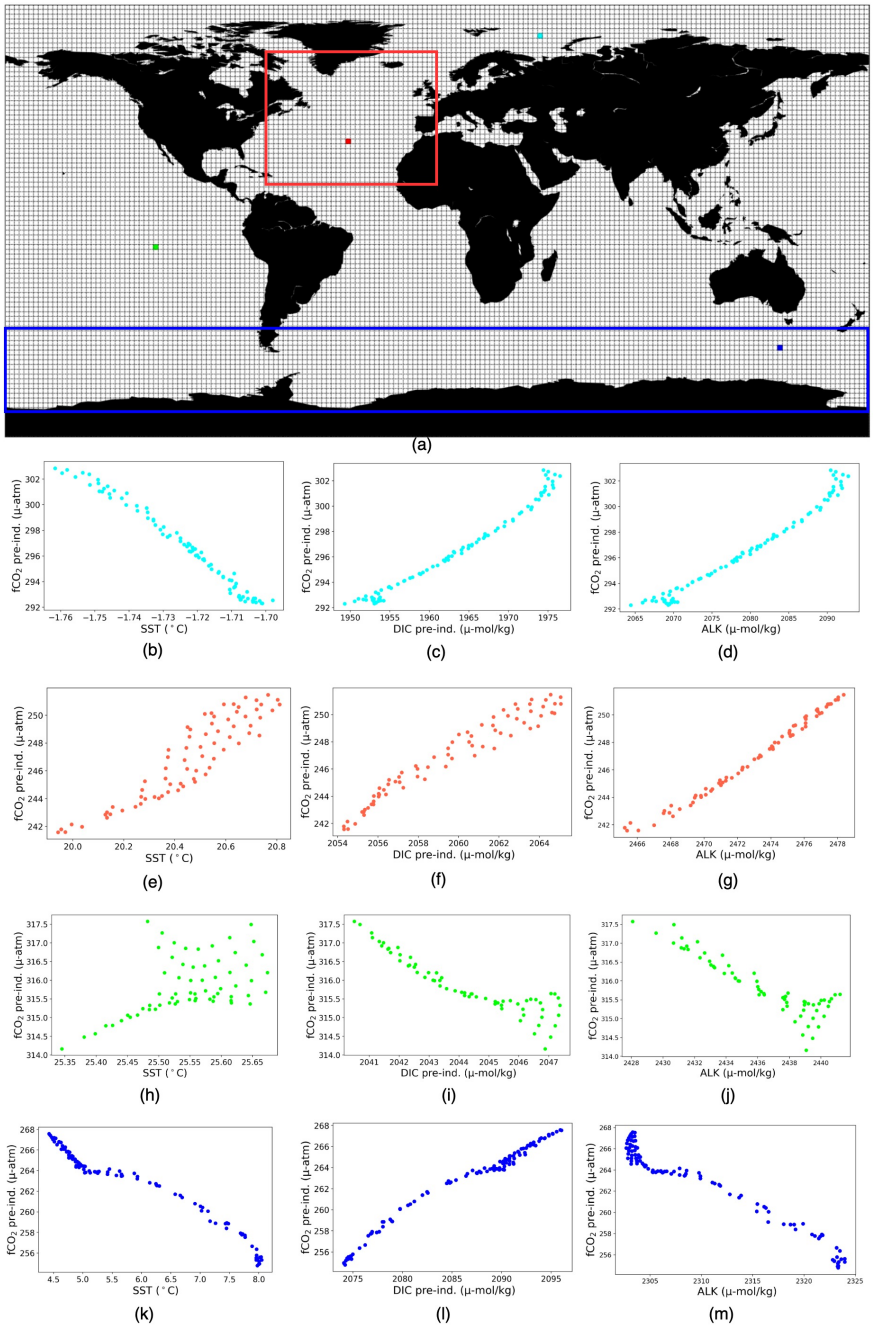

Figure A1. (a) $2^\circ \times 2^\circ$ boxes used for computing multivariate linear regressions, with the contours indicating the two basins - the North Atlantic and Southern Ocean - used for the detailed Sect. 3.2. b-m: scatter plots of fCO_2 against SST (first column), DIC (second column) and ALK (third column) over the four boxes indicated as colors in panel a.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

Exp. No.	$\Delta Dist$	ΔVar	No. of Clusters	BIC Score
1	20	0.1	9	95.54
2	20	0.5	6	76.20
3	20	0.8	5	73.14
4	20	1.0	5	73.14
5	**25**	**0.1**	**7**	**65.76**
6	25	0.5	4	46.42
7	25	0.8	3	43.36
8	25	1.0	3	43.36
9	28	0.1	7	65.76
10	28	0.5	4	46.42
11	28	0.8	3	43.36
12	28	1.0	3	43.36
13	30	0.1	7	65.76
14	30	0.5	4	46.42
15	30	0.8	3	43.36
16	30	1.0	3	43.36

Table B1. Cluster (Carbon biomes) selection for January 2009. BIC Scores calculated over different combinations of clustering parameters $\Delta Dist$ and ΔVar for hierarchical clustering on on normalized regression coefficients of SST, DIC, and ALK.

