JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2026JH001326

Estimating Carbon Pools in the European Shelf Sea Environment: Replacing Reanalysis by Model-Informed Machine Learning?

Jozef Skákala[1,2]

[1]Plymouth Marine Laboratory, Plymouth, UK, [2]National Centre for Earth Observation, Plymouth, UK

Key Points:

- We present a marine-model-consistent, machine-learning-based reconstruction of carbon pools from observations
- By feeding reanalysis inputs into the machine learning (ML) model, its predictions reproduce several of the carbon pools from the same reanalysis
- We propose using ML both as an efficient alternative to reanalysis and as a tool to simulate what-if scenarios

Correspondence to:

J. Skákala,
jos@pml.ac.uk

Citation:

Skákala, J. (2026). Estimating carbon pools in the European shelf sea environment: Replacing reanalysis by model-informed machine learning? *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2026JH001326. https://doi.org/10.1029/2026JH001326

Received 6 MAR 2026
Accepted 14 MAY 2026

Author Contributions:

Conceptualization: Jozef Skákala
Formal analysis: Jozef Skákala
Funding acquisition: Jozef Skákala
Investigation: Jozef Skákala
Methodology: Jozef Skákala
Resources: Jozef Skákala
Software: Jozef Skákala
Validation: Jozef Skákala
Visualization: Jozef Skákala
Writing – original draft: Jozef Skákala
Writing – review & editing: Jozef Skákala

Abstract Shelf seas are important for the economy and the carbon cycle, but shelf sea observations for carbon pools are often sparse or highly uncertain. An alternative can be provided by carbon reanalyses (whether assimilating proxy variables, such as chlorophyll-*a*, or directly carbon), but these are often expensive to run. We propose to use a computationally cheap ensemble of neural networks (i.e., deep ensemble) to learn the relationship between the directly observable (atmospheric, riverine, and ocean) variables and marine carbon pools from a coupled physics-biogeochemistry model. The deep ensemble was trained on a North-West European Shelf (NWES) physical-biogeochemistry model free run simulation. After training, the deep ensemble was run using inputs from the NWES reanalysis instead of the free run, demonstrating that it can efficiently predict several NWES carbon pools (e.g., detritus, zooplankton, and heterotrophic bacteria) in much better agreement with the reanalysis than the free run, while also providing uncertainty information. We further show that the deep ensemble performs similarly well when it is driven directly by the observations assimilated into the reanalysis, with the limitation that carbon pools can then be predicted only at the observed locations and times. We focus on explainability of the results and demonstrate potential use of the deep ensembles for future climate what-if scenarios. We suggest that model-informed machine learning presents a viable alternative to expensive reanalyses and could complement observations, wherever they are missing and/or highly uncertain.

Plain Language Summary The ocean absorbs approximately 30% of carbon emitted into the atmosphere. Monitoring marine carbon pools is essential to understand how carbon cycles within the ocean, including cross-shelf exchange between shelf seas and the open ocean, export to the deep sea and burial in sediments. Shelf seas, owing to land-sea exchange, the solubility pump, and high biological productivity, play a disproportionate role in carbon uptake and cycling. However, observations of different carbon pools in shelf seas are often sparse or uncertain, and marine ecosystem models can also exhibit substantial biases and uncertainties. Models and observations can be combined through data assimilation to produce reanalyses of shelf sea carbon pools, but this approach is computationally expensive. Here, we propose deriving a range of carbon pools (detritus, dissolved organic carbon, zooplankton, heterotrophic bacteria, and dissolved inorganic carbon) from more directly observable variables using machine learning (ML) to reproduce their relationships from a coupled physics–biogeochemistry model. We demonstrate that such ML models provide a cost-effective alternative to reanalysis where carbon observations are sparse or highly uncertain. The low computational cost of using ML for inference offers additional advantages, including enabling straightforward uncertainty quantification and facilitating simulations of a wide range of future climate what-if scenarios.

1. Introduction

About 30% of carbon emitted into the atmosphere ends up absorbed by the ocean, where it circulates in a multitude of organic and inorganic forms (Friedlingstein et al. (2024)). The inorganic carbon is assimilated by autotrophs during photosynthesis, which is followed by marine food web interactions distributing it between many other pools, such as higher trophic level species (e.g., zooplankton, heterotrophic bacteria, fish), non-living organic forms (e.g., detrital, dissolved organic carbon (DOC)), or the dissolved and particulate inorganic carbon (e.g., Emerson and Hedges (2008)). Part of the marine carbon gets eventually deposited to the seafloor and buried, which helps to mitigate the anthropogenically driven climate change (Volk and Hoffert (1985)). Understanding the ocean carbon cycle is therefore essential for better understanding Earth's climate response to atmospheric carbon emissions, both in past and future projections. Although the coastal oceans and shelf seas cover only $10-15\%$ of the global ocean, their impact on the global carbon cycle is disproportionately large (e.g., Roobaert

et al. (2019, 2024); Cao et al. (2020); Chau et al. (2022); Dai et al. (2022)). The North-West European Shelf (NWES), comprising a range of seas including the North Sea, Celtic Sea, Irish Sea, and the English Channel, is of major importance for the economy and the carbon cycle, as it is a highly biologically productive area with a significant impact on carbon sequestration and transport (Borges et al. (2006); Jahnke (2010); Legge et al. (2020)).

A major source of information for the NWES are ocean surface observations derived from the satellite measurements. The biogeochemistry variable most typically derived from satellite Ocean Color (OC) is the surface chlorophyll-*a* pigment concentrations (e.g., Groom et al. (2019); Sathyendranath et al. (2019)), which contain only indirect information about the NWES carbon pools. A variety of methods to estimate carbon from satellites more directly have been proposed for a range of pools (Brewin et al. (2021)), including total particulate organic carbon (POC; e.g., Evers-King et al. (2017); Le et al. (2018)), total phytoplankton carbon (e.g., Roy et al. (2017); Sathyendranath et al. (2020)), dissolved organic carbon (DOC; e.g., Matsuoka et al. (2017); Laine et al. (2024)), as well as indirect approaches to estimate zooplankton carbon (e.g., Strömberg et al. (2009); Behrenfeld et al. (2019)) and carbon associated with some bacterial species (Grimes et al. (2014); Racault et al. (2019)). However these satellite algorithms have been most commonly developed for the global ocean, representing mostly open ocean conditions. Some shelf sea and coastal products exist for specific pools (e.g., DOC, Mannino et al. (2008); Matsuoka et al. (2017)), but have been developed for areas far from the NWES. Thus, apart of the relatively high uncertainty associated with many of those products, they might not be particularly suitable for the NWES. Beyond these satellite products the only available observations are in situ measured data. These cover quite well specific variables related to dissolved inorganic carbon (DIC), such as CO_2 fugacity and partial pressure (Bakker et al. (2014)), which are typically governed by well-established thermodynamic relationships and can be either directly measured or robustly derived from direct observations. However organic carbon pools are typically operationally defined, meaning their quantification depends on methodological choices (e.g., filtration thresholds, oxidation techniques, or analytical protocols). This leads to greater heterogeneity and sparsity across data sets, and those data become too limited to provide us with a more detailed understanding of organic NWES carbon pools (for more general overview of in situ observational capacity in marine biogeochemistry see e.g. Telszewski et al. (2018)).

A more comprehensive representation of the system is provided by marine biogeochemistry models (Wakelin et al. (2012)). However, in the NWES these models can exhibit substantial biases and uncertainties. For example, the European Regional Seas Ecosystem Model (ERSEM; Baretta et al. (1995); Butenschön et al. (2016)), which is used operationally in the region, shows pronounced seasonal errors in phytoplankton phenology. Specifically, it simulates overly intense and delayed spring blooms, alongside near-vanishing phytoplankton concentrations during winter (see Skákala et al. (2022, 2024) for discussion). Because phytoplankton form the base of the marine food web, such biases propagate through the ecosystem and influence carbon cycling (Skákala et al. (2022, 2024)). An intermediate approach between sparse observations and imperfect free-running models (unconstrained by the observations through data assimilation) is provided by the NWES Copernicus reanalysis (Kay et al. (2021, 2019)). This framework constrains the model primarily with satellite observations while delivering a range of outputs for carbon pools. In the NWES reanalysis, ERSEM biogeochemistry is constrained through the assimilation of satellite phytoplankton functional type (PFT) chlorophyll-*a* (Brewin et al. (2017); Skákala et al. (2018)). This substantially reduces key biases in phytoplankton phenology (Skákala et al. (2018); Kay et al. (2021)), making the reanalysis outputs generally more reliable than those from the free-running model. Given the scarcity of robust and comprehensive observations, we treat the reanalysis carbon pool estimates in this study as the data set closest to "ground truth." Nevertheless, generating such reanalyses is computationally expensive, particularly for complex biogeochemical models. To maintain computational feasibility, several simplifications are implemented. Most notably, only biogeochemical variables directly related to the assimilated PFT chlorophyll-*a* product, i.e., phytoplankton biomass, are directly constrained. Other variables (e.g., non-phytoplankton carbon pools) are impacted only through the model dynamical adjustment (Skákala et al. (2018)).

In this paper, we propose an alternative approach. In the spirit of satellite retrieval algorithms, we use a machine learning (ML) model to learn the relationship between directly observable satellite variables (such as sea surface temperature and ocean-color-derived PFT chlorophyll-*a*) and unobserved carbon pools. (For examples of ML models estimating carbon pools see Sauzède et al. (2020); Zemskova et al. (2022); C. Li et al. (2024); Laine et al. (2024); Zhang et al. (2025).) However, in our case, the relationship is learned directly from the underlying NWES coupled physics–biogeochemistry model. In the coupled model such relationship emerges from a plethora

of simulated processes, ranging from hydrodynamics to complex ecosystem interactions. Using the coupled model to inform the ML has several advantages: (a) there is no shortage of training data, that is, the map from the observable variables to the carbon pools can be learned from the existing free run simulations, providing gap-free and abundant outputs, (b) the coupled model is specifically calibrated for the NWES domain, and (c) any carbon pool outputted by the model can be in theory predicted, including vertical distributions.

The disadvantage of this approach is the assumption that the model-simulated relationship between the established satellite-observable variables (together with a range of model forcing data) and the derived carbon pools is sufficiently realistic. This means that any biases in the carbon pools modeled in the free run will be primarily due to model misrepresenting the dynamics of the observable variables (e.g., showing large seasonal biases in chlorophyll-*a*), rather than in the modeled relationship between those variables and the carbon pools. This is a strong assumption, however, it can be argued that similarly strong assumptions are effectively used in the NWES reanalysis itself, when many of the carbon pools are left to dynamically respond within the same coupled model to the assimilation updates to PFT biomass components and physical variables. On the other hand, data assimilation provides information about the uncertainty in the estimated carbon pools. This uncertainty comes from errors in the observational data and from different sources of uncertainty in the coupled physics–biogeochemistry model. These include uncertainty in model forcing, initial and boundary conditions, model structure, and model parameter values. We assume that the last two sources (model structure and parameter values) are major contributors to uncertainty in how the model maps observable variables to carbon pools. Our ML model learned only one version of this map (e.g., related to a specific set of marine model parameter values), and it cannot directly represent this uncertainty. We therefore conducted additional ensemble 1D simulations as an initial step toward quantifying the magnitude of this uncertainty, as described in the Results section.

In this paper, we demonstrate that a reasonably simple feed-forward neural network (NN) can successfully learn from the operational physics–biogeochemistry model for the NWES how to derive a range of surface carbon pools from several observable variables, forcing data, and structural variables such as latitude and longitude. The carbon pools are derived at a relatively coarser spatial (35 km) and temporal (10-day) resolution than that of the operational model outputs (7 km and daily), although this coarser resolution is entirely sufficient for regional climate studies. When driven by inputs from the reanalysis rather than from the free-running model on which it was trained, the NN reproduces the reanalysis estimates of several surface ocean carbon pools reasonably well (i.e., substantially better than the free run). The same holds when the NN uses as inputs the observations directly assimilated into the reanalysis, as intended in its application. We provide insights into the explainability of these results and discuss interesting applications, such as running future climate what-if scenarios using lightweight ML models.

