Università degli Studi di Trieste

Dipartimento di Matematica, Informatica e Geoscienze

XXXVI Ciclo Del Dottorato Di Ricerca In
in
Scienze della Terra, Fluidodinamica e Matematica. Interazioni e Metodiche.

Machine Learning applications to data reconstruction in Marine Biogeochemistry

Settore scientifico-disciplinare: INF-01

Candidata:
Pietropolli Gloria

Relatori:
Prof. Luca Manzoni
Dott. Gianpiero Cossarini

Coordinatore:
Prof. Stefano Maset

Anno Accademico 2022/2023

“Soprattutto, non affrettare il viaggio;
fa che duri a lungo, per anni, e che da vecchio
metta piede sull’isola, tu, ricco
dei tesori accumulati per strada
senza aspettarti ricchezze da Itaca.

Itaca ti ha dato il bel viaggio,
senza di lei mai ti saresti messo
sulla strada: che cos’altro ti aspetti?”

Konstantinos Kavafis

Abstract

Driven by the increase of greenhouse gas emissions, climate change is causing significant shifts in the Earth's climatic patterns, profoundly affecting our oceans.

In recent years, our capacity to monitor and understand the state and variability of the ocean has been significantly enhanced, thanks to improved observational capacity, new data-driven approaches, and advanced computational capabilities. Contemporary marine analyses typically integrate multiple data sources: numerical models, satellite data, autonomous instruments, and ship-based measurements. Temperature, salinity, and several other ocean essential variables, such as oxygen, chlorophyll, and nutrients, are among the most frequently monitored variables. Each of these sources and variables, while providing valuable insights, has distinct limitations in terms of uncertainty, spatial and temporal coverage, and resolution.

The application of deep learning offers a promising avenue for addressing challenges in data prediction, notably in data reconstruction and interpolation, thus enhancing our ability to monitor and understand the ocean. This thesis proposes and evaluates the performances of a variety of neural network architectures, examining the intricate relationship between methods, ocean data sources, and challenges. A special focus is given to the biogeochemistry of the Mediterranean Sea.

A primary objective is predicting low-sampled biogeochemical variables from high-sampled ones. For this purpose, two distinct deep learning models have been developed, each specifically tailored to the dataset used for training. Addressing this challenge not only boosts our capability to predict biogeochemical variables in the highly heterogeneous Mediterranean Sea region but also allows the increase in the usefulness of observational systems such as the BGC-Argo floats. Additionally, a method is introduced to integrate BGC-Argo float observations with outputs from an existing deterministic marine ecosystem model, refining our ability to interpolate and reconstruct biogeochemical variables in the Mediterranean Sea.

As the development of novel neural network methods progresses rapidly, the task of establishing benchmarks for data-driven ocean modeling is far from complete. This work offers insights into various applications, highlighting their strengths and limitations, besides highlighting the importance relationship between methods and datasets.

List of Publications

Thesis co-author Acknowledgments

All the contributions presented in this thesis have been published or have been accepted for an open discussion on EGUsphere. I would like to extend my deepest gratitude to all the co-authors –listed above– who helped me produce the work for this thesis.

In particular: Chapter 4 is based on [1], coauthored by Luca Manzoni from the University of Trieste and Cossarini Gianpiero from National Institute of Oceanography and Experimental Geophysics (OGS). Moreover, Chapter 4 is partially based on [2], a joint work with Carolina Amadio, Anna Teruzzi, Luca Manzoni, Gianluca Coidessa, and Gianpiero Cossarini, all from OGS. The complete version of the latter article can be found in Appendix A. Chapter 5 is based on [3], a joint work with Luca Manzoni and Gianpiero Cossarini. Chapter 6 is based on [4], coauthored by Luca Manzoni and Gianpiero Cossarini.

General co-author Acknowledgments

Throughout my three years of doctoral studies, I have contributed to several publications. While not all are included in this thesis, those marked with ⋆ form the foundational basis of this dissertation. Publications not incorporated were excluded primarily because their focus diverged from the central theme of this thesis. Nonetheless, each of these works aligns with my educational journey, rooted deeply in computer science and algorithms.

Peer-Reviewed Journal Publications

- [5] Castelli, M., Manzoni, L., Mariot, L., Menara, G., **Pietropolli, G. (2022)**. The Effect of Multi-Generational Selection in Geometric Semantic Genetic Programming. *Applied Sciences*.

- [6] Leporati, A., Manzoni, L., Mauri, G., **Pietropolli, G.**, Zandron, C. (2023). Inferring P systems from their computing steps: An evolutionary approach. *Swarm and Evolutionary Computation*.

- ⋆ [1] **Pietropolli, G.**, Manzoni, L., Cossarini, G. (2023). Multivariate Relationship in Big Data Collection of Ocean Observing System. *Applied Sciences*.

- ⋆ [3] **Pietropolli, G**., Manzoni, L., Cossarini, G. (2023). PPCon 1.0: Biogeochemical Argo Profile Prediction with 1D Convolutional Networks. *EGUsphere*.

- [7] Nadizar, G., **Pietropolli, G.** (2023). A grammatical evolution approach to the automatic inference of P systems. *Journal of Membrane Computing*.

- ⋆ [2] Amadio, C., Teruzzi, A., **Pietropolli, G.**, Manzoni, L., Coidessa, G., Cossarini, G. (2023). Combining Neural Networks and Data Assimilation to enhance the spatial impact of Argo floats in the Copernicus Mediterranean biogeochemical model. *EGUsphere*.

- [8] **Pietropolli, G.**, Menara, G., Castelli, M. (2023). A Genetic Programming Based Heuristic to Simplify Rugged Landscapes Exploration. *Emerging Science Journal*.

Peer-Reviewed Conference Publications

- [9] **Pietropolli, G.**, Manzoni, L., Paoletti, A., Castelli, M. (2022, April). *Combining geometric semantic gp with gradient-descent optimization.* Genetic Programming: 25th European Conference, Held as Part of EvoStar 2022 (**EuroGP 2022**).

- ⋆ [4] **Pietropolli, G.**, Cossarini, G., Manzoni, L. (2022, September). *GANs for Integration of Deterministic Model and Observations in Marine Ecosystem.* In Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence (**EPIA 2022**).

- [10] Ferreira, J., Castelli, M., Manzoni, L., **Pietropolli, G.** (2023, March). *A Self-Adaptive Approach to Exploit Topological Properties of Different GAs' Crossover Operators.* In Genetic Programming: 26th European Conference, Held as Part of EvoStar 2023 (**EuroGP 2023**).

- [11] **Pietropolli, G.**, Camerota Verdù, F. J., Manzoni, L., Castelli, M. (2023, July). *Parametrizing GP Trees for Better Symbolic Regression Performance through Gradient Descent.* In Proceedings of the Companion Conference on Genetic and Evolutionary Computation (**GECCO 2023**).

Contents

List of Figures

List of Tables

A Gnan Liliano Claudio,

Ho aspettato a dedicarti qualcosa
Perchè volevo dedicarti qualcosa di abbastanza grande

Chapter 1

Introduction

Climate change, often termed global warming, represents a persistent alteration in Earth's climatic patterns, primarily driven by a surge in greenhouse gas emissions ([13]). Human activities are the predominant contributors to this escalation, as emphasized in various studies ([14, 15]).

Oceans play a crucial role in mitigating climate change by absorbing approximately 93% of the excess heat generated by human activities. This function positions them as a primary buffer against global warming. However, they are vulnerable to human-induced perturbations, yielding complex repercussions ([16–18].