Appendix B: Cluster Selection based on BIC Scores

Normalized input for adaptive clustering: The input to the clustering algorithm is the normalized regression coefficients. Clustering algorithms are sensitive to the scale of variables. Variables with larger scales can dominate the grouping procedure and distort the results (Han et al., 2006). Normalizing variables ensures that all variables contribute equally to the agglomeration process, regardless of their original scales. Additionally, normalized data can lead to more meaningful clusters by focusing on relative differences between data points rather than absolute values. This gives us clusters that better capture the underlying structure of the data and are easier to interpret.

BIC Scores: BIC is a statistical measure utilized for model preference. BIC is based on the likelihood function and penalizes models for complexity to avoid overfitting. The BIC score for each clustering solution c (where $|c|$ is the total number of clusters found) is computed using Eq. (B1):

$$BIC = -2 \cdot \Sigma \log(likelihood(c)) + \log(n) \cdot k \tag{B1}$$

where $log(likelihood)$ is the natural logarithm of the likelihood of the data per cluster c, k is the number of parameters or degrees of freedom in the model, and n is the number of data points. The likelihood measures how well the clusters explain the

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

observed data. BIC comprises a penalty term for the number of parameters (in our case, three). This penalty term discourages overly complex models and helps prevent overfitting. The clustering solution with the lowest BIC score is chosen as the optimal solution. Lower BIC scores indicate a better trade-off between model fit and model complexity.

If we cut the dendrogram using a global parameter, we would have ended up with either too much fragmentation at subpolar and high latitudes (e.g., selecting a distance of 20) or too little structure (e.g., a cut at 40). Using the adaptive method delineated in 2.3, we achieve a reasonable amount of clusters reflective of their underlying data structure. We rely on BIC scores to select a pair of $\Delta Dist$ and ΔVar. BIC scores can be used in clustering to determine the optimal number of groups by comparing the fit of different cluster solutions. First, the output of the Distance-Variance cluster selection methodology for different combinations of $\Delta Dist$(ranging from 20 to 30) and ΔVar(ranging from 0 to 1.0) is carried out. For each selection, we receive a different number of clusters. Second, the likelihood of the input data (regression coefficients of the drivers) given the model is calculated for each clustering scenario.

We conducted 16 experiments for regression coefficients of January 2009 with different parameter choices as shown in table-B1. The clusters obtained through these pairs of $\Delta Dist$ and ΔVar have between 40 and 100 BIC scores. These experiments show that the BIC scores decrease as $\Delta Dist$ increases. With $\Delta Dist$=20.0 and ΔVar=0.1, we get 9 clusters with the highest BIC scores (this implies a clustering model with higher complexity and overfitting). With $\Delta Dist$=30.0 and ΔVar=1.0, we get the lowest BIC scores, but the number of carbon uptake provinces is 3, i.e., too low. We also see that for $\Delta Dist$=25.0, 28.0, 30.0, and ΔVar=0.1, we get a comparatively lower BIC score and 7 clusters. We opted for $\Delta Dist$=25.0 and ΔVar=0.1, as we can already distinguish distinct clusters with the lowest difference in the distance with the fixed difference in variance. We selected the clustering solution with the low BIC score to find the simplest model that best fits the data. BIC scores help identify the optimal number of clusters that effectively capture the underlying structure of the data without overfitting. These scores provide a principled approach to model selection in clustering, balancing goodness of fit and model complexity.

Appendix C: Visualization of shifting biomes between January and December in 2009

Fig. C1 highlights how the carbon regimes detected in January 2009 have spatially evolved in the next 11 months. The dendrogram map (a) has all seven clusters detected. The maps (b) to (l) have six clusters or carbon biomes, as predicted by the FNN model.