2. Methodology

2.1. The Physics-Biogeochemistry Coupled Model

The model used in NWES Copernicus reanalysis (Kay et al. (2019, 2021)) that we are referring to in this study is the physical model Nucleus for European Modeling of the Ocean (NEMO, Madec et al. (2015)), coupled to ERSEM (Baretta et al. (1995); Butenschön et al. (2016)) through Framework for Aquatic Biogeochemical Models (FABM, Bruggeman and Bolding (2014)). The NEMO ocean physics component is a finite difference, hydrostatic, primitive equation ocean general circulation model, here used with a 7 km (AMM7) NWES configuration using the terrain following $z* - \sigma$ coordinates with 51 vertical layers (O'Dea et al. (2017); Siddorn and Furner (2013)). ERSEM is a high complexity ecosystem model representing elemental cycles of carbon, nitrogen, phosphorus and silicon using variable stoichiometry. It represents four PFTs which are mostly size-based (diatoms, nanophytoplankton, microphytoplankton and picophytoplankton), zooplankton in three functional types (mesozooplankton, microzooplankton and heterotrophic nanoflaggelates) and as a decomposer it includes heterotrophic bacteria (Butenschön et al. (2016)). The non-living organic matter is represented in three detrital forms (large, medium-size, small) and three forms of dissolved organic matter (labile, semi-labile and semi-refractory) (Butenschön et al. (2016)). ERSEM uses variable stoichiometry to represent biomass in terms of carbon, nitrogen, and phosphorus, and for phytoplankton also chlorophyll-*a*, with silicon additionally included for diatoms. ERSEM includes carbonate system as per Artioli et al. (2012), representing DIC and total alkalinity as two independent state variables (from which one can derive pH and pCO_2 as diagnostic variables).

10.1029/2026JH001326

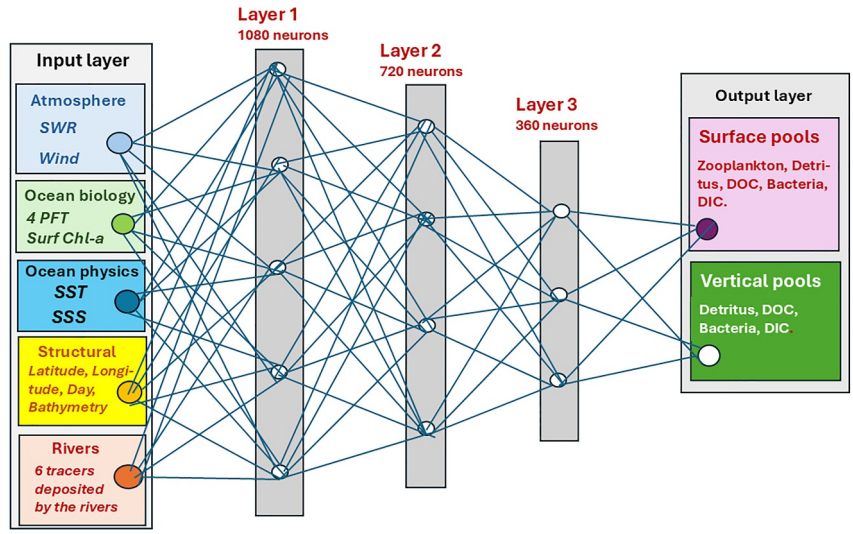

Figure 1. A schematic representation of the feedforward neural network model, its inputs and outputs. For simplicity the schematic representation merges the same types of inputs and outputs into one node on the diagram (i.e., in reality 18 inputs and 9 outputs were used rather than 5 input and 2 output nodes displayed in the schematic Figure). The abbreviations used are, SST: sea surface temperature, SSS: sea surface salinity, SWR: short-wave radiation, DOC: dissolved organic carbon, DIC: dissolved inorganic carbon.

2.2. The Machine Learning Model

We employed a fully connected feed-forward NN implemented using TensorFlow Keras, consisting of three hidden layers with a total of 2160 neurons, arranged in a decreasing configuration of 1,080, 720, and 360 neurons per layer (see Figure 1). The NN model used Adam optimizer, root mean squared cost function and random initialization of weights based on normal distribution. To reduce overfitting and optimize performance we have used dropout function with 30% of neurons randomly switched off at each learning step. This architecture was chosen after extensive testing of NN performance across range of different architectures (e.g., numbers of hidden layers) and hyperparameter values.

We have used a 15-member ensemble of the NN model realizations (i.e., deep ensemble) to boost the performance and introduce some estimate of (mostly epistemic) uncertainty. The ensemble members were naturally distinguished through the random model parameters, such as weight initialization, dropout function and other random procedures, such as splitting the training data into batches (batch size of 32 was used). The ensemble members used for training the same number of epochs. The coefficient of variation of the final training loss across ensemble members was 5.8%, indicating moderate optimization variability. Prediction was taken from the ensemble mean and uncertainty was evaluated as the ensemble standard deviation.

The deep ensemble used as input features (a) observable ocean variables, that is, sea surface PFT chlorophyll, sea surface temperature (SST) and sea surface salinity, (b) "structural" inputs, that is, latitude and longitude, bathymetry and the day of the year, (c) atmospheric inputs, that is, incoming short-wave radiation, surface wind speed, and (d) riverine discharge data variables such as freshwater runoff, nutrients, oxygen, DIC (see also schematic representation in Figure 1). The riverine inputs were considered non-zero in the 50 km neighborhood of the river mouths, as described in Banerjee and Skákala (2025). The deep ensemble predicted a number of carbon pools in the output layer, that is, for (a) total surface zooplankton, (b) total surface detritus, (c) total surface dissolved organic matter, (d) surface heterotrophic bacteria, (e) surface dissolved inorganic matter, and vertically averaged values for all these pools except for zooplankton. All input features and predicted outputs were standardized using z-score normalization $\hat{x}_i = \left(x_i - \mu_x\right)/\sigma_x$, where μ_x and σ_x denote the mean and standard deviation computed over the training data set. This standardization mitigates scale differences between variables and prevents features or outputs with larger numerical ranges from disproportionately influencing the optimization process.

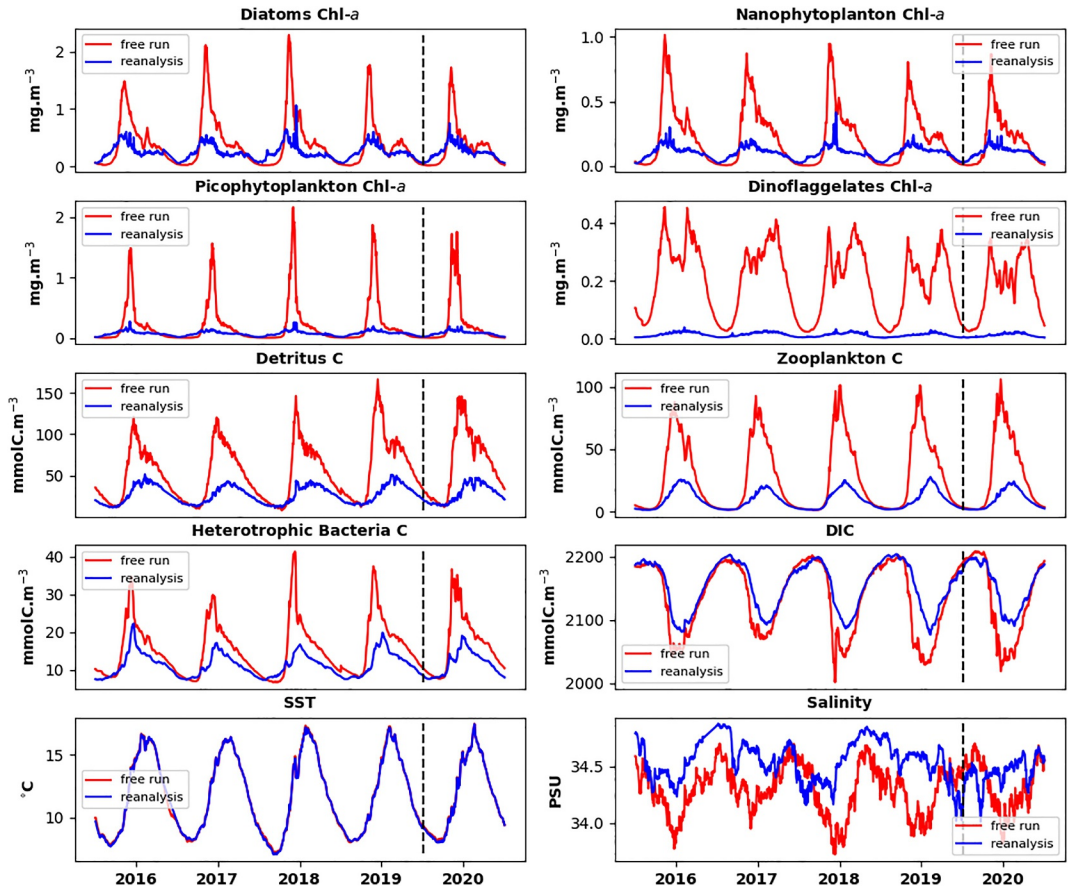

Figure 2. Time series (2016–2020) of domain-averaged values from the training and validation data set derived from the free run (red) and the test data from the reanalysis (blue). The panels compare the key biogeochemical and physics variables of interest: (i) surface phytoplankton functional type chlorophyll, directly corrected by data assimilation in the reanalysis (first two rows), (ii) those (predicted by the neural network) surface carbon pools which were outputted by both the free run and the reanalysis (rows three and four) and (iii) physics variables (SST and surface salinity), both assimilated in the reanalysis (bottom row). The vertical dashed lines show the separation between the training and validation data from the free run.

The training and validation (training-validation) data were: (a) for the ocean variables provided by the 5 years 2016–2020 NEMO-FABM-ERSEM free simulation (the time series of the domain averaged data are shown in Figure 2), (b) for the atmospheric variables by the corresponding 2016–2020 ERA-5 atmospheric product (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5), forcing the oceanic model simulation, and (b) for the river data by the riverine data -set originating from Lenhart et al. (2010), also used to force the oceanic model simulation. The 2016–2020 data set was split into the 2016–2019 period used as the training data set and 2020 years for validation. The training–validation split was chosen to separate the data sets temporally, thereby maximizing their independence and reducing the influence of temporal autocorrelation.

The NEMO-FABM-ERSEM configuration providing the simulations for the training-validation data has been described in Skákala et al. (2020). Similarly to the Copernicus reanalysis the model horizontal resolution was 7 km with 51 vertical grid layers using a terrain-following $z^* - \sigma$ coordinate system. The model was forced by atmospheric data from the ERA-5 reanalysis and used lateral boundary conditions from the GloSea5 Seasonal Forecasting System (MacLachlan et al. (2015)); and from a reanalysis produced by the Danish Meteorological Institute for the Copernicus Marine Service. The simulation used NEMO model version 3.6, a development of the CO5 configuration explained in detail by O'Dea et al. (2017). The light in the ERSEM model was forced by a spectral bio-optical model described in Skákala et al. (2020), and both NEMO and ERSEM were forced by an updated version of the river discharge data set from Lenhart et al. (2010). The simulation delivering the training-validation data set has been performed in Higgs et al. (2024), providing daily and 7 km resolution outputs for all the relevant carbon pools. The training-validation data have been produced by coarsening the simulation outputs

(through averaging) to a 10-day temporal scale and 35 km (5 × 5 model pixel) spatial scale. The deep ensemble inference is consequently applied only to such coarser spatial and temporal scales, which are however considered sufficient to monitor carbon pools on the NWES. These coarser scales were chosen to reduce the data redundancy/duplication due to their spatial and temporal correlations, improving the efficiency of the training and validation of the ML model. The overall number of data points in the coarsened training-validation data was over 500,000.

The performance of the relatively lightweight feed-forward NN (with three hidden layers and ∼2,200 neurons) could be improved by exploiting the spatial and temporal structure of the inputs, for example, by using more sophisticated architectures such as convolutional neural networks (CNNs; Z. Li et al. (2021)), graph neural networks (GNNs; Scarselli et al. (2008)), or long short-term memory networks (LSTMs; Hochreiter and Schmidhuber (1997); Yu et al. (2019)). However, to predict carbon pools directly from observations, these architectures would need to handle substantial time-varying gaps caused by cloud cover, atmospheric disturbances and low solar zenith angle, including seasonal gaps and the near-absence of data over large parts of the NWES in winter. Such seasonally biased data gaps—potentially correlated with the target ocean variables—may pose significant challenges for these architectures, especially if the training data set is reasonably limited (e.g., Zhou et al. (2020)). Another challenge is how to generalize models trained on gap-free free-run simulations (both spatially and temporally) so that they can make accurate predictions when using satellite data as input, which frequently contain substantial gaps. Finally, models that rely on spatial and temporal relationships (such as CNNs or GNNs) and are trained on the model free run may underperform when applied to reanalysis inputs because the spatial and temporal consistency of reanalysis variables is degraded by the intermittent availability of assimilated observations, leading to highly uneven impacts across the model domain. For these reasons, at this initial proof-of-concept stage, we chose to use a simpler perceptron NN that predicts carbon pools using only inputs from the same location as the target value. In future work, however, it will be desirable to explore more complex architectures and/or explore alternative very recently developed physics-informed approaches (Raissi et al. (2019)) to predict target data from incomplete observations, such as 4DVarNet (Fablet et al. (2021)).