Firstly, rising global temperatures, influenced by anthropogenic factors, result in oceanic warming ([19, 20]). These changes in ocean temperatures and chemistry exert a profound impact on marine ecosystems, leading to shifts in species distributions, altered migration patterns, and, in some cases, the bleaching and mortality of coral reefs ([21]). These disruptions can have cascading effects throughout the entire marine food web. The effects of increasing temperatures extend to the melting of polar ice sheets and glaciers. This melting is a significant contributor to sea-level rise, endangering coastal habitats, and ecosystems, and intensifying the risk of coastal inundation ([22, 23]).

Additionally, greenhouse gas emissions increase atmospheric CO_2 levels. The oceans absorb this excess CO_2, leading to a decrease in pH levels, resulting in ocean acidification ([24, 25]). This phenomenon poses severe threats to marine life, especially to species like corals and shellfish that depend on calcium carbonate to form their shells and skeletons ([26]).

Lastly, climate change can exacerbate extreme weather events, such as hurricanes and typhoons, with devastating consequences for coastal areas and the ocean environment ([27]).

Analyzing the complex interplay between human activities, oceanic processes, and marine ecosystems provides essential knowledge for defining strategies that ensure sustainable ocean health.

The Essential Ocean Variables (EOVs) present a structured approach to observing and monitoring global oceans. Defined by the Global Ocean Observing System (GOOS, [28]), these variables capture critical metrics for interpreting the ocean's dynamic nature and its reactions to external shifts, notably the consequences of climate change ([29, 30]). Key EOVs include sea surface temperature, which plays a pivotal role in understanding the ocean's climatic influence, and sea surface salinity, a key indicator of ocean circulation and freshwater input. Additionally, the pH of seawater, ocean oxygen levels, and nutrient concentrations are vital for tracking ocean acidification, ecosystem health, and nutrient cycling. These EOVs, along with others such as chlorophyll concentration, ocean color, and biological diversity, are essential as part of a global effort to comprehensively observe and comprehend the ocean's pivotal role in our changing world.

Given the essential nature of these variables, marine analysis makes use of a diverse spectrum of data sources to measure them, each with its unique advantages and disadvantages. These sources vary considerably in terms of coverage and the quality of predictions they provide. Ship-based observations, with their origins dating back to the 18th century, offer invaluable historical data ([31]). However, they come with limitations, particularly in terms of their sparse coverage, especially in the deeper layers of the ocean. In comparison, autonomous instruments, introduced after the year 2000, provide real-time data collection across a wider expanse compared to ship-based measurements ([32–35]). Nonetheless, they may exhibit less precision. Satellite observations give a panoramic view of the surface of oceans, enabling estimations of parameters like sea surface temperature and chlorophyll ([36, 37]). However, as said before, their scope is restricted to surface-level data. Numerical models, in the form of ocean physical-biogeochemical models (OBMs), utilize mathematical equations to simulate intricate ocean processes ([38–40]). These models can be used to better understand the oceans, provide forecasting, and are able to reconstruct entire domains. Nevertheless, they are susceptible to errors stemming from lack and simplifications in their mathematical formulations. In essence, the integration of these diverse data sources presents both opportunities to better understand the oceans and challenges stemming from the nature of the data and our knowledge of the oceans' processes.

With the increase in the volume of new data, studying the ocean started to require the capabilities of Artificial Intelligence (AI), a domain that has witnessed rapid advancements in recent decades. Within AI, neural networks (NN) are one of the most prominent tools and Deep Learning (DL), particularly stands out for its success and the wide range of possible applications ([41–43]).

DL techniques are known for their ability to learn from (almost) raw data without the need for complex feature construction steps. Over recent years, DL has received significant attention – even from the general public – for its ability to be applied to large and complex datasets obtaining human-level results on multiple tasks. This makes them useful for tasks where manual extraction of patterns would be too labor-intensive or too complex ([44–46]).

Given the complexity and multi-dimensionality of oceanographic data, DL techniques emerge as apt tools, offering valuable modeling capabilities and the ability to handle vast, intricate datasets effectively ([47–49]). Compared to conventional statistical methods, DL excels in processing data characterized by many dimensions, complex inter-variable relationships, and occasional missing values. Thus, integrating these techniques into oceanography promises to augment current methodologies in a significant way.

1.1 Modeling the Mediterranean Sea Biogeochemistry

Our research centers on the Mediterranean Sea basin, often referred to as a miniature ocean. ([50]). Despite its relatively small size, the Mediterranean contains oceanographic processes emblematic of larger oceans, making it a natural laboratory for oceanographic studies.

One defining feature of the Mediterranean is its semi-enclosed nature, mostly surrounded by land, with limited connections to the Atlantic Ocean through the Strait of Gibraltar. This aspect gives rise to several unique oceanographic features.

As stated above, one key characteristic is its limited connection to the Atlantic Ocean, which translates to a diminished water exchange ([51]). This factor is instrumental in the creation of distinct deep water masses, including the Mediterranean Deep Water. Moreover, the Mediterranean is characterized by elevated salinity levels ([52]), surpassing the average salinity found in oceans. This heightened salinity, attributed to the imbalance between high evaporation rates and freshwater influx, makes the Mediterranean an ideal subject for exploring saltwater dynamics and their influence on circulation patterns. In fact, the Mediterranean exhibits complex circulation patterns ([53]) driven by wind, temperature, and density differences. Additionally, owing to its geographical confinement and diverse coastal landscapes, the Mediterranean exhibits significant biogeochemical variations ([54]). Scientists often use it to investigate the effects of nutrients, pollution, and climate change on marine ecosystems in enclosed seas. Furthermore, the Mediterranean's pronounced sensitivity to climatic indicators renders it a crucial zone for the study of climate variability and its long-term shifts ([55]): analyzing historical climate records through sediment cores and proxies provides information about past climate variations.

Transitioning from the general oceanographic attributes of the Mediterranean, our research is about its biogeochemistry, a field that intertwines biological, geological, and chemical processes to understand the composition, causes, and consequences of chemical substances found in the marine environment ([56]). Biogeochemistry focuses on how living organisms, including microorganisms, plants, and animals, influence and are influenced by the Earth's chemical environment ([57]). This interplay determines the distribution and abundance of marine life and affects the ocean's elemental cycles, including carbon, nitrogen, and phosphorus ([58]).

Key variables in biogeochemical studies encompass nutrients (e.g., nitrate, phosphate), trace metals, organic matter, dissolved gases (e.g., oxygen, carbon dioxide), and various isotopic tracers ([59]). Each of these components provides information about the different biogeochemical processes, from primary production to organic matter decomposition, and offers clues regarding the health and functioning of marine ecosystems.

In this context, our research concentrates on the use of machine learning (ML) techniques to understand the biogeochemical processes of the Mediterranean and to provide information about the variables integral to its biogeochemistry.

1.2 Approach

Oceanography spans a large variety of data types, each characterized by distinct spatial coverage of the marine domain, presenting both opportunities to increase our understanding of the oceans and challenges due to the different ways data are collected.

Numerical models simulate various marine variables across the entire domain, yielding three-dimensional (3D) datasets that simulate oceanic processes. In contrast, satellite data provides two-dimensional (2D) surface-level snapshots, capturing metrics like sea surface temperature and chlorophyll. However, satellite data can only capture surface measurements and depend on cloud cover and atmospheric conditions. Autonomous instruments, which measure at fixed locations while varying depth, produce one-dimensional (1D) vertical profiles. Ship-based measurements, with their inherent data sparsity, capture zero-dimensional (0D) data points. These isolated points, while crucial, necessitate tailored techniques for effective integration into comprehensive oceanographic studies.