Appendix D: FNN Model Construction and Evaluation

There are different components of implementing FNNs - activation functions, loss functions, optimizers, and evaluation criteria. Activation functions are an integral part of neural network construction. We have chosen the Rectified Liner Unit, aka ReLU, for our input and hidden layers. The function returns the input value itself if it is positive. It returns 0 if the input value is negative. This has been shown in Eq. (D1). Depending on the outcome, the FNNs learn the importance of one variable (in our case, coefficients of SST, DIC, and ALK) over the other while predicting the biome labels. ReLU allows the network to learn

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

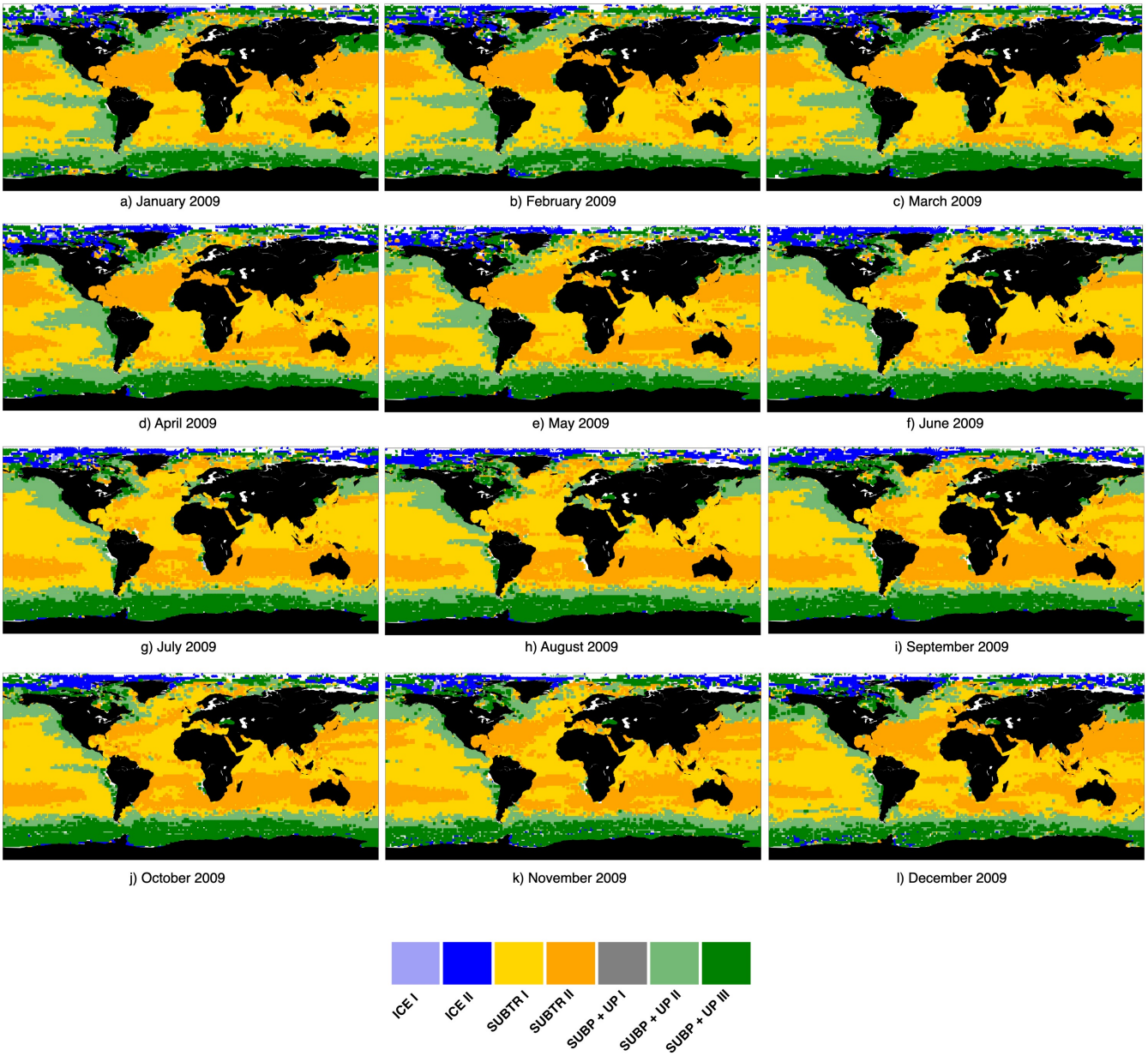

Figure C1. Carbon biomes over 12 months in 2009. Map (a) is the outcome of the adaptive hierarchical clustering algorithm, while maps from (b) to (l) are the output of the FNN model.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

more informative and discriminative features from the input data, producing enhanced performance. It provides non-linearities without the vanishing gradient issue compared with sigmoid or tanh activation functions. This enables faster learning and better optimization of the model parameters.