2.3. The Multi-Decadal Reanalysis

The capability of the ML model to predict carbon pools using observational data as inputs was subsequently tested on a data set provided by Copernicus multi-decadal reanalysis (Kay et al. (2019, 2021)). The reanalysis is based on a very similar NEMO-FABM-ERSEM model configuration as the free run used for training-validation. The NEMO-FABM-ERSEM model was constrained in the reanalysis through daily assimilation of SST from European Space Agency (ESA) Climate Change Initiative (CCI) v1.1 product, in situ SST from International Comprehensive Ocean-Atmosphere Data Set, temperature and salinity profiles from EN4 data (Good et al. (2013)), and PFT (log-)chlorophyll from ESA CCI v3.1 data (Sathyendranath et al. (2019)). The data were assimilated using NEMOVAR (Mogensen et al. (2012)), based on 3DVar approach similar to the one described in King et al. (2018); Skákala et al. (2018). Within ERSEM the assimilation of PFT chlorophyll-*a* updates only the PFT biomass (all components chlorophyll-*a*, carbon, nitrogen, phosphorus, and for diatoms silicon), based on the forecast stoichiometry. The remaining ERSEM variables are not directly constrained by the assimilation and evolve only during the subsequent model simulation through dynamical adjustment. Figure 2 shows that the assimilated PFT chlorophyll-*a*, as well as the simulated surface carbon pools, differ substantially between the reanalysis and the free run that provided the training and validation data for the NN. This discrepancy is well documented in the literature (e.g., Skákala et al. (2018)). Consequently, the reanalysis data constitute a highly non-trivial test of NN performance.

The reanalysis carbon can be validated with independent data, wherever they are available. For the relatively abundant ICES pCO2 data set this was already done in Kay et al. (2021), demonstrating a good skill. We have done here an additional comparison of the reanalysis outputs with OC-CCI v4.2 satellite product for the total POC (Stramski et al. (2008); Evers-King et al. (2017)), comprising aggregate carbon of phytoplankton, zooplankton, detritus and bacteria pools. The results are shown in Figure A1 of the Appendix A. They indicate that the POC in the reanalysis has significant negative biases relative to the satellite product, but the way the POC values are distributed across the domain is to a degree similar between both products. Satellite-derived DOC from Laine et al. (2024) was also available, unfortunately DOC was not outputted by the reanalysis and could not be compared to the satellite observations. However issues (major positive biases in hundreds of percents of the observed values) with ERSEM DOC have been found (Clark et al. (2025)) and have been also confirmed here (not

shown). Although DOC pool is included here to demonstrate the ML capability to learn it from the model, it is recognized that as an end-product it would likely have at this stage only limited value.

2.4. Experiments and Validation Metrics

The model performance on the independent data produced by the reanalysis was evaluated by comparing the spatial distributions of 2016–2020 time-averaged reanalysis and predicted carbon pools, indicating the overall biases of the prediction. A separate metrics of Bias-Corrected Root-Mean Square Difference (BC-RMSD) was applied, defined as

$$\text{BC-RMSD} = \sqrt{\langle (X_1 - X_2 - (\langle X_1 \rangle - \langle X_2 \rangle))^2 \rangle}, \tag{1}$$

where X_1, X_2 are the two compared data sets. The BC-RMSD % improvement generated by data 1 (with BC-RMSD_1) relative to data 2 (with BC-RMSD_2) is defined as

$$\text{BC-RMSD}_{imp} = 100 \cdot \frac{\text{BC-RMSD}_1 - \text{BC-RMSD}_2}{\text{BC-RMSD}_2}. \tag{2}$$

To demonstrate the ultimate goal of this work—namely, predicting carbon pools directly from observations—we conducted an additional experiment in which the SST and PFT chlorophyll-*a* inputs from the reanalysis were replaced with the corresponding satellite observations assimilated into the reanalysis. Because the satellite products contain substantial data gaps, it was necessary to define a minimum data-availability threshold for aggregating the observations to the 35 km spatial and 10-day temporal resolution required by the NN. For PFT chlorophyll-*a*, the minimum threshold was set to 10% of the maximum possible number of OC-CCI observations within a 35 km, 10-day grid cell, given the native ∼5 km daily resolution of the OC-CCI product. For SST, the threshold was reduced to 5% in order to retain a sufficient number of samples. This lower threshold is justified because the assimilated merged CCI v1 SST product already combines data from multiple satellites, thereby reducing observational uncertainty. After applying these processing steps, just over 50,000 observation-based samples remained for NN testing over the 2016–2020 period—approximately an order of magnitude fewer than in the reanalysis-based test data set.

We have also used the deep ensemble to predict hypothetical what-if scenarios. Two scenarios were chosen for the ensemble prediction: (a) One scenario, where PFT chlorophyll is scaled down from the 2016–2020 reanalysis value gradually to zero, maintaining the same PFT community structure and spatio-temporal distributions as in the reanalysis. This meant taking $\gamma.\text{Chl}(x,t)$, with $\text{Chl}(x,t)$ being the reanalysis chlorophyll-*a* and γ a scaling parameter lowered from 1 to 0. (b) Another scenario, where the ratio of large phytoplankton (sum of diatoms and dinoflaggelates) to total phytoplankton chlorophyll was gradually scaled from the 2016–2020 reanalysis down to zero, but maintaining the same total chlorophyll concentration and spatio-temporal distributions as in the 2016–2020 reanalysis. This means we used $\gamma_1.\text{Chl}_1(x,t)$ and $\gamma_2.\text{Chl}_2(x,t)$, where $\text{Chl}_1(x,t)$ is the large phytoplankton chlorophyll-*a* (diatoms and dinoflaggelates) and $\text{Chl}_2(x,t)$ is the smaller phytoplankton chlorophyll-*a* (nanophytoplankton and picophytoplankton). The γ_1 was a scaling parameter gradually reduced from 1 to 0 and γ_2 another parameter proportionally increased above 1 to maintain that

$$\gamma_1.\text{Chl}_1(x,t) + \gamma_2.\text{Chl}_2(x,t) = \text{Chl}_1(x,t) + \text{Chl}_2(x,t) = \text{Chl}(x,t).$$

In both cases the NN model inputs other than PFT chlorophyll were kept the same as in the reanalysis. These scenarios (defined by the NN inputs) were motivated by certain aspects of future climate projections for the NWES (Wakelin et al. (2015)), but are obviously major simplifications, for example, there is no guarantee that the scenarios are sufficiently self-consistent, as they did not come from a model simulation.

3. Results and Discussion

The deep ensemble performance on validation data, as shown in Table 1, is very good, with R^2 of the surface pools between 0.83 and 0.89, and for the vertically averaged pools in the 0.9–0.92 range, except for the vertically averaged DIC, where it was lower ($R^2 = 0.68$).

Table 1
Ensemble-Averaged Skill of the Deep Ensemble on the 2020 Free-Run Validation Data, Measured for Each Variable Using the R^2 Score, the RMSD, and the Percentage RMSD Relative to the Natural Variability of That Variable, as Quantified by the Standard Deviation (σ) of the Validation Data

	Surface C					Vertically averaged C		
Metric	Detritus	DOC	Zooplankton	Het. Bacteria	DIC	DOC	Het. Bacteria	DIC
R^2	0.83	0.87	0.86	0.85	0.89	0.92	0.9	0.68
RMSE (mmolC/m^3)	28.2	252.6	12.8	4.3	24.4	124.5	2.2	94.2
RMSE σ-norm. (%)	41.5	36	37.1	38.3	33.4	28.6	31.6	56.1

Figure 3 shows the performance of the deep ensemble on both, 2016–2020 free run training-validation data (comparing first and second rows), and the reanalysis test data (comparing third and fourth rows). The Figure compares the deep ensemble biases, as well as biases among the free run (first row) and the reanalysis (third row). It is demonstrated that except for DIC (fifth column) the surface concentrations of the carbon pools in the reanalysis are substantially lower than in the free run (see also Figure 2). This is due to a reduction in phytoplankton concentrations caused by the assimilation of satellite data (see Skákala et al. (2018, 2020)), which propagates to the other organic carbon pools. The deep ensemble, trained on the free-run data, is capable of capturing this pattern, as can be seen by comparing rows three and four in Figure 3. In fact, for detrital matter the deep ensemble predicts a larger reduction than the reanalysis, whereas for zooplankton and bacteria the reduction is slightly smaller. However, in all cases the deep ensemble predictions are much closer to the reanalysis than the free run. Figure 3 also presents, in its bottom row, the uncertainty estimates obtained from the deep ensemble. It

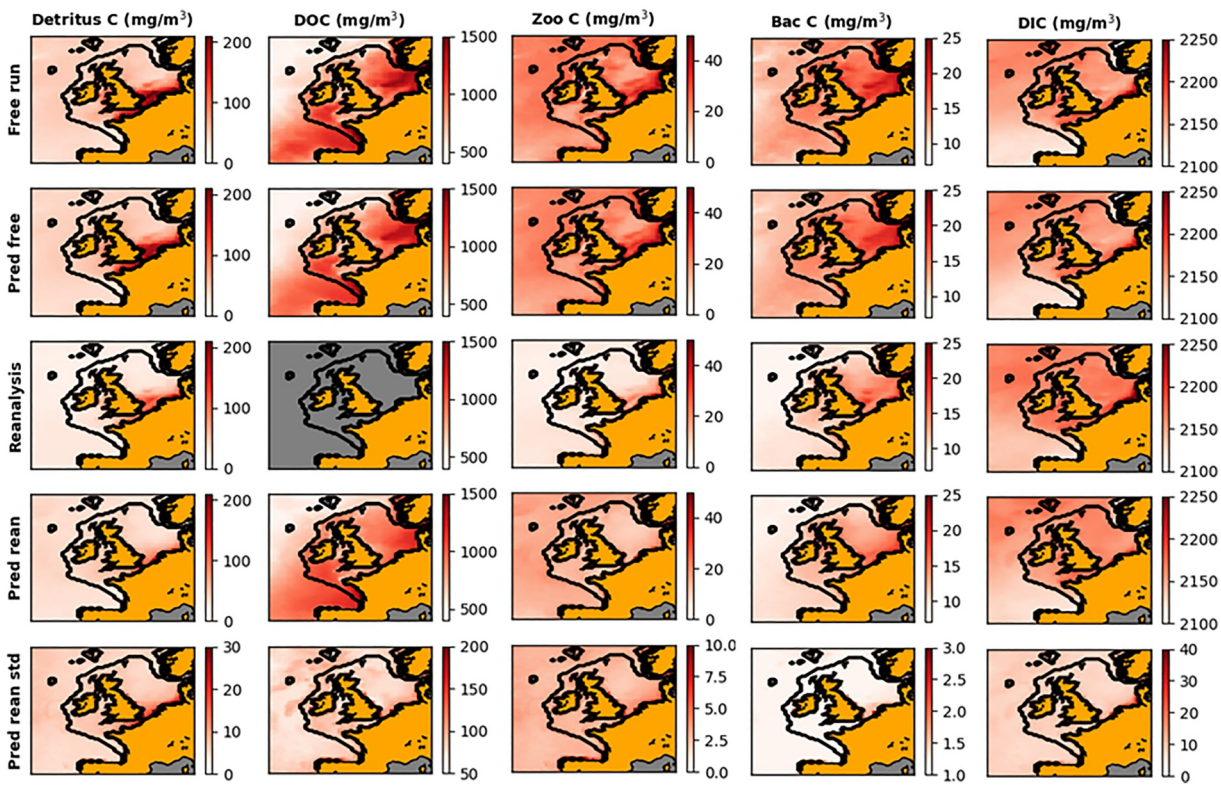

Figure 3. The 2016–2020 average values of different estimated carbon pools comprising ocean surface concentrations of detritus (first column), dissolved organic carbon (DOC) (second column), zooplankton (third column), heterotrophic bacteria (fourth column) and dissolved inorganic carbon (fifth column). Compared are the NEMO-FABM-ERSEM free run providing the training and validation data (first row), the prediction of these data by the mean of the deep ensemble (second row), the Copernicus reanalysis concentrations from Kay et al. (2021) (third row) and the analog of these reanalysis concentrations predicted by the mean of the deep ensemble from the reanalysis inputs (fourth row). In the bottom, fifth row the panels show the uncertainties of the estimated concentrations from the fourth row obtained as the standard deviation of the deep ensemble averaged in time. The reanalysis DOC is masked, since it was not available in the reanalysis outputs.