Due to the large variety of the kind of data available, it is important to design different neural network architectures that can exploit the strengths and mitigate the inherent biases of each data type—be it measurement inaccuracies, sampling biases, or data collection disparities, and, most important, the spatial dimensionality of the data.

This research aims to create a framework that can address oceanographic questions across different data dimensions. This extends beyond DL model creation; it encompasses the rationale behind choosing specific DL architectures tailored to distinct research challenges.

This study initiates with an analysis of the 0D scenario, targeting the prediction of low-sampled biogeochemical variables from high-sampled ones using a training dataset derived from cruise ship measurements. To work with this data format we employ a straightforward architecture: a multilayer perceptron model that works with point-like input-output pairs.

Transitioning to the 1D domain, the focus shifts to vertical profiles sourced from autonomous instruments. A 1D convolutional neural network is proposed to process these profiles, accepting vectorial inputs and producing vectorial outputs.

Both applications aim to estimate low-sampled marine variables leveraging high-sampled ones. Yet, the distinct datasets underpinning the DL models —specifically the 0D samples in the former and the 1D samples in the latter — dictate a shift in the architecture used.

As the last step, the research broadens to cover the full 3D Mediterranean domain. The proposed DL model forecasts this domain through 2D horizontal maps at varied depths and is based on 2D convolutional neural networks. Training data combines inputs from numerical simulations, satellite observations, and autonomous instrument observations. The aim is to craft a DL model able to predict biogeochemical variables, synthesizing the missing data from these diverse data sources.

For every application, the choice of network architecture is guided by its ability to model the specific features and biases inherent in the data. Detailed explanations for these architectural choices are provided in dedicated chapters.

In essence, this study focuses on DL techniques tailored to handle different data characteristics and dimensions in oceanography. Moving from 0D to 1D, and then employing 2D input representations to model a 3D domain, the aim is to understand the strengths and weaknesses of various neural network architectures. Through this approach, it is possible to find the advantages and limitations of these architectures in their application to the oceanographic domain.

1.3 Contributions

This section provides an overview of the different tasks tackled in this thesis, highlighting the three primary investigations undertaken. A detailed discussion of each of them will be presented in the subsequent chapters. Specifically: *MLP for the Prediction of Biogeochemical Variables* will be the topic of Chapter 4; *PPCon: 1D CNN for Predicting Nutrient Profiles* will be the topic of Chapter 5; and *GANs for Integrating Deterministic Model and Observations.* will be the topic of Chapter 6.

1.3.1 MLP for the Prediction of Biogeochemical Variables.

This research addresses the 0D problem, characterized by both input and output features in a scalar form. The goal is the development of DL techniques for predicting relationships among biogeochemical variables, emphasizing the prediction of low-sampled variables from high-sampled ones. The dataset comprises bottle measurements from scientific cruises, offering high-precision data due to laboratory-based analyses. However, the limited (and highly sparse) number of samples limits the ability to find patterns or relationships among data ([60, 61]). Given the absence of vertical or horizontal relations in this dataset, a straightforward DL architecture suffices for predictions.

Consequently, a Multilayer Perceptron is employed ([62]) to predict marine variables, such as nutrient concentrations and carbonate system variables, based on temporal and locational data, and high-frequency parameters like temperature, salinity, and oxygen levels. To improve upon existing models, we employ a larger dataset for training (EMODnet, [63]) and make appropriate adjustments to the network structure. This approach introduces two innovations with respect to the state-of-the-art in oceanography: first, we derive confidence intervals for predictions by combining multiple NNs. Second, we implement a two-step evaluation process to assess the quality of the dataset and eliminate spurious data before training. The resultant model demonstrates better prediction accuracy compared to existing models.

In order to provide an example of how this research can be useful in the oceanographic field, this chapter also proposed a practical application of the proposed model for the data assimilation task.

1.3.2 PPCon: 1D CNN for Predicting Nutrient Profiles.

This research extends the previously introduced task of inferring low-sampled variables from high-sampled ones.

Instead of ship measurements, this study uses samples from autonomous Argo floats ([64]). While obtaining dense oceanic measurements across temporal and spatial scales remains intricate, Biogeochemical (BCG) Argo profiling floats have revolutionized subsurface oceanic data acquisition.

Argo's distinction from bottle samples lies in its vertical resolution. Unlike bottle samples, which capture data every tens or hundreds of meters, floats measure at intervals spanning mere meters, yielding dense profile data such as temperature, chlorophyll, and oxygen profiles. Each profile—a curve plotting variable values against depth—can be represented as a vector.

Given this data format, a DL architecture that actually makes use of the vertical profile data format can enhance the prediction quality. Multilayer Perceptrons, while efficient, have difficulties capturing the innate shapes of biogeochemical variable profiles. Hence, their predictions often manifested irregularities, like abrupt transitions between two values at very similar depths.

To actually use the unique structure of vertical profiles, a one-dimensional Convolutional Neural Network model is the one introduced for nutrient profile forecasting ([65]). We present the PPCon (Predict Profiles Convolutional) model—trained on BCG Argo float data—to predict nitrate, chlorophyll, and backscattering (bbp700) values using data parameters like date, location, and associated profiles of temperature, salinity, and oxygen. The quality of the results is validated both quantitatively and visually, emphasizing its advantage in generating smoother, more accurate profiles with reduced errors over its MLP counterpart.

1.3.3 GANs for Integrating Deterministic Model and Observations.

The main objective of this study is to define and implement a DL model that can reproduce the biogeochemical variables within the Mediterranean Sea's 3D domain. To achieve this we employ 2D horizontal maps, each one representing the marine domain at various depth levels. These maps serve as a bridge between in-situ and satellite observations and the outputs from an existing deterministic marine ecosystem model.

Monitoring marine ecosystems traditionally makes use of two methods: direct observations (either in-situ or satellite-derived) and deterministic models. Observations, while potentially precise, may not offer complete spatial or temporal coverage. In contrast, deterministic models provide a comprehensive view of marine ecosystems but may contain inaccuracies.

This research introduces a DL technique that combines information from both the deterministic models and in-situ and satellite observations. The inputs and outputs of this DL model are 2D sections of the Mediterranean and can be visualized as instantaneous pictures of specific variables at varying depths within the sea. Such a structure lends itself to a 2D convolutional-based architecture, which can make use of the spatial structure inherent in these sections.

Specifically, this study makes use of an inpainting DL model built on Generative Adversarial Networks. Two DL models using this inpainting approach are proposed: EmuMed which acts as a surrogate for the deterministic model by mirroring its outputs; and InpMed which is an evolved version of EmuMed that augments the model by assimilating data from in-situ and satellite observations.

Empirical evaluations reveal EmuMed's ability to replicate the deterministic model's results. In the subsequent phase of training, the use of observation refines the entire model. This results in marine fields more closely aligned with the picture presented by actual observations.

1.4 Thesis Outline

Following this introductory chapter (Chapter 1), this thesis is structured as follows:

- **Chapter 2** (*Neural Network*) introduces the main principles and mechanics of NNs. The chapter elaborates on essential topics, such as NN training dynamics and backpropagation. It also provides an overview of significant machine learning paradigms and delves into specific NN architectures used in this research, i.e. Multilayer Perceptrons, Convolutional Neural Networks, and Generative Adversarial Networks. The chapter concludes by explaining the concept of Relational Inductive Biases, a critical construct for understanding the inherent capabilities of artificial NN in finding complex relationships, especially within data-intensive domains like environmental, oceanographic, and geophysical datasets.
- **Chapter 3** (*Machine Learning for Ocean Biogeochemistry*) discusses the intersection of machine learning methodologies and oceanographic applications. The chapter delineates primary oceanographic data sources and shows the inherent complexities of working with these datasets. It then provides a structured review of recent

progress in the use of machine learning techniques for various oceanographic data and tasks.