$$f(x) = max(0, x) \tag{D1}$$

Our task is tracking the carbon biomes, i.e., the goal is to classify a set of regression coefficients to a particular cluster label. Thus, for the final layer in our FNN model, we use the Softmax function. Softmax is well-suited for our multi-class classification task, as its objective is to provide one of several mutually exclusive cluster labels to input target-driver relationship data. Softmax also provides individual probabilities for each label, capturing uncertainties if necessary. These probabilities convey the FNN model's confidence or certainty about each cluster/biome label, permitting a probabilistic interpretation of the predictions. We pair Softmax (Eq. (D2)) with the cross-entropy loss function (Eq. (D3)), which measures the distinction between the predicted probability distribution and the actual distribution of cluster/biome labels. Mathematically, let $z = (z_1, z_2, ..., z_k)$ be the vector of scores produced by the neural networks for K number of cluster labels. The softmax function is applied to each element of z to produce the vector of predicted probabilities over K classes, $\hat{y} = (\hat{y}_1, \hat{y}_2, .., \hat{y}_K)$ as shown in Eq. (D2)):

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{D2}$$

where, $i = 1, 2, ..., K$

Let $y = (y_1, y_2, ..., y_K)$ be the one-hot encoded vector for the actual cluster label(i.e., a vector with a 1 in the position corresponding to the true biome and 0s elsewhere). The cross-entropy loss L between the predicted probabilities $\hat{y}$ and the true labels y is computed as Eq. (C3):

$$L(y, \hat{y}) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i) \tag{D3}$$

where, $L(y, \hat{y})$ is the cross-entropy loss, K represents the total number of class labels or cluster labels, $\hat{y}$ is the vector representing the predicted probability distribution over the K clusters, $\hat{y}_i$ is the predicted probability of class i, and y_i is the true probability of cluster label i (1 if the class is the true label, 0 otherwise).

FNNs also require an optimizer. An optimizer is an algorithm used to adjust the parameters (such as weights and biases) of the neural network during the training process. The objective of an optimizer is to minimize the loss function, which measures the difference between the predicted cluster label of the neural network and the actual cluster label. These optimizers help the neural network learn the fundamental patterns in the data and improve its performance by iteratively adjusting the parameters based on the gradient of the loss function. We employ ADAM (Adaptive Moment Estimation) Kingma and Ba (2017) in our cluster prediction/tracking task. It is an adaptive learning rate optimization algorithm that combines the advantages of two

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

other popular optimizers: AdaGrad Duchi et al. (2011) and RMSProp Ruder (2016). ADAM speeds up the training process and improves generalization performance(i.e., lowered overfitting). The evaluation metrics used in our FFN model - accuracy (in percentage), precision, and recall are defined as Eq. (D4), (D5), and (D6), respectively.

$$Accuracy = \frac{\text{No. of correctly predicted instances}}{\text{Total no. of instances in the dataset}} \times 100 \tag{D4}$$

$$Precision = \frac{\text{True positives}}{\text{True Positives} + \text{False Positives}} \tag{D5}$$

$$Recall = \frac{\text{True positives}}{\text{True Positives} + \text{False Negatives}} \tag{D6}$$

In order to optimize and select a robust feed-forward neural network (FNN), our goal was to recognize the most constructive amalgam of hyperparameters that would maximize the predictive accuracy, precision, and recall of the deep learning model. Hence, we experimented with various hyperparameter configurations, including diverse batch sizes, optimizers, learning rates, and numbers of training epochs. We evaluated batch sizes of 16, 32, and 64 with optimizers such as Adam and Stochastic gradient descent (SGD) with learning rates of 0.01 and 0.001. Additionally, we extended the number of training epochs from 50 to 200 to reckon the influence of training duration on model performance. In totality, we ran 240 experiments, and the list of different combinations of hyperparameters can be found on our Github link.

Through this tuning process, we observed that the ADAM optimizer attained the highest test accuracy of 95.9% over 50 epochs with a batch size of 32 and a learning rate set to 0.01. We employed early stopping to prevent overfitting, leading to FNN being trained for 30 epochs. Fig. D1 highlights how well the neural network was trained, and the plot Fig. D1 a emphasizes the change in cross-entropy loss during the process of training and validation. The plots Fig. D1 b-e show how accuracy, precision, and recall, respectively, evolved during training and validation over the 30 epochs. The confusion matrix Fig. D1 e highlights the average test accuracy to be at 94.83%, precision at 95.39%, and recall at 94.5% over seven different clusters.