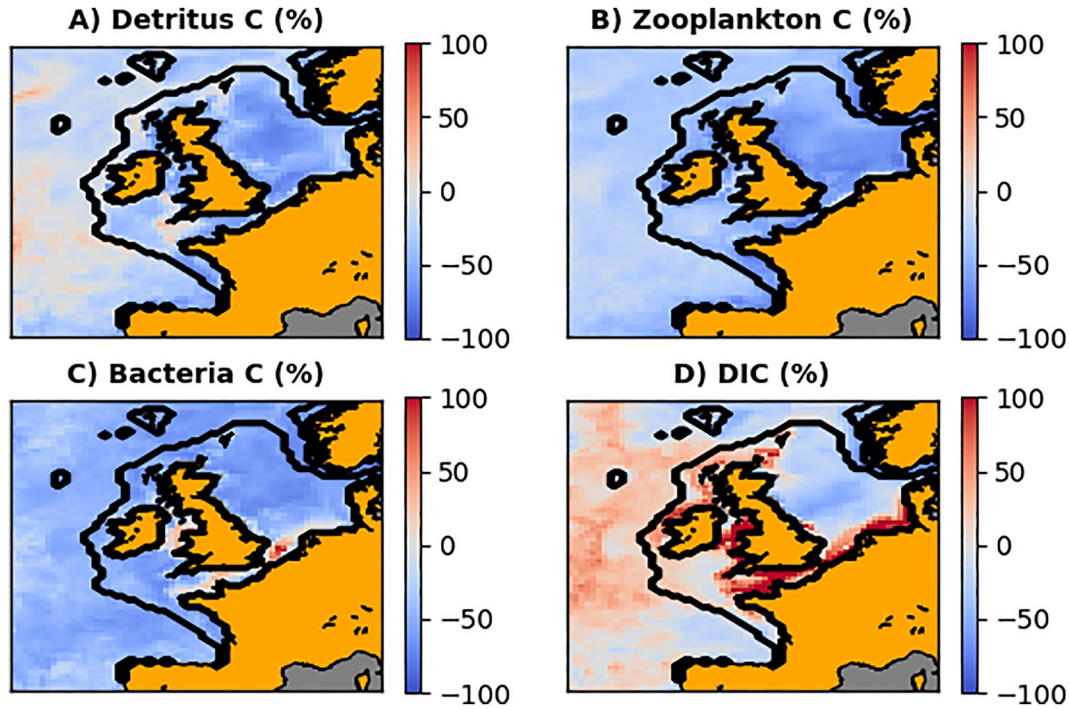

Figure 4. The relative percentage (%) improvement at ocean surface captured by BC-RMSD$_{imp}$ (see Equation 2) as measured relatively to the reanalysis. The blue color means that the deep ensemble mean outperforms the free run and the red color means that it performs worse than the free run. The BC-RMSD's are calculated across the 2016–2020 period.

should be noted that these uncertainties reflect primarily the effects of neural-network optimization (and are therefore predominantly epistemic). They do not account for several important sources of uncertainty in the predicted carbon pools, such as those arising from the input variables or from the learned relationships between the observable variables and the carbon pools.

The comparison in Figure 3 is complemented by BC-RMSD$_{imp}$ metrics from Figure 4, showing the % improvement in BC-RMSD when comparing the BC-RMSD of the deep ensemble prediction using the reanalysis inputs with the BC-RMSD of the free run. The BC-RMSD is in both cases calculated separately for each spatial location and measured relative to the reanalysis data. Figure 4 clearly demonstrates that the deep ensemble prediction from reanalysis inputs substantially outperforms the free run in all surface carbon pools except for DIC (DOC could not be compared due to the lack of reanalysis output). This means for all the surface carbon pools except for DIC, it is preferable to use the deep ensemble prediction to running the free model simulation, and this is true not just for the time-averaged values (Figure 3), but also for the time-series hidden behind the time-averages (Figure 4). We did not identify any clear reason why the model fails to improve DIC relative to the free run, except that the biases between the free run and the reanalysis are substantially smaller for DIC than for the other variables (see Figures 2 and 3). In this context, it should be noted that the regions where the deep ensemble performs worse than the free run for DIC—the southern North Sea and coastal areas around the UK and the Irish Sea (Figure 4d)—correspond to areas where the BC-RMSD of the free run relative to the reanalysis was found to be lowest, indicating its best performance (BC-RMSD <10 mmolC.m^{-3}; these results are not shown). The comparatively small DIC biases in the free run can be attributed to several key drivers that are identical, or very similar, in both the free run and the reanalysis, including air–sea CO_2 fluxes, river discharge, and SST (for SST see Figure 2). Consequently, given the relatively strong performance of the free run for DIC, the deep ensemble faces a substantially more stringent benchmark for further improvement. Since, overall, the deep ensemble performs at a skill level comparable to that of the free run, the deep ensemble DIC product can be considered acceptable.

Curiously deep ensemble failed to improve the vertically averaged pools (not shown here). This can be understood based on SHapley Additive exPlanations (SHAP, Lundberg and Lee (2017)) analysis shown in Figures

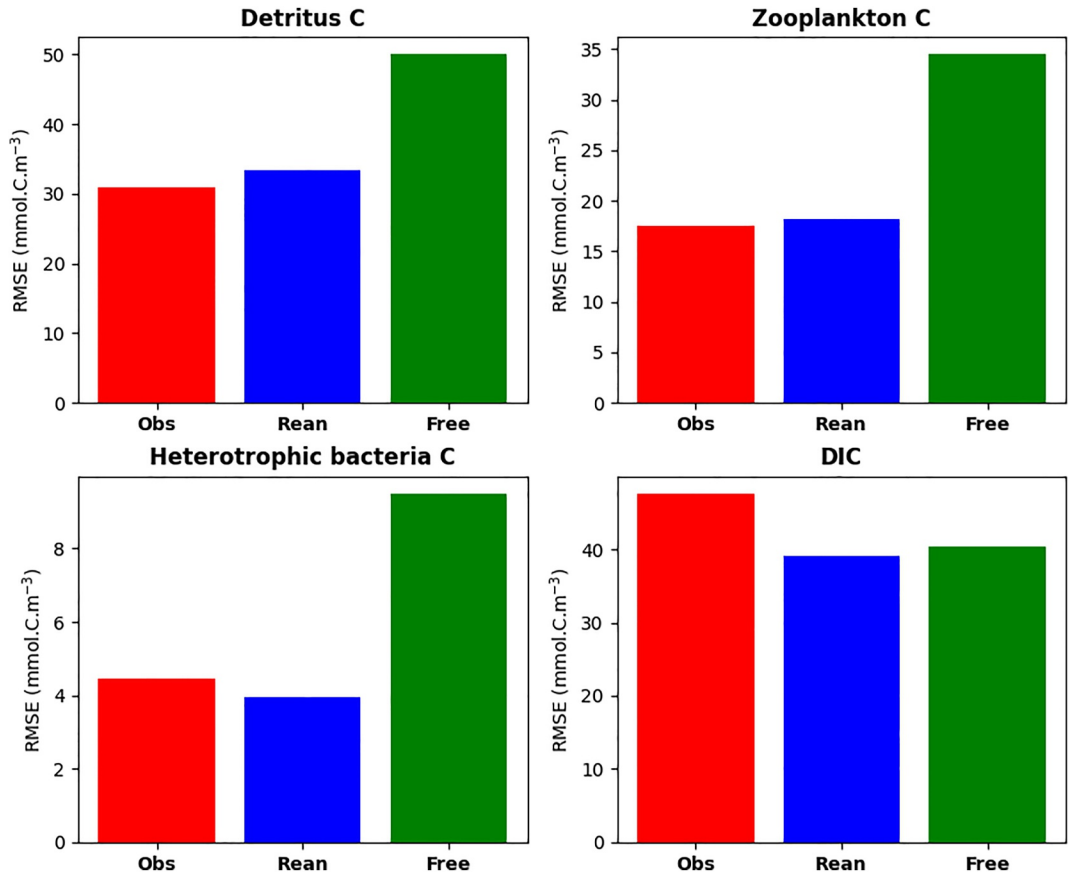

Figure 5. RMSD skill (mmol.C/m^3) comparing performance of deep ensemble prediction of four surface carbon pools, using directly inputs from the assimilated observations (red), reanalysis (blue) and free run (green). The skill was evaluated against reanalysis carbon pools at the locations where satellite observations were available for the 2016–2020 period (amounting to over 50,000 data points, see Section 2.4).

A2–A3 of the Appendix A. Using SHAPS it has been observed that for the vertically averaged variables the structural inputs (coordinates and bathymetry) are more important than they are for the surface variables (Figures A2–A3 of the Appendix A demonstrates this for heterotrophic bacteria carbon and for DOC). High importance of structural variables suggests that the deep ensemble mostly learned the free run climatology of the vertically averaged variables, showing too little flexibility when moving toward reanalysis. We have tried improving upon the structure of the NN models by including time-lagged features, to potentially represent longer time-scales associated with the vertically averaged variables, but there was no marked improvement to the results (not shown here). Finally, the SHAP analysis has shown consistently across all the carbon pools that the two most important groups of variables are oceanic inputs (SST, salinity, PFT chlorophyll) and structural variables, with atmospheric variables being less important and riverine discharge the least important.

Although Figures 2 and 3 show substantial differences between the reanalysis and the free run in both phytoplankton and the predicted carbon pools, one could argue that the reanalysis is not fully independent of the free run used to train the deep ensemble, since both are based on the same underlying physics–biogeochemistry model. To evaluate the skill of the deep ensemble directly against observations (being the ultimate proposed application of the deep ensemble), we therefore applied it to observational data. The validation, expressed in terms of RMSD (Figure 5), indicates that the predictive skill for detritus, zooplankton, and heterotrophic bacteria carbon is comparable when using reanalysis and observational inputs, and substantially better than that of the free run. For DIC, the predictive skill obtained from observational inputs is slightly lower than for the other carbon pools; the underlying causes are likely multifactorial. Nevertheless, these results demonstrate that the deep ensemble is, for the most part, capable of predicting carbon pools from observational data with skill comparable to that achieved using reanalysis inputs (with the exception of DIC). This outcome is not entirely unexpected. Previous studies

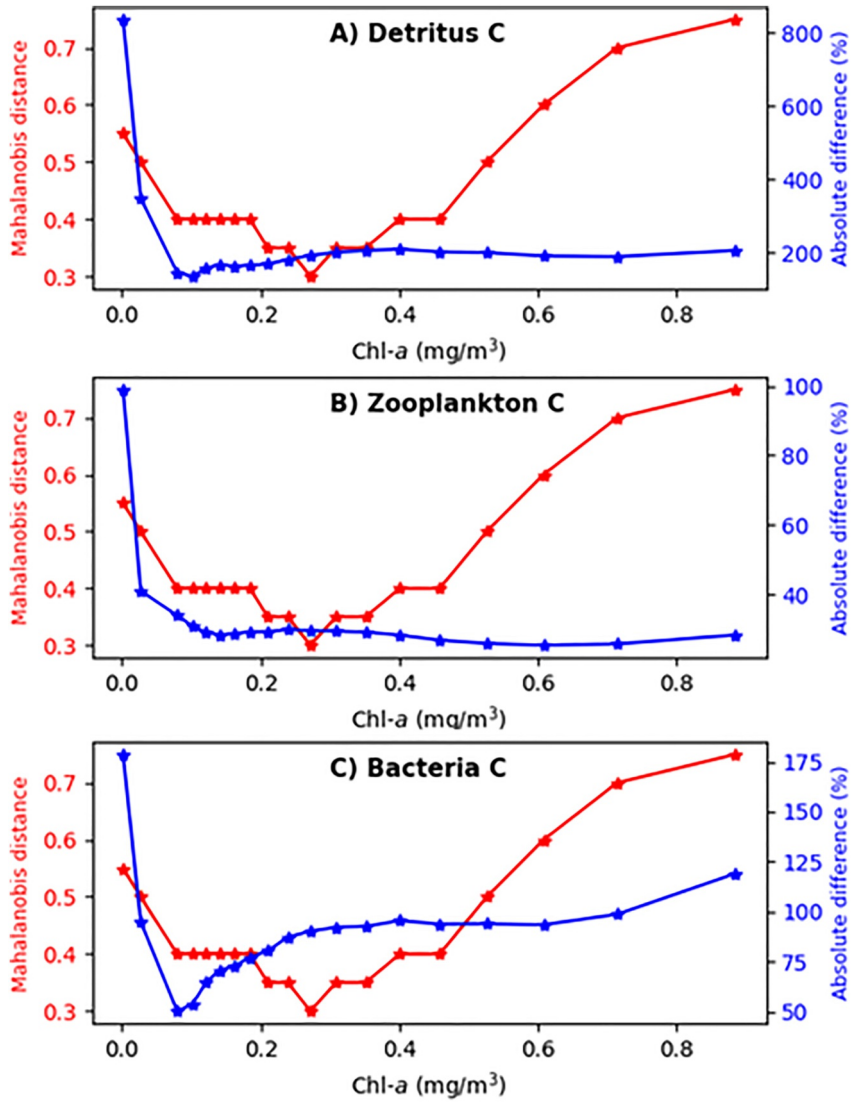

Figure 6. The red line in each plot shows Mahalanobis distance (Chandra (1936); Ghorbani (2019)) of the reanalysis inputs relative to the training and validation (training-validation) data calculated as follows: Chlorophyll-*a* from the reanalysis (*x*-axis) was split into 20 quantiles and for each quantile we calculated median Mahalanobis distance between reanalysis and the training-validation data set. Then we identified where (into which quantile) this median falls within the distribution of Mahalanobis distances within the training-validation data set itself (calculated as distribution of Mahalanobis distances of the training-validation data from the training-validation data set itself). The red *y*-axis then shows those quantiles on the 0–1 range of values, that is, if the value is 0.5 it means that the median distance between reanalysis point and training-validation data set was the same as the median distance among the training-validation data. If it is larger than 0.5, than the reanalysis points were generally further from the training-validation data set than the median and vice versa. The blue line shows for the same quantiles the deep ensemble prediction skill measured against the reanalysis through mean absolute difference relative to the mean concentration (in %).