- **Chapter 4** (*MLP for the Prediction of Nutrients and Carbonate System Variables*) introduces the MLP developed for forecasting nutrient concentrations and carbonate system variables in a 0D context.
- **Chapter 5** (*PPCon: 1D CNN for Predicting Nutrient Profiles*) presents the CNN architecture that takes advantage of the characteristic 1D shape of vertical profiles to improve the predictions of biogeochemical variables.
- **Chapter 6** (*GANs for Integrating Deterministic Models and Observations*) outlines the inpainting architecture utilized to fuse observations into deterministic model output, resulting in an improved ability to reconstruct the 3D field of the Mediterranean basin.

Chapter 2

Neural Network

Artificial Intelligence (AI) is a subfield of computer science with a strong interdisciplinary component that aims to create intelligent agents capable of mimicking human-like cognitive functions, such as learning, reasoning, problem-solving, perception, and decision-making. In the general and long-term goals of AI research, we can find the ability to build machines and algorithms that can analyze complex data, adapt to different scenarios, and perform tasks that typically require human intelligence ([41]). Within the broader domain of AI, two prominent branches have emerged: Machine Learning and its sub-field of Deep Learning.

Machine Learning (ML) is a subfield of AI that focuses on the development of algorithms and statistical models that allow computers to learn patterns and insights from data without being explicitly programmed for a specific task ([66]). In contrast to traditional rule-based programming or symbolic approaches, ML algorithms use data to iteratively adjust their parameters and improve their performance (measured according to some predefine metrics). The key characteristic of ML is that the resulting models can be able to generalize from the data they have been trained on to make predictions on new, unseen data. Common types of ML algorithms include decision trees ([67]), support vector machines ([68]), k-nearest neighbors ([69]), and most notably, artificial neural networks ([70]).

Deep Learning (DL) is a specialized subfield of ML that focuses on using Artificial Neural Networks (ANN) to find patterns in data ([42]). In their general form, ANNs consist of interconnected units called neurons, arranged in layers: there's an input layer to receive data, several hidden layers that process the data, and an output layer for predictions. The term deep in DL refers to the depth of the NN, indicating the presence of multiple hidden layers.

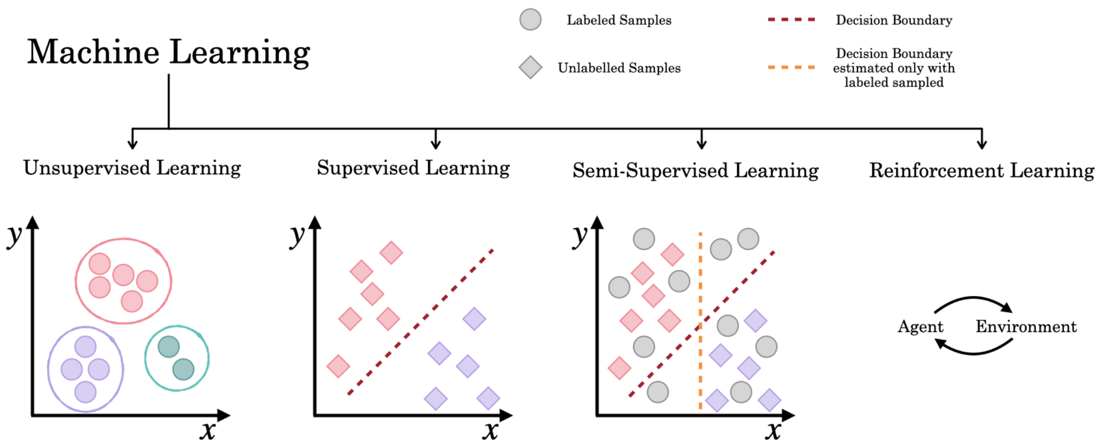


FIGURE 2.1: Overview of the ML Categories

This is in contrast with "classical" (or "shallow") ANN, where only a few hidden layers were used.

Unlike traditional ML algorithms, which often require handcrafted feature engineering, DL algorithms can automatically learn relevant features from raw data, making them highly effective in tasks where feature extraction would be challenging or time-consuming. In fact, these deep architectures enable the system to learn progressively abstract and more complex representations of the data, thus performing this feature engineering step by themselves.

In recent years, DL has obtained significant attention - even from the general public - due to its ability to handle large-scale, unstructured data with exceptional accuracy. These algorithms have been successfully applied to a wide range of applications, such as image recognition ([44]), natural language processing ([45]), recommendation systems ([46]), and many others. It has led to significant advancements in areas such as self-driving cars ([71]), medical diagnosis ([72]), and language translation ([73]).

ML (and so DL) is organized into different categories based on the way the learning process works and on the presence of labeled data during the training phase. Depending on the category, there can be distinct learning methods and architectures that are specific to the task being tackled. The main categories of DL include supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning ([74]). An intuitive representation of the principal categories that constitute ML is reported in Figure 2.1.

Supervised learning is the most common category of DL, where the algorithm learns from labeled examples provided in the training data ([75]). The labeled data consists of input-output pairs, with the model learning to map inputs to the corresponding correct outputs. In supervised learning, the deep NN is trained to minimize the difference between its predicted outputs and the true labels through the use of a loss function.

Unsupervised Learning, in contrast to supervised learning, involves training the model on unlabeled data without explicit output labels ([76]). The objective of unsupervised learning is to discover patterns or relationships within the data or different representations. Unsupervised learning plays a vital role in tasks like generating realistic synthetic data.

Semi-supervised learning combines elements of both supervised and unsupervised learning ([77]). In this approach, the algorithm is provided with a limited amount of labeled data and a more extensive set of unlabeled data. The model uses the labeled data to guide its learning process, while also taking advantage of the additional unlabeled data.

Lastly, reinforcement learning is a different paradigm, where the algorithm learns to interact with an environment and receive feedback (rewards or penalties) based on its actions ([78]). Examples include agents mastering complex games like Go and various Atari videogames without prior game knowledge ([79], [80]), robotic control for tasks like walking and manipulation ([81]), and optimizing recommendation systems in digital platforms ([82]).

This thesis will focus on utilizing supervised and unsupervised learning methods for modeling biogeochemical variables. Consequently, these methods will be the only ones that will be detailed in this Chapter. Specifically, Section 2.1 will provide a more detailed description of the basic notion of the Artificial NN, including a mathematical explanation of its learning process. Within Section 2.2, we present an explanation of the supervised learning techniques and architectures used in this thesis, namely, the Multilayer Perceptron and the Convolutional Neural Network (CNN). On the other hand, Section 2.3 will describe the unsupervised learning method employed in this research, i.e. Generative Adversarial Nets (GANs). In conclusion, Section 2.4 will explain the concept of relational inductive biases.