D1 Limitations behind tracking SUBP+UP I

Our FNN model incorporated RC_{SST}, RC_{DIC}, and RC_{ALK} as inputs, with the target being the aforementioned carbon regimes. The distribution of labelled regimes across $2^\circ \times 2^\circ$ grid boxes is as follows: ICE I (1.03%), ICE II (4.39%), SUBTR I (37.40%), SUBTR II (23.60%), SUBP+UP I (0.55%), SUBP + UP II (18.97%), and SUBP + UP III (14.06%). This indicates an inconsistent distribution of inputs among different labels in our FNN. The trained FNNs were subsequently employed to predict the

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

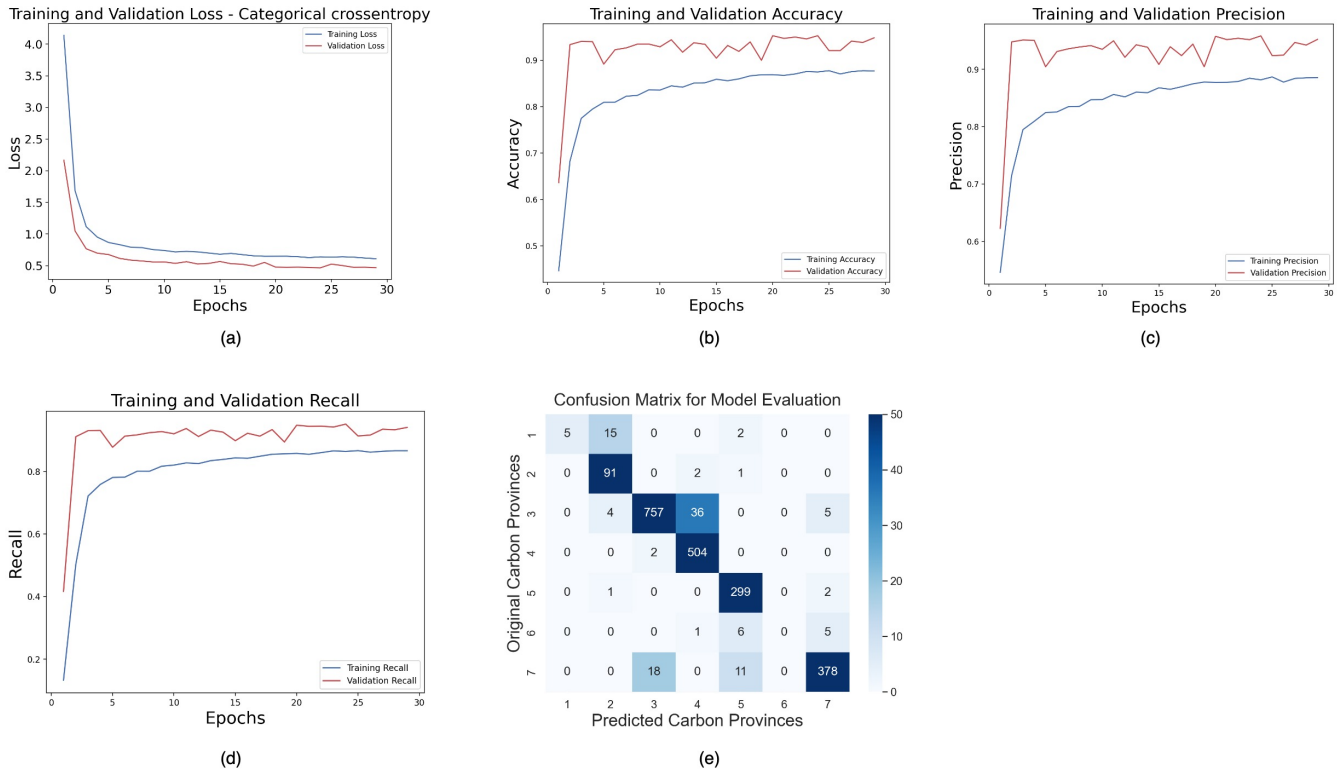

Figure D1. FNN model evaluation: During the training process, we used cross-entropy loss (a), accuracy (b), precision (c), and recall (d) to evaluate the quality of the model. The heatmap (e) is the confusion matrix plotted on a test set from January 2009 regression coefficients. In the confusion matrix, the numbers 1 to 7 represent the carbon biomes - 1: ICE I, 2: ICE II, 3: SUBTR I, 4: SUBTR II, 5: SUBP + UP III, 6: SUBP + UP I, and 7: SUBP + UP II.

seven regimes for 12 months over a span of 61 years (1958-2018). Notably, the prediction/tracking process successfully predicted six regimes, excluding SUBP+UP I. Such behavior in FNNs can arise from training on an imbalanced dataset, leading to a bias toward majority biome labels. In Fig. 3 a, the RCs of SUBP+UP I exhibit closer proximity to SUBP+UP II, with their combination occurring at a very low Euclidean Distance, suggesting minimal variation in the underlying distribution of both biomes. Furthermore, the combination of SUBP+UP I and SUBP+UP II lies closer to SUBP+UP III. As a result, the FNN model struggled to discern the structure of SUBP+UP I due to the scarcity of data points, leading it to associate the underlying RCs more strongly with SUBP+UP II and SUBP+UP III, which exhibit closer proximity in terms of RC distributions. The confusion matrix (Figure D1 e) illustrates the test predictions from January 2009, revealing that biome SUBP+UP I was predominantly predicted to be SUBP+UP II or SUBP+UP III.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