have shown that the biogeochemical analysis state in the Met Office NWES operational system exhibits close agreement with assimilated observations near the observation locations (e.g., Skákala et al. (2018, 2020, 2021)). Although observational data are inherently noisier than reanalysis products, much of this noise is likely reduced through aggregation to the relatively coarse spatial and temporal resolution used for the NN inputs.

There are two main reasons why the deep ensemble could have failed when predicting the carbon pools from the reanalysis (or observations): (a) the NN inputs provided by the reanalysis were too far from the data seen by the NN in the training process, (b) the inherent relationship between the NN model inputs and the predicted NN outputs, learned by the NN, differs between the reanalysis and the model free run. Figure 6 addresses

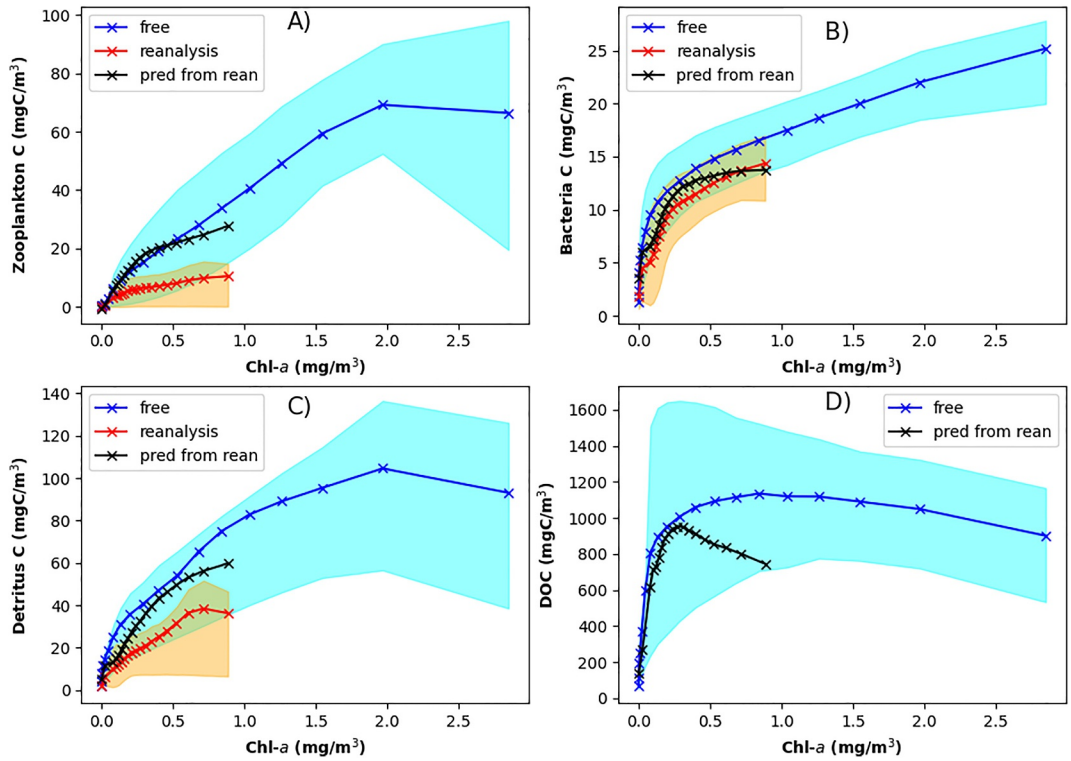

Figure 7. On the *x*-axis we show phytoplankton surface chlorophyll-*a* concentrations (in mg/m^3). The markers on the lines show on the *x*-axis 20 quantiles into which the total chlorophyll-*a* distribution (aggregated across spatial domain and 2016–2020 period) was split. On the *y*-axis the markers show for each chlorophyll-*a* quantile the median values of the corresponding carbon pools: zooplankton (a), heterotrophic bacteria (b), detritus (c) and dissolved organic carbon (d). The shaded colors show the spread of the values for each carbon pool across each chlorophyll-*a* quantile, that is, the spread is defined for each carbon pool as two quartiles around its median value for each chlorophyll-*a* quantile. The differently colored lines and shaded areas account for the model free run (blue), reanalysis (red) and the deep ensemble prediction using reanalysis as inputs (black). For the deep ensemble prediction only the median carbon pools are shown (no shaded areas).

point (a) by showing how Mahalanobis distances (Chandra (1936); Ghorbani (2019)) between the reanalysis and the training-validation data vary across 20 reanalysis chlorophyll quantiles, and how they compare to Mahalanobis distances within the training–validation data set. The Figure also compares these distances with NN performance evaluated against the reanalysis. Figure 6 shows that the reanalysis inputs were not far from what the deep ensemble has seen during training and validation, as their distance from the training and validation data set was on average no significantly larger than the typical distance between the data points within the training and validation data set.

Addressing the point (b) is a little bit more tricky: although the relationship between NN model inputs and the predicted outputs in the reanalysis is provided via the dynamical adjustment of the same biogeochemistry model, whose free run was used to train the NN, there is no guarantee that the map from the NN inputs to the outputs within the reanalysis will not get distorted by the impact of assimilation. An interesting insight is provided by Figure 7 showing the dependence of surface carbon pools on chlorophyll-*a*. Although the dependency plotted in Figure 7 is a major simplification of the true function learned by the deep ensemble, it indicates that the functions between NN inputs and detritus and zooplankton carbon pools might significantly differ between the model free run and the reanalysis. It is also notable that for zooplankton and detritus the deep ensemble predicts similar functional dependence as the free run. This suggests that the deviation between the deep ensemble prediction and the reanalysis in Figures 7a and 7c is due to differences in those functional dependencies between the free run and

the reanalysis, that is, deep ensemble using the function it learned in the free run to create predictions on the space of reduced chlorophyll-*a* concentrations from the reanalysis. Interestingly the same is less true for the heterotrophic bacteria carbon and the DOC (Figures 7b and 7d).

It is possible to use the lightweight NN models to simulate a wide range of what-if scenarios, something hard to do with computationally expensive process-based models. The future climate projections for the NWES indicate important changes to a range of variables included as inputs in the NN (Wakelin et al. (2015)). This includes decline in near-surface primary production and therefore phytoplankton biomass, and changes to phytoplankton community structure, with increased proportion of smaller size-classes in the phytoplankton community. Both these changes are thought to be primarily driven through the increased thermal stratification of the ocean, cutting off the nutrients from the ocean surface (Wakelin et al. (2015)). In Figure 8 we plot the response of ocean surface carbon pools as predicted by the deep ensemble to such changes in the surface phytoplankton (i.e., "what-if" means here "what happens with ocean carbon if such changes occur"). The functional relationships from Figure 8a can be mostly easily understood (except for DOC), as they just indicate an overall decline of organic carbon pools with the decline of phytoplankton biomass. However, the dependence of the same ocean surface carbon pools on the ocean surface phytoplankton community structure (Figure 8b) is in some cases more complex. Whilst the surface detrital carbon monotonically increases with reduced large-size phytoplankton presence likely due to decrease in sinking rates and carbon export, the non-monotonous changes in DOC and zooplankton carbon pools could be influenced by the changes in their own functional type compositions as a function of the modified phytoplankton community structure. It should be noted, however, that the limitations of the NN-based approach are immediately apparent: a decline of phytoplankton biomass to zero (Figure 8a) does not lead to a corresponding reduction of the carbon pools to zero. This indicates that the NN is operating far outside the range over which it was trained and, at some point, begins to underestimate ecosystem changes associated with declining phytoplankton biomass.

There are many reasons to be cautious about the results presented in Figure 8. Even though the what-if scenario input features did not have anomalously large Mahalanobis distance from the training data (not shown here), one can raise serious doubts about how the statistical relationship learned by the deep ensemble in the model hindcast translates into a hypothetical future climate. We have tested this by a series of 1D simulations using Generalized Ocean Turbulence Model (GOTM, Burchard et al. (1999)) coupled to ERSEM through FABM. In those simulations we reduced nutrients by relaxing them toward lowered climatology values, triggering decrease in primary production and phytoplankton in the 1D model. We have trained a deep ensemble based on the simulation data where nutrients were not reduced and we analyzed how well the deep ensemble predicts the carbon pools from the 1D simulations with reduced nutrients. These experiments (see Figure 9) demonstrate that the deep ensemble systematically underestimates the effects of reduced phytoplankton on carbon pools, with the magnitude of this bias increasing as phytoplankton biomass declines. The 1D experiments would then nicely explain why the carbon pools in Figure 8 do not approach zero as the phytoplankton concentration vanishes. However, the deep ensemble was capable to do a decent job in estimating bacteria carbon changes until the phytoplankton biomass was reduced by more than 30% (Figure 9d). This may be explained by the comparatively weaker response of bacterial concentrations to reductions in phytoplankton biomass, relative to the responses of the other carbon pools (at least until phytoplankton declines to approximately half of its baseline concentration; see Figure 9d). The relative stability of bacterial carbon under changing phytoplankton climatology would therefore make it more predictable by an ML model trained on the present-day climate simulation. An interesting feature is the sudden dip in carbon pools relative to phytoplankton biomass reduction around 40% of its present value (Figure 9). We have observed (not shown) that this is due to a sudden change in phytoplankton composition (increase in large species) in such a climate, and interestingly consequences of this change in phytoplankton composition for the carbon pools are broadly correctly predicted by the NN (similar dip is predicted by the NN). Overall, this simple exercise shows that although similar what-if scenarios using deep ensembles should be taken with healthy suspicion, there might be specific cases where the prediction works well within a relatively wide range of scenarios.