2.1 Artificial Neural Networks

Artificial Neural Networks (ANN) ([83]), often referred to as *Neural Networks* (NN), represent a class of computational models inspired by the functioning of biological neurons

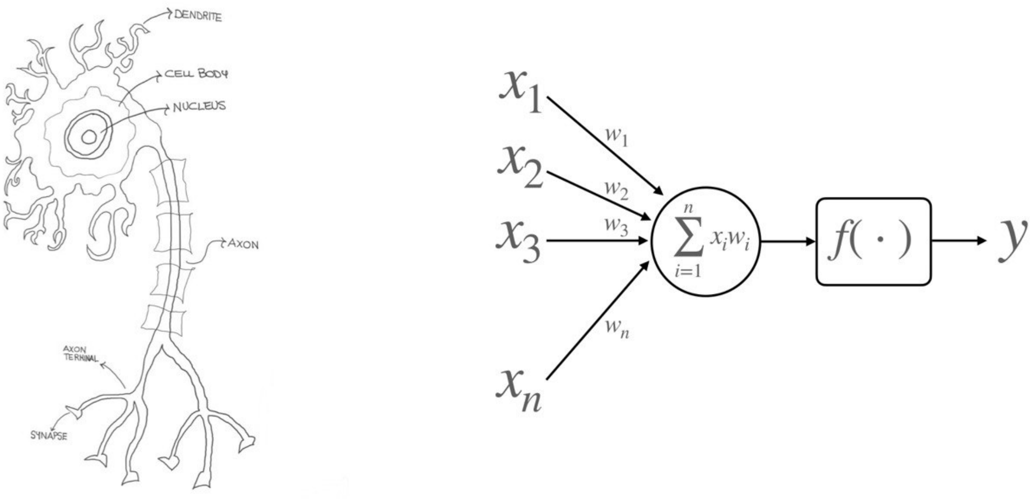


FIGURE 2.2: Working principle of ANNs: on the left is represented a schematic drawing of biological neurons, on the right a schematic drawing of an artificial neuron.

in the human brain. We can place the birth of ANN with the advent of artificial neurons proposed by McCulloch and Pitts [84] in 1943. This concept evolved with Rosenblatt's introduction of the perceptron, an early algorithm designed for pattern recognition using a two-layer network ([85]). The initial excitement was tempered when Minsky and Papert [86] highlighted the perceptron's limitations, especially its inability to solve problems non-linear problems like XOR, leading to a period of stagnation in the NNs domain. The 1980s witnessed a renaissance in the field, driven by the introduction of the backpropagation algorithm, which made training multi-layer networks feasible ([87]). The real acceleration, however, came in the 21st century. The combination of larger datasets, like ImageNet, and increased computational power, particularly by making use of GPUs, set the stage for deep NN to outperform other methods in various tasks ([88]).

As said before, central to ANN is the concept of the artificial neuron, inspired by biological neurons in the human brain. An intuitive visualization of the similarity between a biological neuron and an artificial neuron is represented in Figure 2.2. Artificial neurons, also known as nodes or units, take in input data and process it using trainable weights, denoted as θ_i, where i ranges across all input connections (plus a constant input or *bias*). Once the inputs are multiplied by their corresponding weights, the resulting weighted sum is passed through a (non-linear) activation function ϕ, introducing non-linearity to the network.

Formally, each neuron operates by aggregating the weighted inputs as:

$$z = \sum_{i=1}^{n} \phi_i \cdot x_i + b \tag{2.1}$$

Here, z denotes the weighted sum and b is the bias term. The inputs x_i of the neuron can be either the actual inputs of the entire networks or the outputs of other neurons. Following this summation, the value z goes through a non-linear activation function $f(\cdot)$, yielding the neuron's output:

$$y = f(z) \tag{2.2}$$

By applying non-linear activation functions, feedforward NNs can represent non-linear functions, which are needed to approximate any non-trivial real-world dataset.

By deciding how the artificial neurons are connected, i.e, the architecture of the network, we can tailor the NN to specific tasks, leading to various types of NNs. One of the most common families of architectures is the feed-forward NN, where data flows in one direction, from the input layer through hidden layers to the output layer. Inside this family we can find, for example, CNN, specifically designed to process spatially structured data, such as images and videos. In this case the connections will be only local, taking advantage of the spatial correlation of the data.

2.1.1 Training a Neural Network

In this section, we present a brief overview of the methods employed to train a neural network.

Essential to the training of any ML method is the dataset. In the supervised paradigms, datasets consist of pairs, each containing an input x, the features, and an associated output y, the intended prediction. A NN is then a function f mapping each input x to an output which we expect to be (near to) y. Since the output depends on the set of parameters ϕ of the NN, we denote the function computer by an NN by $f(x;\phi)$.

In order to be able to test correctly the generalization ability (i.e., the prediction ability on unseen data) of the model, datasets are partitioned into:

- Training Set: Used for training the model
- Validation Set: Post initial training, this subset aids in model refinement, hyperparameter adjustments, and mitigates overfitting risks. It provides the possibility

to perform some checks without making use of the test data, which should remain unseen until

- Test Set: Following training and validation, this subset is used for assessing the model on unobserved data, in order to understand if it has learned how to generalize well.

The training phase aims to refine the network's weights θ_i minimizing the difference between its predictions, $f(x;\theta)$, and actual outputs, y. Successful training converges $f(x;\theta)$ to y not only on the training set but also on the test set.

This convergence is evaluated using the *loss function* $\mathcal{L}$, which measures the difference between predicted and true outputs, guiding parameter optimization ([89]). One common loss function in regression problems is the *Root Mean Squared Error* $\mathcal{L}_{rmse}(\theta)$, which is:

$$\min_{\theta} \mathcal{L}_{rmse}(\theta) = \sum_{(x,y)} (f(x;\theta) - y)^2 \tag{2.3}$$

Such minimization of the loss function cannot be performed analytically due to the depth and the complex architectures of NNs. Instead, backpropagation and stochastic gradient descent are used, as detailed in the next section.

Notice that this assumes a fixed network structure, but the NN's architecture and the actual selection of optimization algorithms and regularization measures critically influence the network's adaptability to learn and to adapt to novel data ([90]).

2.1.2 Stochastic Gradient Descent and Backpropagation

Gradient descent (GD) consists of an iterative algorithm optimized to find a function's local minimum ([90]). For NNs, the target for minimization is the defined loss function, $\mathcal{L}$. GD refines the network's weights in alignment with the negative gradient direction, as follows:

$$\theta_{k+1} = \theta_k - \gamma_k \nabla \mathcal{L}(\theta_k) \tag{2.4}$$

In this equation, θ_k denotes current model parameters, γ_k represents the learning rate, dictating the optimization step magnitude, and $\nabla \mathcal{L}(\theta_k)$ the gradient of the loss function computed with respect to the parameter θ_k.

Despite its utility, gradient descent in DL exhibits certain limitations, including issues with vanishing/exploding gradients, susceptibility to local minima, sensitivity

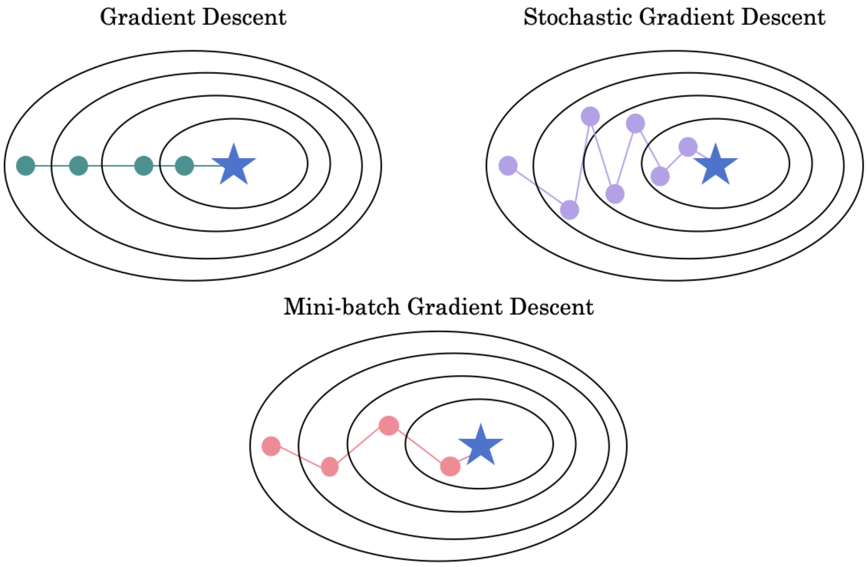


FIGURE 2.3: Illustration of the different gradient-based methods.

to learning rate and initialization, and extensive training time requirements due to the fact that the computation of the loss function requires the iteration across the entire training set.