Author contributions. SM: conceptualization, analysis, investigation, methodology, visualization, writing; LP: ocean model integration, conceptualization, investigation, writing, editing, reviewing, funding acquisition; DK: conceptualization, investigation, writing, editing, reviewing; PK: editing, reviewing, funding acquisition

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. S. Mohanty was supported by the Helmholtz School for Marine Data Science, Helmholtz Association (grant HIDSS-0005). The ocean model simulations were integrated at the North German Supercomputing Alliance (HLRN, now NHR). L. Patara acknowledges funding from the German Research Foundation under grant PA 3075/2-1. S. Mohanty acknowledges using grammar and spell-checking AI tools.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

References

Agarap, A. F.: Deep learning using rectified linear units (relu), arXiv preprint arXiv:1803.08375, 2018.

Barnier, B., Brodeau, L., Le Sommer, J., Molines, J.-M., Penduff, T., Theetten, S., Treguier, A.-M., Madec, G., Biastoch, A., Böning, C. W., et al.: Eddy-permitting ocean circulation hindcasts of past decades, Clivar Exchanges, 42, 8–10, 2007.

Behncke, J., Landschützer, P., and Tanhua, T.: A detectable change in the air-sea CO 2 flux estimate from sailboat measurements, Scientific Reports, 14, 3345, 2024.

Bulgin, C. E., Merchant, C. J., and Ferreira, D.: Tendencies, variability and persistence of sea surface temperature anomalies, Scientific reports, 10, 7986, 2020.

Chien, C.-T., Durgadoo, J. V., et al.: FOCI-MOPS v1–Integration of Marine Biogeochemistry within the Flexible Ocean and Climate Infrastructure version 1 (FOCI 1) Earth system model, Geoscientific Model Development Discussions, 2022.

Couespel, D., Tjiputra, J., Johannsen, K., Vaittinada Ayar, P., and Jensen, B.: Machine learning reveals regime shifts in future ocean carbon dioxide fluxes inter-annual variability, Communications Earth & Environment, 5, 99, 2024.

DeVries, T., Yamamoto, K., Wanninkhof, R., Gruber, N., Hauck, J., Müller, J. D., Bopp, L., Carroll, D., Carter, B., Chau, T.-T.-T., et al.: Magnitude, trends, and variability of the global ocean carbon sink from 1985 to 2018, Global Biogeochemical Cycles, 37, e2023GB007 780, 2023.

Duchi, J., Hazan, E., and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization., Journal of machine learning research, 12, 2011.

Fay, A. R. and McKinley, G. A.: Global open-ocean biomes: mean and temporal variability, EARTH SYSTEM SCIENCE DATA, 6, 273–284, 2014.

Feely, R., Takahashi, T., Wanninkhof, R., McPhaden, M., Cosca, C., Sutherland, S., and Carr, M.-E.: Decadal variability of the air-sea CO2 fluxes in the equatorial Pacific Ocean, Journal of Geophysical Research: Oceans, 111, 2006.

Follows, M. J., Ito, T., and Dutkiewicz, S.: On the solution of the carbonate chemistry system in ocean biogeochemistry models, Ocean Modelling, 12, 290–301, 2006.

Friedlingstein, P., O'sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., Le Quéré, C., Luijkx, I. T., Olsen, A., Peters, G. P., et al.: Global carbon budget 2022, Earth System Science Data Discussions, 2022, 1–159, 2022.

Garcia, H., Weathers, K., Paver, C., Smolyar, I., Boyer, T., Locarnini, M., Zweng, M., Mishonov, A., Baranova, O., Seidov, D., et al.: World Ocean Atlas 2018, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Dissolved Oxygen Saturation., 2019.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep learning, MIT press, 2016.

Gruber, N., Bakker, D. C., DeVries, T., Gregor, L., Hauck, J., Landschützer, P., McKinley, G. A., and Müller, J. D.: Trends and variability in the ocean carbon sink, Nature Reviews Earth & Environment, 4, 119–134, 2023.

Han, J., Kamber, M., and Mining, D.: Concepts and techniques, Morgan Kaufmann, 340, 2006.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 9, 1735–1780, 1997.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., and Saynisch-Wagner, J.: Towards neural Earth system modelling by integrating artificial intelligence in Earth system science, Nature Machine Intelligence, 3, 667–674, 2021.