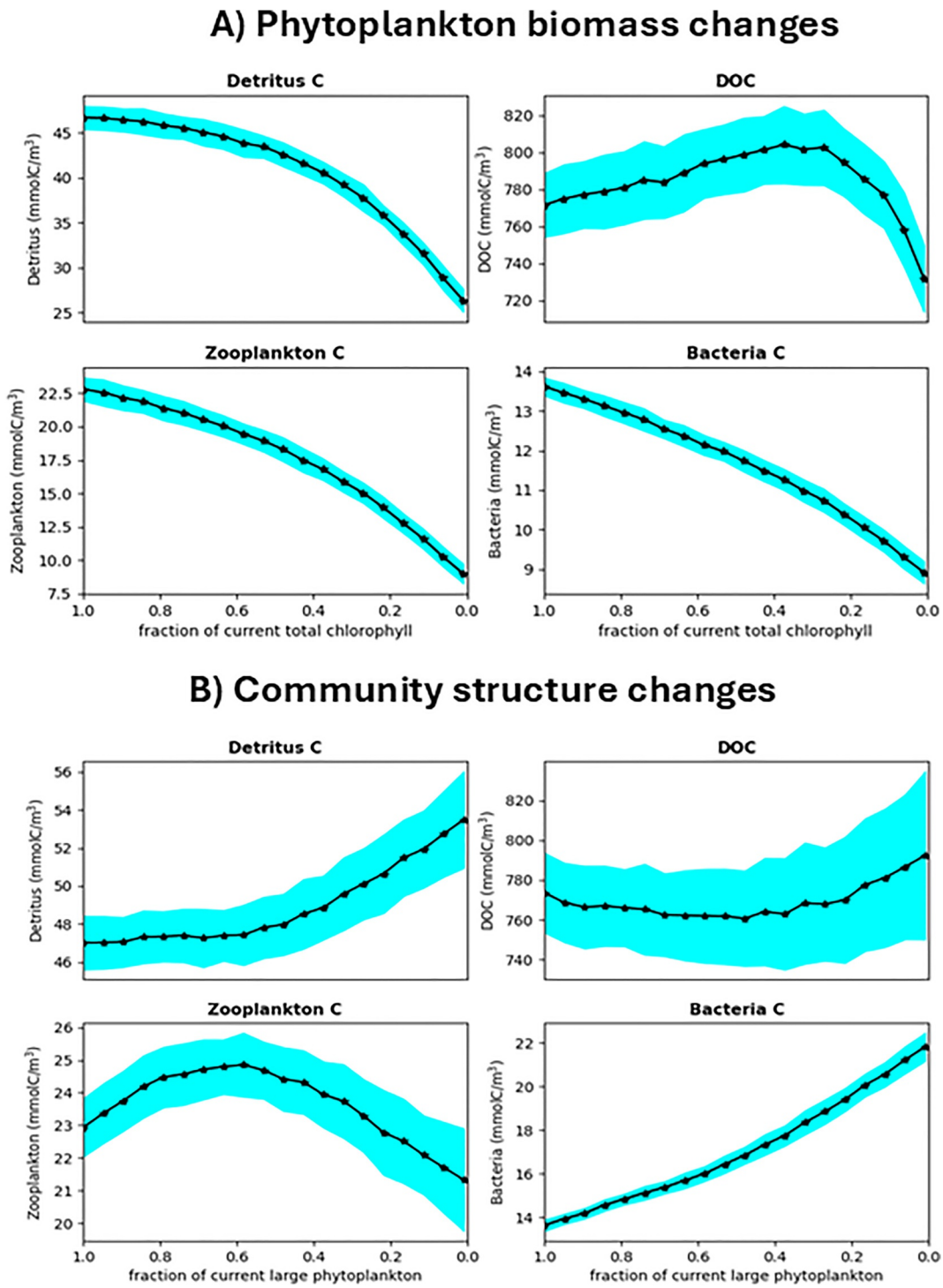

Figure 8. (a) Deep ensemble predicted what-if scenario for how decline in phytoplankton biomass measured as a fraction of its present 2016–2020 reanalysis value (on [0–1] scale) maps into the estimated carbon pools. (b) Similar what-if scenario for changes in the community structure measured by the ratio in chlorophyll-*a* biomass between macrophytoplankton (diatoms and dinoflaggelates) and total phytoplankton, relative to the average reanalysis 2016–2020 ratio (again on a [0–1] scale). The shaded areas show in both cases the uncertainty derived from the deep ensemble standard deviation.

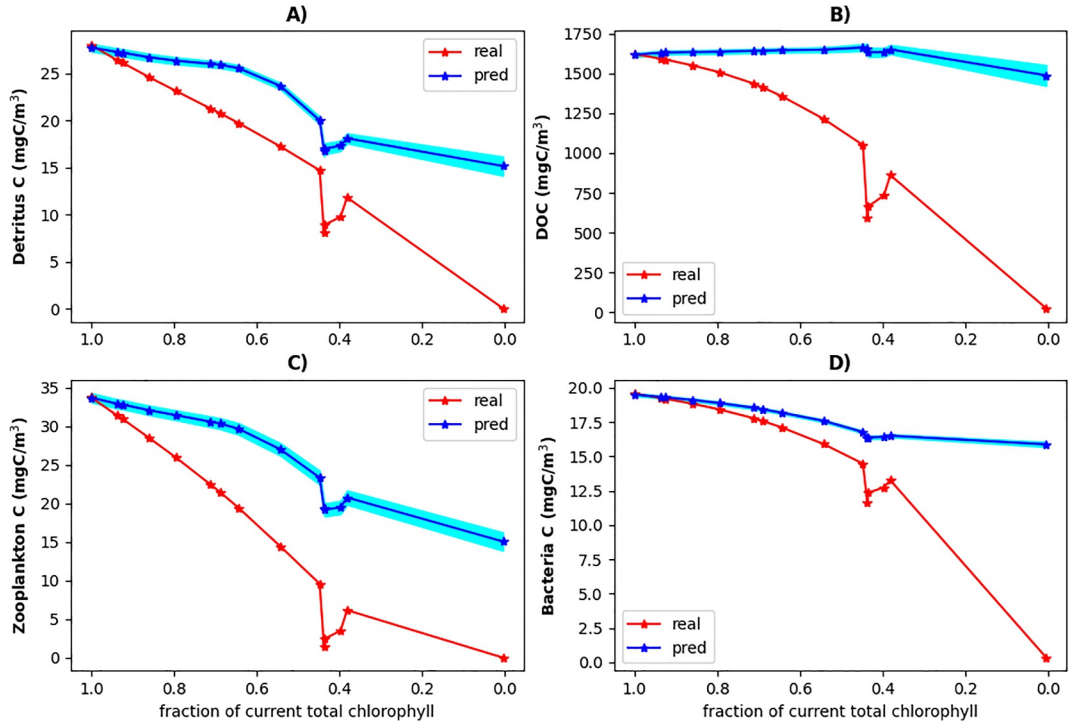

Figure 9. 1D experiments forced by nutrient relaxation to a state with reduced surface phytoplankton concentrations representing changed climate. In red is the relationship between time-averaged total chlorophyll-a (*x*-axis) and time-averaged carbon pools (*y*-axis) from the 1D model runs. Each 1D simulation was run for 20 years with 5 years used as spin-up and 15 years used to calculate the averages for each "climate" state. The simulations were run at L4 location in the western English Channel with atmospheric forcing taken from real 2001–2020 data (hence the only forcing changing the climate state was the nutrient relaxation). In blue is deep ensemble prediction of the surface carbon pools, using the inputs from the 1D runs. The deep ensemble was trained on the simulation with the present state nutrients. The Figure indicates in the simplified 1D setting how much is the deep ensemble capable to generalize from the current climate to different climate states.

Finally, we took initial steps toward estimating the uncertainty in the marine physics–biogeochemistry model–guided, ML-learned function that maps observable variables to carbon pools. The uncertainty in this function is expected to arise mainly from two sources: (a) model structural deficiencies (e.g., errors in the structural form of equations or unresolved variability) and (b) highly uncertain model parameter values (for discussion of this in the context of ERSEM, see Skákala et al. (2024)). Addressing structural deficiencies was beyond the scope of this study; however, we performed initial analyses to estimate the impact of parameter uncertainty. To this end, we used the same 1D model configuration at the L4 location as in the climate scenario experiments, but ran a 100-member ensemble in which five ERSEM model parameters (such as diatom maximum productivity at reference temperature, or the diatom maximum nitrogen-to-carbon ratio) were perturbed. These parameters were selected based on the sensitivity analysis of Ciavatta et al. (2025). The ensemble simulations were conducted using the Ensemble and Assimilation Tool described by Bruggeman and Bolding (2014). The five parameters were sampled from uniform distributions within ±30% of their reference values. One hundred different ML models were then trained using outputs from the 100 ensemble members, which differed only in their parameter values. The 100 ML models were subsequently applied to the same input data, and we evaluated the spread in their predictions of carbon pools. In this simplified setting, prior uncertainty in the model parameters resulted in a typical standard deviation of surface organic carbon pools of approximately 30%–40% of their mean value (not shown). We recommend extending this highly simplified 1D analysis in future work to obtain a more realistic assessment of uncertainty in the ML models.

4. Conclusions and the Next Steps

In this work, we demonstrate that a variety of surface carbon pools (detrital, DOC, zooplankton, heterotrophic bacteria carbon and DIC) in the Northwest European Shelf (NWES) can be estimated (on 35 km and 10-day scale) using model-informed ML. This approach has the potential to serve as a computationally highly efficient substitute for reanalysis systems by predicting the carbon pools directly from atmospheric, riverine data and satellite observations. The ML model, based on a deep ensemble of relatively simple and easy-to-train perceptron neural networks, can generate predictions spanning multiple years of data within seconds on a desktop computer. In contrast, reanalysis requires approximately 1 hour of wall-clock time on a supercomputer using O(100) cores to simulate a single day. However, reanalysis retains a key advantage in its ability to provide more detailed outputs in terms of both variables, spatial and temporal coverage and resolution. The extent to which this level of detail can be reproduced by ML-based approaches remains an open question.

Although the deep ensemble enables rapid predictions, it may require substantial resources for generating training data. In our case, these data are readily available from existing, freely accessible coupled marine physics–biogeochemistry model simulations. The trained deep ensemble was then applied to estimate surface carbon pools (along with their associated uncertainties) directly from observations of physical variables and OC–derived PFT chlorophyll, as well as atmospheric, riverine, and structural data. We have shown that the deep ensemble agrees much better with the reanalysis than the free run for all predicted surface pools where the free run exhibits large biases relative to the reanalysis (i.e., all except DIC).

Such model-informed ML could complement the existing satellite algorithms for some of the carbon pools in regions where they perform poorly, or provide carbon pools where such satellite algorithms are entirely absent. For example, in the case of NWES there are no satellite-derived products for detrital, zooplankton and heterotrophic bacteria carbon (and very little in situ data as well), so the reanalysis assimilating physics data and OC PFT chlorophyll-*a* is taken as the best available estimate of those pools. Wherever the existing spatial and temporal coverage by the satellite data is considered sufficient, we propose to use the ML model developed in this study as an alternative to the relatively computationally costly NWES reanalysis, or a real-time monitoring system. Furthermore, we recommend considering similar tools also in different regions. The degree to which the ML model developed here for the NWES is transferrable to those different regions remains to be determined in the future work. If transferability is an issue, the development of NN models will depend on the availability of model simulations for those new regions, or it will incur additional costs to construct a new training data set.

As the surface carbon pools carry only limited information about carbon cycle, estimating vertical profiles of those pools is quite essential, for example, to get answers to questions such as how much carbon is exported into the deep ocean. The approach adopted in this work was able to only learn the seasonal climatology of the vertically averaged carbon pools. Although there are likely hard limits on how much one can dynamically predict vertically averaged pools from the surface inputs, the relative simplicity of the ML architecture adopted here might have been a contributing factor to ML capturing only the seasonal climatology of the vertically averaged pools. Even though implementing time-lagged features did not seem to help, there is a chance that improvement can be made with more sophisticated models and longer data sets. As part of increasing model complexity, one could also consider replacing the structural ML inputs with new flow-dependent features, for example, capturing the underlying hydrodynamics. Such new features would however likely require using observations from profiles, significantly constraining the spatial domain where the ML model can be applied, or a 3D physical model run, making the applicability of the ML model significantly more demanding in terms of the required inputs into the model (relying on 3D hydrodynamic model runs).

Finally, other applications of the existing ML surface carbon prediction can be imagined, these include implementing assimilation of the ML-derived surface carbon alongside the currently assimilated variables, similarly to what was done with nitrate in Banerjee et al. (2025). If this was done "online" with the ML prediction cycled with the DA (see the discussion in Banerjee et al. (2025)) we could improve the speed with which the model adjusts itself to assimilation of standardly provided satellite data (such as OC chlorophyll-*a*). We suggest to explore these routes in the future.

Appendix A: Figures

Figure A1.