A refined version, *Stochastic Gradient Descent* (SGD), offers a stochastic gradient approximation ([91]). For a loss function defined as $\mathcal{L}(\theta) = \frac{1}{n}\sum_i^n \mathcal{L}i(\theta)$, with each $\mathcal{L}i$, weights can be updated per:

$$\theta_{k+1} = \theta_k - \gamma_k \nabla \mathcal{L}_i(\theta_k) \tag{2.5}$$

Conceptually, SGD adds noise to GD, implying that:

$$\mathbb{E}[\nabla \mathcal{L}_i(\theta_k)] = \nabla \mathcal{L}(\theta_k) \tag{2.6}$$

Thus, expected SGD updates align with gradient steps:

$$\mathbb{E}[\theta_{k+1}] = \theta_k - \gamma_k \nabla \mathcal{L}(\theta_k) \tag{2.7}$$

SGD brings several advantages, including computational efficiency due to its reduced cost and the potential for avoiding convergence to suboptimal local minima through the introduction of noise (or annealing) ([92]). However, while SGD and its variants can aid in optimization, care must be taken to avoid a common problem that emerges when training NNs: overfitting. Overfitting arises when a model while trying to minimize the loss on the training data, captures not just the underlying patterns but also the noise or random fluctuations present in the data. As a result, while the model might demonstrate excellent performance on the training data, it fails to generalize well to new, unseen data. Regularization techniques, such as dropout or L2 regularization, are often employed to combat overfitting, ensuring that the model remains robust and performs consistently on new unseen data ([93]).

A balanced approach is to employ *mini-batches* instead of individual instances ([94]). Here, data subsets drive efficient and stochastic parameter updates:

$$\theta_{k+1} = \theta_k - \gamma_k \frac{1}{|B_i|} \sum_{j \in B_i} \nabla \mathcal{L}_j(\theta_k) \tag{2.8}$$

A visual representation distinguishing the convergence behaviors of these gradient methods can be found in Figure 2.3.

The standard technique for computing gradients in NNs, enabling systematic updates of model parameters and facilitating successful training of deep architectures, is based on *backpropagation* ([95]). This algorithm recursively applies the chain rule for derivatives, moving backward post a forward network pass. Each training iteration updates NN parameters in proportion to the cost function's partial derivative concerning the current parameter. The gradient for the l-th layer concerning weight θ_{ij} is:

$$\frac{\partial \mathcal{L}}{\partial \theta_{ij}^{(l)}} = \frac{\partial \mathcal{L}}{\partial f_i^{(l)}} \cdot \frac{\partial f_i^{(l)}}{\partial f_j^{(l-1)}} \cdot \frac{\partial f_j^{(l-1)}}{\partial \theta_{ij}^{(l)}} \tag{2.9}$$

Backpropagation allows to determine the gradient of neurons that are not directly connected to the outputs, allowing to train the hidden layers of a NN ([96]).

The versatility of gradient-based optimization doesn't end with the use of mini-batches. Several enhancements and adaptations to GD and SGD have been proposed to improve their convergence properties and address some of their intrinsic limitations.

For instance, variants like *Momentum* and *Adaptive Gradient Algorithms* (e.g., Adagrad, RMSprop, and Adam) incorporate past gradients to adjust the current one or adaptively tune the learning rate during training, respectively. Such strategies aim to expedite convergence and avoid local minima ([97]).

Furthermore, the choice of initialization for the NN's weights, such as Xavier or He initialization, plays an essential role in ensuring the efficacy of the gradient descent process. Improper weight initialization can either slow down the training considerably or push the model into undesirable regions of the weight space ([98]).

Regularization techniques, such as dropout, weight decay, and early stopping, have also been used to enhance generalization, particularly in deeper architectures. These mechanisms primarily serve to mitigate overfitting, ensuring the model remains robust and can generalize beyond its training data ([93]).

2.2 Supervised Learning

In this Thesis two supervised architectures will be used: the Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN). For a more detailed overview of supervised learning paradigms is available in [99].

2.2.1 Multilayer Perceptron

Multilayer Perceptrons (MLPs) are one of the oldest architectures used for NNs ([62]). The architecture of an MLP consists of multiple layers: an *input layer*, several *hidden layers*, and an *output layer*. Data is initially given in input through the (aptly named) input layer, processed in subsequent layers (the hidden layers), and then transformed into predictions in the output layer. Each hidden layer is composed to M_i neurons (the width of the layer) having as inputs all the outputs of the

Hence, when we consider of the function computed by the ℓ-th layer as

$$y_\ell = f(\phi_\ell \cdot x_{\ell-1} + b_\ell) \tag{2.10}$$

Here, y_ℓ represents the output vector of layer ℓ. ϕ_ℓ denotes the weight matrix connecting nodes from layer $\ell - 1$ to layer ℓ, while b_ℓ is the bias vector for layer ℓ.

As we can see, in MLP, like un any feed-forward network, the data traverses each layer successively until it reaches the output layer.

It is important to remember that the choice of activation function can significantly affect the MLP's performance. While traditional activation functions like the sigmoid or hyperbolic tangent were widely used, modern DL architectures often prefer the Rectified Linear Unit (ReLU) due to its ability to mitigate the vanishing gradient problem, ensuring more effective training in deeper networks ([98]).

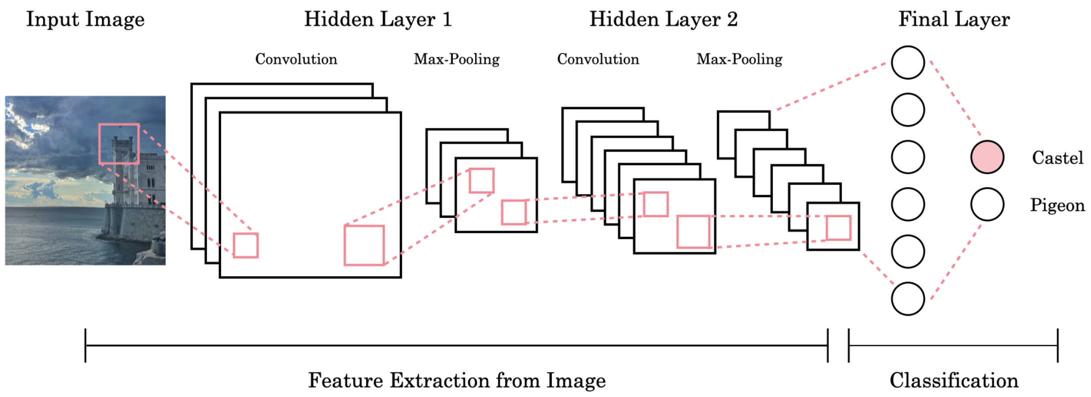


FIGURE 2.4: A CNN sequence to classify images.

2.2.2 Convolutional Neural Network

Convolutional Neural Networks (CNNs) have seen widespread use in DL, taking their name from the mathematical convolution operation between matrices ([100]). Analogous to MLPs, CNNs too comprise neurons whose weights are changed during the learning phase, and where the information flows in only one direction ([101]). CNNs are designated to work with patterns that typically appear in data with a strong spatial correlation, like visual data. Hence, CNNs can use fewer parameters than fully connected layers. ([102]).