Jones, D. and Ito, T.: Gaussian mixture modeling describes the geography of the surface ocean carbon budget., 2019.

Key, R. M., Olsen, A., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., et al.: Global ocean data analysis project, version 2 (GLODAPv2), Ornl/Cdiac-162, Ndp-093, 2015.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, 2017.

Krasting, J. P., De Palma, M., Sonnewald, M., Dunne, J. P., and John, J. G.: Regional sensitivity patterns of Arctic Ocean acidification revealed with machine learning, Communications Earth & Environment, 3, 91, 2022.

Kriest, I. and Oschlies, A.: MOPS-1.0: towards a model for the regulation of the global oceanic nitrogen budget by marine biogeochemical processes, Geoscientific Model Development, 8, 2929–2957, 2015.

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

Landschützer, P., Gruber, N., and Bakker, D. C.: Decadal variations and trends of the global ocean carbon sink, Global Biogeochemical Cycles, 30, 1396–1417, 2016.

Lauderdale, J. M., Dutkiewicz, S., Williams, R. G., and Follows, M. J.: Quantifying the drivers of ocean-atmosphere CO2 fluxes, Global Biogeochemical Cycles, 30, 983–999, 2016.

Lauvset, S. K., Key, R. M., Olsen, A., Van Heuven, S., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., et al.: A new global interior ocean mapped climatology: The 1× 1 GLODAP version 2, Earth System Science Data, 8, 325–340, 2016.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, nature, 521, 436–444, 2015.

Levitus, S., Boyer, T. P., Conkright, M. E., O'Brien, T., Antonov, J., Stephens, C., Stathoplos, L., Johnson, D., and Gelfeld, R.: World ocean database 1998. Volume 1, Introduction, 1998.

Lin, F., Bai, B., Bai, K., Ren, Y., Zhao, P., and Xu, Z.: Contrastive Multi-view Hyperbolic Hierarchical Clustering, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, edited by Raedt, L. D., pp. 3250–3256, International Joint Conferences on Artificial Intelligence Organization, 2022.

Longhurst, A.: Seasonal cycles of pelagic production and consumption, Progress in oceanography, 36, 77–167, 1995.

Lovenduski, N. S., Gruber, N., Doney, S. C., and Lima, I. D.: Enhanced CO2 outgassing in the Southern Ocean from a positive phase of the Southern Annular Mode, Global Biogeochemical Cycles, 21, 2007.

Madec, G.: NEMO ocean engine, Note du Pôle modélisation, Inst, Pierre-Simon Laplace, p. 406, 2016.

Meinshausen, M., Vogel, E., Nauels, A., Lorbacher, K., Meinshausen, N., Etheridge, D. M., Fraser, P. J., Montzka, S. A., Rayner, P. J., Trudinger, C. M., et al.: Historical greenhouse gas concentrations for climate modelling (CMIP6), Geoscientific Model Development, 10, 2057–2116, 2017.

Mikaloff Fletcher, S., Gruber, N., Jacobson, A. R., Gloor, M., Doney, S., Dutkiewicz, S., Gerber, M., Follows, M., Joos, F., Lindsay, K., et al.: Inverse estimates of the oceanic sources and sinks of natural CO2 and the implied oceanic carbon transport, Global Biogeochemical Cycles, 21, 2007.

Mohanty, S., Kazempour, D., Patara, L., and Kröger, P.: Detection and Tracking of Dynamic Ocean Carbon Uptake Regimes Built Upon Spatial Target-Driver Relationships via Adaptive Hierarchical Clustering, in: 2023 IEEE 19th International Conference on e-Science (e-Science), pp. 1–10, IEEE, 2023a.

Mohanty, S., Kazempour, D., Patara, L., and Kröger, P.: Interactive Detection and Visualization of Ocean Carbon Regimes, in: Proceedings of the 18th International Symposium on Spatial and Temporal Data, pp. 171–174, 2023b.

Müllner, D.: Modern hierarchical, agglomerative clustering algorithms, 2011a.

Müllner, D.: Modern hierarchical, agglomerative clustering algorithms, 2011b.

Oliver, T. H., Heard, M. S., Isaac, N. J., Roy, D. B., Procter, D., Eigenbrod, F., Freckleton, R., Hector, A., Orme, C. D. L., Petchey, O. L., et al.: Biodiversity and resilience of ecosystem functions, Trends in ecology & evolution, 30, 673–684, 2015.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

Orr, J. C., Najjar, R. G., Aumont, O., Bopp, L., Bullister, J. L., Danabasoglu, G., Doney, S. C., Dunne, J. P., Dutay, J.-C., Graven, H., et al.: Biogeochemical protocols and diagnostics for the CMIP6 Ocean Model Intercomparison Project (OMIP), Geoscientific Model Development, 10, 2169–2199, 2017.