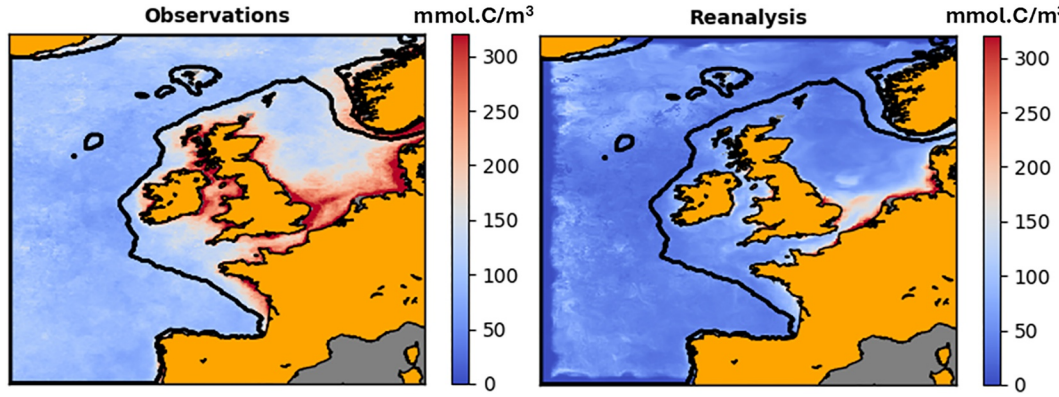

Figure A1. Comparing the Ocean Color–Climate Change Initiative v4.2 satellite-derived product for total particulate organic carbon concentrations (in mgC/m^3), broadly representing aggregate across phytoplankton, zooplankton, bacteria and detrital matter (left-hand panel), with the corresponding Copernicus reanalysis output (right-hand panel). The panels show temporally averaged values across the 2016–2017 period. The reanalysis was masked wherever there were missing satellite data, to ensure like-to-like comparison.

Figure A2.

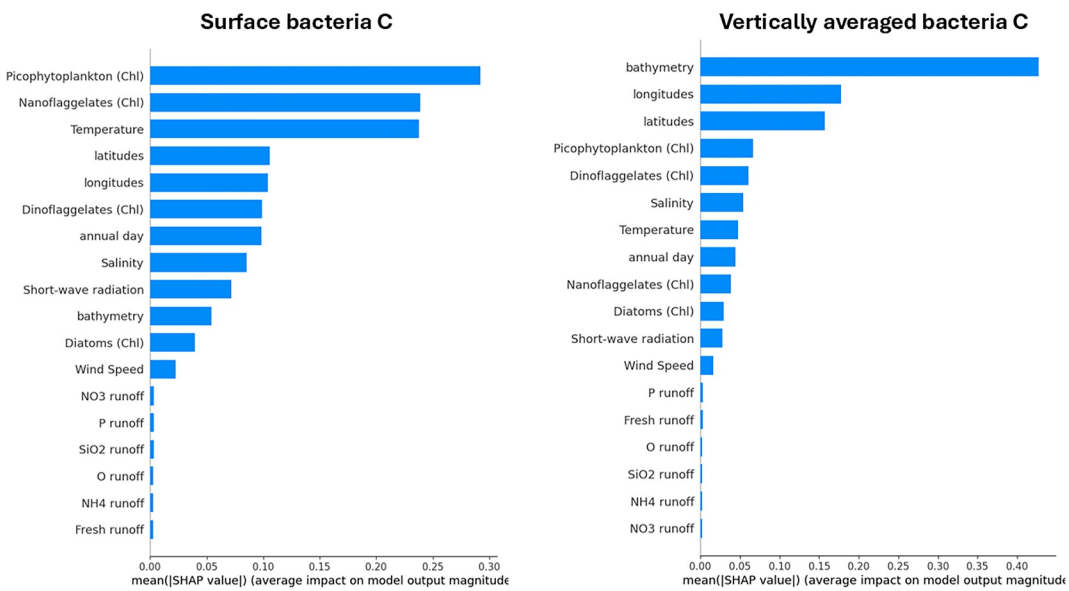

Figure A2. SHapley Aditive exPlanations (SHAP) analysis (shown absolute values) for the surface heterotrophic bacteria carbon (left) and the vertically averaged heterotrophic bacteria carbon (right).

Figure A3.

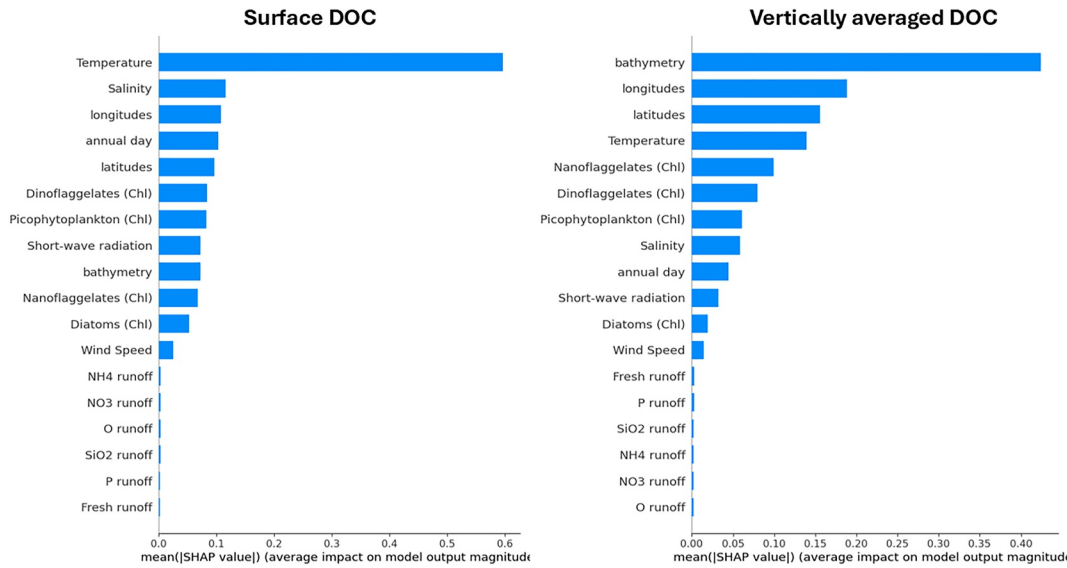

Figure A3. SHAP analysis (shown absolute values) for the surface dissolved organic carbon (DOC) (left) and the vertically averaged DOC (right).

Conflict of Interest

The author declares no conflicts of interest relevant to this study.

Availability Statement

The simulation and reanalysis data are all stored and available through the Met Office MASS system. The full reanalysis can be downloaded under ID: rosie_u-bn555_jan98_pft3.1_bias_r133976_frdf and partly can be obtained from https://data.marine.copernicus.eu/product/NWSHELF_MULTIYEAR_BGC_004_011/description (Kay et al., 2019). The processed version of the 2016–2020 reanalysis can be downloaded from Skákala (2025b). The processed version of the free run can be downloaded from Skákala (2025a), or obtained in full under ID: u-ci193 (for instructions please see https://help.jasmin.ac.uk/docs/mass/). The NN model and the related code (Skákala (2025c)) used in this study can be obtained through Zenodo on https://doi.org/10.5281/zenodo.20276280.

Acknowledgments
This work was funded by the Horizon Europe project The New Copernicus Capability for Tropic Ocean Networks (NECCTON, grant agreement no.101081273). I also acknowledge financial support from the UK Natural Environment Research Council, including the single centre national capability programme—Climate Linked Atlantic Sector Science (Climate Linked Atlantic Sector Science, 379NE/R015953/1), Atlantic Climate and Environment Strategic Science (Atlantis), and National Centre for Earth Observation (NCEO). I would like to thank David Ford and Richard Renshaw for providing me with access to the Copernicus reanalysis data on the Met Office HPC MonSOON2 storage system MASS. I would also like to thank Helen Powley and David Moffat for discussions and Deep S. Banerjee for providing me with processed river data inputs.

References

Artioli, Y., Blackford, J. C., Butenschön, M., Holt, J. T., Wakelin, S. L., Thomas, H., et al. (2012). The carbonate system in the north sea: Sensitivity and model validation. *Journal of Marine Systems*, *102*, 1–13. https://doi.org/10.1016/j.jmarsys.2012.04.006

Bakker, D., Pfeil, B., Smith, K., Hankin, S., Olsen, A., Alin, S., et al. (2014). An update to the surface ocean co 2 atlas (socat version 2). *Earth System Science Data*, *6*(1), 69–90. https://doi.org/10.5194/essd-6-69-2014

Banerjee, D., & Skákala, J. (2025). Improved understanding of nitrate trends, eutrophication indicators, and risk areas using machine learning. *Biogeosciences*, *22*(15), 3769–3784. https://doi.org/10.5194/bg-22-3769-2025

Banerjee, D., Skakala, J., & Ford, D. (2025). Assimilation of machine learning-predicted nitrate to improve the quality of phytoplankton forecasting in the shelf sea environment. *Quarterly Journal of the Royal Meteorological Society*, e70156. preprint. https://doi.org/10.1002/qj.70156

Baretta, J., Ebenhöh, W., & Ruardij, P. (1995). The European regional seas ecosystem model, a complex marine ecosystem model. *Netherlands Journal of Sea Research*, *33*(3–4), 233–246. https://doi.org/10.1016/0077-7579(95)90047-0

Behrenfeld, M., Gaube, P., Della Penna, A., O'malley, R., Burt, W., Hu, Y., et al. (2019). Global satellite-observed daily vertical migrations of ocean animals. *Nature*, *576*(7786), 257–261. https://doi.org/10.1038/s41586-019-1796-9

Borges, A., Schiettecatte, L.-S., Abril, G., Delille, B., & Gazeau, F. (2006). Carbon dioxide in European coastal waters. *Estuarine, Coastal and Shelf Science*, *70*(3), 375–387. https://doi.org/10.1016/j.ecss.2006.05.046

Brewin, R., Ciavatta, S., Sathyendranath, S., Jackson, T., Tilstone, G., Curran, K., et al. (2017). Uncertainty in ocean-color estimates of chlorophyll for phytoplankton groups. *Frontiers in Marine Science*, *104*. https://doi.org/10.3389/fmars.2017.00104

Brewin, R., Sathyendranath, S., Platt, T., Bouman, H., Ciavatta, S., Dall'Olmo, G., et al. (2021). Sensing the ocean biological carbon pump from space: A review of capabilities, concepts, research gaps and future developments. *Earth-Science Reviews*, *217*, 103604. https://doi.org/10.1016/j.earscirev.2021.103604

Bruggeman, J., & Bolding, K. (2014). A general framework for aquatic biogeochemical models. *Environmental Modelling & Software*, *61*, 249–265. https://doi.org/10.1016/j.envsoft.2014.04.002

Burchard, H., Bolding, K., & Villarreal, M. R. (1999). *Gotm, a general ocean turbulence model: Theory, implementation and test cases*. Space Applications Institute.

Butenschön, M., Clark, J., Aldridge, J. N., Allen, J. I., Artioli, Y., Blackford, J., et al. (2016). Ersem 15.06: A generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels. *Geoscientific Model Development*, *9*(4), 1293–1339. https://doi.org/10.5194/gmd-9-1293-2016

Cao, Z., Yang, W., Zhao, Y., Guo, X., Yin, Z., Du, C., et al. (2020). Diagnosis of co2 dynamics and fluxes in global coastal oceans. *National Science Review*, *7*(4), 786–797. https://doi.org/10.1093/nsr/nwz105

Chandra, M. P. (1936). On the generalised distance in statistics. In *Proceedings of the national institute of sciences of India* (Vol. 2, pp. 49–55).

Chau, T. T. T., Gehlen, M., & Chevallier, F. (2022). A seamless ensemble-based reconstruction of surface ocean pco 2 and air–sea co 2 fluxes over the global coastal and open oceans. *Biogeosciences*, *19*(4), 1087–1109. https://doi.org/10.5194/bg-19-1087-2022

Ciavatta, S., Lazzari, P., Alvarez, E., Bertino, L., Bolding, K., Bruggeman, J., et al. (2025). Control of simulated ocean ecosystem indicators by biogeochemical observations. *Progress in Oceanography*, *231*, 103384. https://doi.org/10.1016/j.pocean.2024.103384

Clark, J., Kay, S., Samuelsen, A., Conchon, A., Albernhe, S., Butenschon, M., et al. (2025). Neccton: Technical specification of the pelagic lower trophic level products. *Zenodo*. https://doi.org/10.5281/zenodo.10057818

Dai, M., Su, J., Zhao, Y., Hofmann, E. E., Cao, Z., Cai, W.-J., et al. (2022). Carbon fluxes in the coastal ocean: Synthesis, boundary processes, and future trends. *Annual Review of Earth and Planetary Sciences*, *50*(1), 593–626. https://doi.org/10.1146/annurev-earth-032320-090746

Emerson, S., & Hedges, J. (2008). *Chemical oceanography and the marine carbon cycle*. Cambridge University Press.