CNNs have been used for tasks such as image classification to accurately categorize visual content ([103]), medical image analysis to aid in precise diagnostics ([104]), autonomous driving systems, enabling vehicles to perceive their environment ([105]), and many more.

In CNNs, the organization of neurons is dictated by the spatial characteristics of input data. The three integral spatial dimensions are: (i) height (M) which captures the vertical features; (ii) width (N) which identifies horizontal patterns; (iii) depth (channels), which is used for multi-channel data, such as color channels in images.

Mathematically, a convolution for a 2-D spatial domain is defined as:

$$y(m,n) = \sum_{i=1}^{M} \sum_{j=1}^{N} x(m+i, n+j) w(i,j) \tag{2.11}$$

Here, $y(m,n)$ is the output pixel's convolution from the input pixel $x(m,n)$ and a filter $w(i,j)$, considering dimensions M (height) and N (width) ([106]).

The architecture of Convolutional Neural Networks (CNNs) usually contains three types of layers, each with specific functions:

- Convolutional Layer: These layers are adept at recognizing local patterns in data. With small spatial but full-depth kernels, convolutional layers operate through scalar products between their weights and the input region. By convolving these filters across the input, 2D activation maps are generated. Key hyperparameters for tuning include:
  - Depth: This determines the number of neurons within the layer, with each neuron connecting to a specific region of the input.
  - Stride: The stride parameter dictates the number of units by which the filter or kernel shifts.
  - Zero-padding: Involves adding zeros around the border of an input image.
- Pooling Layer: The primary role of pooling layers is downsampling along the spatial dimensions of the input. This downsampling helps in reducing the number of parameters within the activation maps, making the network computationally efficient.
- Fully-Connected Layers: Similar to standard ANNs, fully connected layers aim to produce class scores based on the network's activations. These scores are crucial for classification tasks. It's worth noting that Rectified Linear Units (ReLU) activation functions are often applied between fully-connected layers, which can enhance the network's overall performance.

For further details about CNN structure and functioning the reader can refer to [101]. A representative illustration of a typical CNN architecture for image classification task is reported in Figure 2.4.

Through these transformations, CNNs sequentially process the input data using convolution and downsampling techniques, ultimately generating class scores for classification and regression tasks.

In the case of data where there is only one spatial or temporal dimension, 1D CNNs can be used. Unlike 2D CNNs, which process two-dimensional spatial input, 1D CNNs focus on sequential or time-series data. This distinction manifests in the kernels they use: while 2D CNNs employ two-dimensional kernels, 1D variants utilize one-dimensional kernels, scanning and processing input data linearly. 1-D CNNs are employed in the areas of biomedical data categorization, infrastructure health analysis, and power electronics anomaly detection. Their architecture's simplicity offers allow for real-time processing and cost-efficient hardware integrations ([65]).

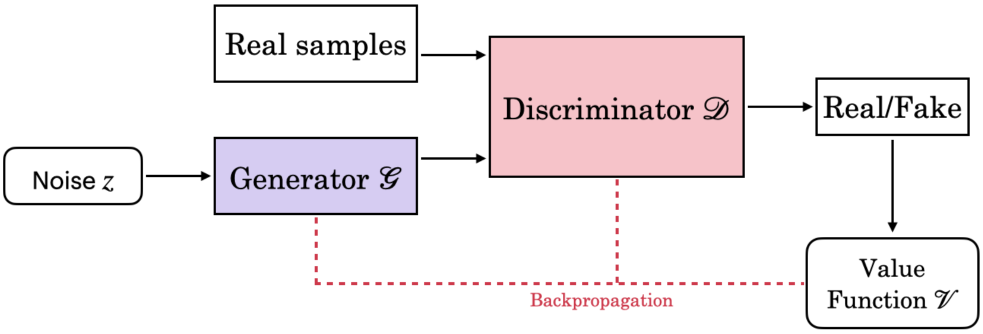


FIGURE 2.5: Diagram of a Generative Adversarial Network.

2.3 Unsupervised Learning

In unsupervised learning, the dataset consists of unlabeled examples $(x_i)_{i=1}^{N}$, where each x_i represents a set of features. Unsupervised learning is a less clearly defined branch of ML, with various algorithms used for quite different tasks. The main idea behind unsupervised learning is to find useful patterns in a dataset that lacks labeled information ([107]). Common examples of unsupervised learning include clustering ([108]) and dimensionality reduction ([109]). Another approach is generative modeling ([110]). In generative modeling, we work with training examples x drawn from an unknown distribution $p_{data}(x)$. The goal is to learn a model, $p_{model}(x)$, that comes as close as possible to approximating $p_{data}(x)$. One popular method for generative modeling is Generative Adversarial Networks ([111]), which we'll also employ in our study.

2.3.1 Generative Adversarial Nets

Generative adversarial networks (GANs) can be used to learn the distribution of the training dataset. Once this distribution is accurately modeled, GANs can extrapolate and generate new, unseen samples that align with the learned data distribution ([107]).

GANs have demonstrated remarkable versatility and impact, finding applications across a wide spectrum of fields. Notably, GANs have been used for diverse tasks such as image processing to enhance visual quality ([112]), medical diagnostics to aid in accurate disease detection ([113]), and even face detection for improved biometric security ([114]). In addition, GANs have also been applied to different domains like text transformation for language manipulation ([115]), data augmentation to enrich training datasets ([116]), anomaly detection for identifying unusual patterns ([117]), and super-resolution techniques for generating high-quality images from low-resolution inputs ([118]).

Algorithm 1 Generic Generative Adversarial Network (GAN) Algorithm

```
Hyperparameters:
Batch size $b$
Number of epochs $n_{tot}$
Number of epochs $n_{\mathcal{D}}$ for $\mathcal{D}$
Learning rate $(\lambda_{\mathcal{G}}, \lambda_{\mathcal{D}})$ for $(\mathcal{G}, \mathcal{D})$
Training loop:
for $n = 1 \dots n_{tot}$ do
    for all $n_d = 1 \dots n_{\mathcal{D}}$ do
        Sample a MB of $\{x_i\}_{i=1}^{b}$ from $X$, and $\{z_i\}_{i=1}^{b}$ from $P_z$
        Compute $\mathcal{D}$ loss $L_{\mathcal{D}}$ as
            $$\frac{1}{b}\sum_{i=1}^{b} \log(D(x_i)) + \log(1 - D(G(z_i)))$$
        Perform a GD step on $\theta_D$ to minimize $L_D$ with lr $\lambda_{\mathcal{D}}$
    end for
    Sample a MB of $\{z_i\}_{i=1}^{b}$ from $P_z$
    Compute the $\mathcal{G}$ loss $L_{\mathcal{G}}$ as
        $$\frac{1}{b}\sum_{i=1}^{b} log(1 - D(G(z_i)))$$
    Perform a GD step on $\theta_G$ to maximize $L_G$ with lr $\lambda_{\mathcal{G}}$
end for
```

A GAN's architecture is composed of two neural networks: the Generator and the Discriminator. Their symbiotic relationship is governed by adversarial learning principles, where each model continually refines its strategies in response to the other.

Considering the widespread popularity and extensive use of GANs since they were first introduced, various different versions have been developed over the years to serve specific purposes ([119]).