Patara, L., Visbeck, M., Masina, S., Krahmann, G., and Vichi, M.: Marine biogeochemical responses to the North Atlantic Oscillation in a coupled climate model, Journal of Geophysical Research: Oceans, 116, 2011.

Pérez, F. F., Mercier, H., Vázquez-Rodríguez, M., Lherminier, P., Velo, A., Pardo, P. C., Rosón, G., and Ríos, A. F.: Atlantic Ocean CO2 uptake reduced by weakening of the meridional overturning circulation, Nature Geoscience, 6, 146–152, 2013.

Pfeil, B., Olsen, A., Bakker, D. C., Hankin, S., Koyuk, H., Kozyr, A., Malczyk, J., Manke, A., Metzl, N., Sabine, C. L., et al.: A uniform, quality controlled Surface Ocean CO 2 Atlas (SOCAT), Earth System Science Data, 5, 125–143, 2013.

Prend, C. J., Hunt, J. M., Mazloff, M. R., Gille, S. T., and Talley, L. D.: Controls on the boundary between thermally and non-thermally driven pCO2 regimes in the South Pacific, Geophysical Research Letters, 49, e2021GL095 797, 2022.

Reygondeau, G., Cheung, W. W., et al.: Climate change-induced emergence of novel biogeochemical provinces, Frontiers in Marine Science, 7, 657, 2020.

Ruder, S.: An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747, 2016.

Sallée, J.-B., Pellichero, V., Akhoudas, C., Pauthenet, E., Vignes, L., Schmidtko, S., Garabato, A. N., Sutherland, P., and Kuusela, M.: Summertime increases in upper-ocean stratification and mixed-layer depth, Nature, 591, 592–598, 2021.

Sarmiento, J. L. and Gruber, N. P.: Ocean biogeochemical dynamics, in: Ocean Biogeochemical Dynamics, Princeton university press, 2006.

Schmittner, A., Oschlies, A., Matthews, H. D., and Galbraith, E. D.: Future changes in climate, ocean circulation, ecosystems, and biogeochemical cycling simulated for a business-as-usual CO2 emission scenario until year 4000 AD, Global biogeochemical cycles, 22, 2008.

Sonnewald, M., Dutkiewicz, S., Hill, C., and Forget, G.: Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces, Science advances, 6, eaay4740, 2020.

Sutton, A. J., Williams, N. L., and Tilbrook, B.: Constraining Southern Ocean CO2 flux uncertainty using uncrewed surface vehicle observations, Geophysical Research Letters, 48, e2020GL091 748, 2021.

Swart, N. C., Fyfe, J. C., Gillett, N., and Marshall, G. J.: Comparing trends in the southern annular mode and surface westerly jet, Journal of Climate, 28, 8840–8859, 2015.

Takahashi, T., Olafsson, J., Goddard, J. G., Chipman, D. W., and Sutherland, S.: Seasonal variation of CO2 and nutrients in the high-latitude surface oceans: A comparative study, Global Biogeochemical Cycles, 7, 843–878, 1993.

Takahashi, T., Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R., Feely, R. A., Sabine, C., et al.: Global sea–air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects, Deep Sea Research Part II: Topical Studies in Oceanography, 49, 1601–1622, 2002.

Takahashi, T., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G., Chavez, F., Sabine, C., et al.: Climatological mean and decadal change in surface ocean pCO2, and net sea–air CO2 flux over the global oceans, Deep Sea Research Part II: Topical Studies in Oceanography, 56, 554–577, 2009.

Tsujino, H., Urakawa, S., Nakano, H., Small, R. J., Kim, W. M., Yeager, S. G., Danabasoglu, G., Suzuki, T., Bamber, J. L., Bentsen, M., et al.: JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do), Ocean Modelling, 130, 79–139, 2018.

Vichi, M., Allen, J. I., Masina, S., and Hardman-Mountford, N. J.: The emergence of ocean biogeochemical provinces: A quantitative assessment and a diagnostic for model evaluation, Global Biogeochemical Cycles, 25, 2011.

https://doi.org/10.5194/egusphere-2024-1369
Preprint. Discussion started: 23 May 2024
© Author(s) 2024. CC BY 4.0 License.

Wanninkhof, R.: Relationship between wind speed and gas exchange over the ocean revisited, Limnology and Oceanography: Methods, 12, 351–362, 2014.

Williams, R. G. and Follows, M. J.: Ocean dynamics and the carbon cycle: Principles and mechanisms, Cambridge University Press, 2011.