Evers-King, H., Martinez-Vicente, V., Brewin, R. J., Dall'Olmo, G., Hickman, A. E., Jackson, T., et al. (2017). Validation and intercomparison of ocean color algorithms for estimating particulate organic carbon in the oceans. *Frontiers in Marine Science*, *4*, 251. https://doi.org/10.3389/fmars.2017.00251

Fablet, R., Beauchamp, M., Drumetz, L., & Rousseau, F. (2021). Joint interpolation and representation learning for irregularly sampled satellite-derived geophysical fields. *Frontiers in Applied Mathematics and Statistics*, *7*, 655224. https://doi.org/10.3389/fams.2021.655224

Friedlingstein, P., O'sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Landschützer, P., et al. (2024). Global carbon budget 2024. *Earth System Science Data Discussions*, *2024*, 1–133.

Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. Facta Universitatis. *Series: Mathematics and Informatics*, 583–595. https://doi.org/10.22190/fumi1903583g

Good, S. A., Martin, M. J., & Rayner, N. A. (2013). En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, *118*(12), 6704–6716. https://doi.org/10.1002/2013jc009067

Grimes, D. J., Ford, T. E., Colwell, R. R., Baker-Austin, C., Martinez-Urtaza, J., Subramaniam, A., & Capone, D. G. (2014). Viewing marine bacteria, their activity and response to environmental drivers from orbit: Satellite remote sensing of bacteria. *Microbial Ecology*, *67*(3), 489–500. https://doi.org/10.1007/s00248-013-0363-4

Groom, S., Sathyendranath, S., Ban, Y., Bernard, S., Brewin, R., Brotas, V., et al. (2019). Satellite ocean colour: Current status and future perspective. *Frontiers in Marine Science*, *6*, 485. https://doi.org/10.3389/fmars.2019.00485

Higgs, I., Skákala, J., Bannister, R., Carrassi, A., & Ciavatta, S. (2024). Investigating ecosystem connections in the shelf sea environment using complex networks. *Biogeosciences*, *21*(3), 731–746. https://doi.org/10.5194/bg-21-731-2024

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jahnke, R. A. (2010). Global synthesis1. In *Carbon and nutrient fluxes in continental margins: A global synthesis* (pp. 597–615). Springer.

Kay, S., McEwan, R., & Ford, D. (2019). *North west european shelf production centre northwestshelf_analysis_forecast_bio_004_011, quality information document*. Copernicus Marine Environment Monitoring Service.

Kay, S., McEwan, R., & Ford, D. (2021). North west European shelf production centre nwshelf_multiyear_bio_004_011. *CMEMS Report*, *3*, 21.

King, R. R., While, J., Martin, M. J., Lea, D. J., Lemieux-Dudon, B., Waters, J., & O'Dea, E. (2018). Improving the initialisation of the met office operational shelf-seas model. *Ocean Modelling*, *130*, 1–14. https://doi.org/10.1016/j.ocemod.2018.07.004

Laine, M., Kulk, G., Jönsson, B. F., & Sathyendranath, S. (2024). A machine learning model-based satellite data record of dissolved organic carbon concentration in surface waters of the global open ocean. *Frontiers in Marine Science*, *11*, 1305050. https://doi.org/10.3389/fmars.2024.1305050

Le, C., Zhou, X., Hu, C., Lee, Z., Li, L., & Stramski, D. (2018). A color-index-based empirical algorithm for determining particulate organic carbon concentration in the ocean from satellite observations. *Journal of Geophysical Research: Oceans*, *123*(10), 7407–7419. https://doi.org/10.1029/2018jc014014

Legge, O., Johnson, M., Hicks, N., Jickells, T., Diesing, M., Aldridge, J., et al. (2020). Carbon on the northwest European shelf: Contemporary budget and future influences. *Frontiers in Marine Science*, *7*, 143. https://doi.org/10.3389/fmars.2020.00143

Lenhart, H.-J., Mills, D. K., Baretta-Bekker, H., Van Leeuwen, S. M., Van Der Molen, J., Baretta, J. W., et al. (2010). Predicting the consequences of nutrient reduction on the eutrophication status of the north sea. *Journal of Marine Systems*, *81*(1–2), 148–170. https://doi.org/10.1016/j.jmarsys.2009.12.014

Li, C., Wu, H., Yang, C., Cui, L., Ma, Z., & Wang, L. (2024). Advanced machine learning models for estimating the distribution of sea-surface particulate organic carbon (poc) concentrations using satellite remote sensing data: The mediterranean as an example. *Sensors*, *24*(17), 5669. https://doi.org/10.3390/s24175669

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/tnnls.2021.3084827

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

MacLachlan, C., Arribas, A., Peterson, K., Maidens, A., Fereday, D., Scaife, A., et al. (2015). Global seasonal forecast system version 5 (glosea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1072–1084. https://doi.org/10.1002/qj.2396

Madec, G., Bourdallé-Badie, R., Bouttier, P. A., Bricaud, C., Bruciaferri, D., Calvert, D., et al. (2015). Nemo ocean engine.

Mannino, A., Russ, M. E., & Hooker, S. B. (2008). Algorithm development and validation for satellite-derived distributions of doc and cdom in the us middle Atlantic bight. *Journal of Geophysical Research*, *113*(C7). https://doi.org/10.1029/2007jc004493

Matsuoka, A., Boss, E., Babin, M., Karp-Boss, L., Hafez, M., Chekalyuk, A., et al. (2017). Pan-arctic optical characteristics of colored dissolved organic matter: Tracing dissolved organic carbon in changing arctic waters using satellite ocean color data. *Remote Sensing of Environment*, *200*, 89–101. https://doi.org/10.1016/j.rse.2017.08.009

Mogensen, K., Balmaseda, M. A., & Weaver, A. (2012). The nemovar ocean data assimilation system as implemented in the ecmwf ocean analysis for system 4.

O'Dea, E., Furner, R., Wakelin, S., Siddorn, J., While, J., Sykes, P., et al. (2017). The co5 configuration of the 7 km Atlantic margin model: Large-scale biases and sensitivity to forcing, physics options and vertical resolution. *Geoscientific Model Development*, *10*(8), 2947–2969. https://doi.org/10.5194/gmd-10-2947-2017

Racault, M.-F., Abdulaziz, A., George, G., Menon, N., C, J., Punathil, M., et al. (2019). Environmental reservoirs of vibrio cholerae: Challenges and opportunities for ocean-color remote sensing. *Remote Sensing*, *11*(23), 2763. https://doi.org/10.3390/rs11232763

Raissi, M., Perdikaris, P., & Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Roobaert, A., Laruelle, G. G., Landschützer, P., Gruber, N., Chou, L., & Regnier, P. (2019). The spatiotemporal dynamics of the sources and sinks of co2 in the global coastal ocean. *Global Biogeochemical Cycles*, *33*(12), 1693–1714. https://doi.org/10.1029/2019gb006239

Roobaert, A., Resplandy, L., Laruelle, G. G., Liao, E., & Regnier, P. (2024). Unraveling the physical and biological controls of the global coastal co2 sink. *Global Biogeochemical Cycles*, *38*(3), e2023GB007799. https://doi.org/10.1029/2023gb007799

Roy, S., Sathyendranath, S., & Platt, T. (2017). Size-partitioned phytoplankton carbon and carbon-to-chlorophyll ratio from ocean colour by an absorption-based bio-optical algorithm. *Remote Sensing of Environment*, *194*, 177–189. https://doi.org/10.1016/j.rse.2017.02.015

Sathyendranath, S., Brewin, R. J., Brockmann, C., Brotas, V., Calton, B., Chuprin, A., et al. (2019). An ocean-colour time series for use in climate studies: The experience of the ocean-colour climate change initiative (oc-cci). *Sensors*, *19*(19), 4285. https://doi.org/10.3390/s19194285

Sathyendranath, S., Platt, T., Kovač, Ž., Dingle, J., Jackson, T., Brewin, R. J., et al. (2020). Reconciling models of primary production and photoacclimation. *Applied Optics*, *59*(10), C100–C114. https://doi.org/10.1364/ao.386252

Sauzède, R., Johnson, J., Claustre, H., Camps-Valls, G., & Ruescas, A. (2020). Estimation of oceanic particulate organic carbon with machine learning. *ISPRS Annals of Photogrammetry*, *2*, 949–956. https://doi.org/10.5194/isprs-annals-v-2-2020-949-2020

Scarselli, F., Gori, M., Tsoi, A., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605

Siddorn, J., & Furner, R. (2013). An analytical stretching function that combines the best attributes of geopotential and terrain-following vertical coordinates. *Ocean Modelling*, *66*, 1–13. https://doi.org/10.1016/j.ocemod.2013.02.001

Skákala, J. (2025a). Nemo-fabm-ersem free run 2016-2020 simulations [Dataset]. *GitHub*. Retrieved from https://github.com/JOZSKA/ML_for_carbon_pools

Skákala, J. (2025b). Nemo-fabm-ersem reanalysis 2016-2020 simulations [Dataset]. *GitHub*. Retrieved from https://github.com/JOZSKA/ML_for_carbon_pools

Skákala, J. (2025c). Nn code and weights [Software]. *GitHub*. Retrieved from https://github.com/JOZSKA/ML_for_carbon_pools

Skákala, J., Bruggeman, J., Brewin, R. J., Ford, D. A., & Ciavatta, S. (2020). Improved representation of underwater light field and its impact on ecosystem dynamics: A study in the north sea. *Journal of Geophysical Research: Oceans*, *125*(7), e2020JC016122. https://doi.org/10.1029/2020jc016122

Skákala, J., Bruggeman, J., Ford, D., Wakelin, S., Akpınar, A., Hull, T., et al. (2022). The impact of ocean biogeochemistry on physics and its consequences for modelling shelf seas. *Ocean Modelling*, *172*, 101976. https://doi.org/10.1016/j.ocemod.2022.101976

Skákala, J., Ford, D., Brewin, R. J., McEwan, R., Kay, S., Taylor, B., et al. (2018). The assimilation of phytoplankton functional types for operational forecasting in the northwest European shelf. *Journal of Geophysical Research: Oceans*, *123*(8), 5230–5247. https://doi.org/10.1029/2018jc014153

Skákala, J., Ford, D., Fowler, A., Lea, D., Martin, M. J., & Ciavatta, S. (2024). How uncertain and observable are marine ecosystem indicators in shelf seas? *Progress in Oceanography*, *224*, 103249. https://doi.org/10.1016/j.pocean.2024.103249

Skákala, J., Ford, D. A., Bruggeman, J., Hull, T., Kaiser, J., King, R. R., et al. (2021). Towards a multi-platform assimilative system for ocean biogeochemistry. *Journal of Geophysical Research: Oceans*, *126*(4), e2020JC016649. https://doi.org/10.1029/2020jc016649

Stramski, D., Reynolds, R. A., Babin, M., Kaczmarek, S., Lewis, M. R., Röttgers, R., et al. (2008). Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern south Pacific and eastern Atlantic oceans. *Biogeosciences*, *5*(1), 171–201. https://doi.org/10.5194/bg-5-171-2008

Strömberg, K. P., Smyth, T. J., Allen, J. I., Pitois, S., & O'Brien, T. D. (2009). Estimation of global zooplankton biomass from satellite ocean colour. *Journal of Marine Systems*, *78*(1), 18–27. https://doi.org/10.1016/j.jmarsys.2009.02.004

Telszewski, M., Palacz, A., & Fischer, A. (2018). Biogeochemical in situ observations–motivation, status, and new frontiers. *New Frontiers in Operational Oceanography*, 131–160.

Volk, T., & Hoffert, M. I. (1985). Ocean carbon pumps: Analysis of relative strengths and efficiencies in ocean-driven atmospheric co2 changes. *The carbon cycle and atmospheric CO2: Natural variations Archean to present*, *32*, 99–110.

Wakelin, S., Artioli, Y., Butenschön, M., Allen, J. I., & Holt, J. T. (2015). Modelling the combined impacts of climate change and direct anthropogenic drivers on the ecosystem of the northwest European Continental shelf. *Journal of Marine Systems*, *152*, 51–63. https://doi.org/10.1016/j.jmarsys.2015.07.006

Wakelin, S., Holt, J., Blackford, J., Allen, I., Butenschön, M., & Artioli, Y. (2012). Modeling the carbon fluxes of the northwest European continental shelf: Validation and budgets. *Journal of Geophysical Research*, *117*(C5). https://doi.org/10.1029/2011jc007402

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Zemskova, V., He, T., Wan, Z., & Grisouard, N. (2022). A deep-learning estimate of the decadal trends in the Southern ocean carbon storage. *Nature Communications*, *13*(1), 4056. https://doi.org/10.1038/s41467-022-31560-5

Zhang, Z., Chen, P., Zhang, S., Huang, H., Pan, Y., & Pan, D. (2025). A review of machine learning applications in ocean color remote sensing. *Remote Sensing*, *17*(10), 1776. https://doi.org/10.3390/rs17101776

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., et al. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001