Some of these versions include:

- *Vanilla GAN* which is a basic type of GAN using simple MLP networks for the Discriminator and Generator.
- *Fully-Connected GAN* (FCGAN), which uses Fully Connected NNs for both the Discriminator and Generator.
- *Laplacian Pyramid GAN* (LAPGAN), an adaptation optimized for tasks in unsupervised learning scenarios ([120]).
- *Generative Recurrent Adversarial Networks* (GRAN), which introduces a method where generation happens step by step in a recurring manner ([121]).

- *Conditional GAN* (CGAN), where both the Discriminator and Generator are influenced by extra information ([122]).

2.3.1.1 GANs: a Game Between Two ML Models

The concept of GANs is based on a game between two ML models, the Generator and the Discriminator, usually implemented using NNs. While the Generator produces samples aiming for authenticity, the Discriminator evaluates these samples, distinguishing between genuine and generated data.

Algorithm 1 summarizes the step in the training procedure and Figure 2.5 shows the standard diagram for a GAN architecture.

The Generator $\mathcal{G}$ implicitly defines $p_{model}(x)$ by generating samples from the distribution. It relies on a prior distribution $p(z)$ for an input vector z, which serves as a source of randomness within the system. Its primary role is to learn a function $\mathcal{G}(z)$ that transforms this random noise z into realistic samples.

The Discriminator $\mathcal{D}$ assesses samples x and provides an estimate $\mathcal{D}(x)$ of whether the sample is real (drawn from the training distribution) or fake (created by running the generator).

Each player incurs a cost:

$$\text{Obj}_{\mathcal{G}}(\theta_{\mathcal{G}}, \theta_{\mathcal{D}}) \tag{2.12}$$

for the Generator and

$$\text{Obj}_{\mathcal{D}}(\theta_{\mathcal{G}}, \theta_{\mathcal{D}}) \tag{2.13}$$

for the Discriminator. Then, each player attempts to minimize its own cost. The Discriminator's cost encourages the Discriminator to correctly classify data as real or fake, while the Generator's cost encourages the Generator to produce samples that the discriminator incorrectly identifies as real. In order to win the game, the two participants need to continuously optimize themselves to improve the generation ability and the discrimination ability, respectively. This interplay between the Generator and Discriminator drives the GAN training process.

2.3.1.2 Training GANs

The adversarial optimization process improves the performance of $\mathcal{G}$ and $\mathcal{D}$ gradually ([123]).

We train $\mathcal{D}$ to maximize the probability of assigning the correct label to real and generated samples, and simultaneously, we train $\mathcal{G}$ to minimize:

$$\log(1 - \mathcal{D}(\mathcal{G}(z))) \tag{2.14}$$

Mathematically speaking, $\mathcal{D}$ and $\mathcal{G}$ play the following two-player min-max game with value function:

$$\min_{\theta_{\mathcal{G}}} \max_{\theta_{\mathcal{D}}} \mathcal{V}(\theta_{\mathcal{G}}, \theta_{\mathcal{D}}) \tag{2.15}$$

where the loss function $\mathcal{V}(\theta_{\mathcal{G}}, \theta_{\mathcal{D}})$ is formulated as:

$$\mathcal{V}(\theta_{\mathcal{G}}, \theta_{\mathcal{D}}) = \mathbf{E}_{x \sim p_{data}(x)}[\log \mathcal{D}(x)] + \mathbf{E}_{z \sim p_z(z)}[\log(1 - \mathcal{D}(\mathcal{G}(z)))] \tag{2.16}$$

Here, x is sampled from the real data distribution $p_{data}(x)$, z is sampled from the prior distribution $p_z(z)$, and $\mathbf{E}(\cdot)$ represents the expectation.

In practice, the training phase is implemented following an iterative process so that: when the Discriminator's weights are updated, the Generator's weights are held constant; and when the Generator's weights are updated, the Discriminator's weights are held constant.

Since the introduction of GANs in [111], numerous improvements in the GANs architecture have been developed. These variations encompass enhancements in model structures, theoretical extensions, and novel applications. For a more in-depth and comprehensive analysis, readers can refer to [123].

2.3.2 Inpainting

Image inpainting, or image completion, is a common task in computer vision, aiming to restore missing or corrupted pixels in images. This task appears in diverse scenarios spanning photo editing, computational photography, and image-based rendering, among others ([124–126]).

The main difficulty of image inpainting lies in synthesizing pixels that are both visually consistent and contextually congruent with the existing image content, ensuring the completed image appears authentic and meaningful ([127]).

Inpainting techniques are specific to the task that inpainting should solve. For image restoration, the objective might be to rectify scratches or overlaying text. In photo

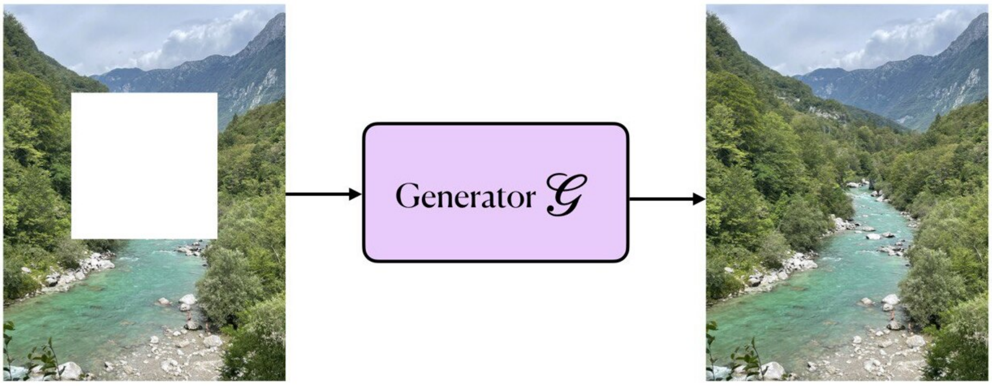


FIGURE 2.6: Visual intuition of how the Generator reconstructs input images with a hole in the inpainting technique.

editing contexts, object removal is often the primary aim. In cases concerning image coding or transmission, the focus shifts to recovering lost data segments. Moreover, in the niche of virtual painting restoration, emphasis is laid on eradicating imperfections like scratches ([128]).

Historically, diffusion-based synthesis emerged as a prevalent technique for image inpainting. This method capitalizes on propagating proximate image features into the affected zones for completion ([129]). While adequate for addressing minor imperfections, its efficacy diminishes with larger or more complex image gaps.

Alternatively, patch-based methods can address more sophisticated inpainting challenges, filling even considerable voids in images ([130, 131]). Yet, these techniques often fall short when confronted with intricate structures or when tasked with creating entirely novel image elements. Their reliance on textures extracted exclusively from the source image can sometimes limit their creative scope.

With the advent of DL, CNNs have been used for image completion tasks ([132]), and recent advancements in GANs have increased the applicability of neural models for the task of inpainting ([133–135]).

The widespread use of GANs in image inpainting arises from their ability to understand and reconstruct missing or corrupted segments of an image, ensuring the filled regions seamlessly integrate both visually and semantically with the original content ([136]).

In the context of inpainting, the Generator typically adopts an encoder-decoder structure. The encoder captures the high-level features of the intact portions of the image, while the decoder leverages this encoded information to recreate the missing or corrupted parts ([137]).

Component	Entities	Relations	Rel. inductive bias	Invariance
Fully connected	Units	All-to-all	Weak	
Convolutional	Grid elements	Local	Locality	Spatial translation
Recurrent	Timesteps	Sequential	Sequentiality	Time translation
Graph network	Nodes	Edges	Arbitrary	Node, edge permutations

TABLE 2.1: Relational inductive biases in standard DL building blocks ([12]).

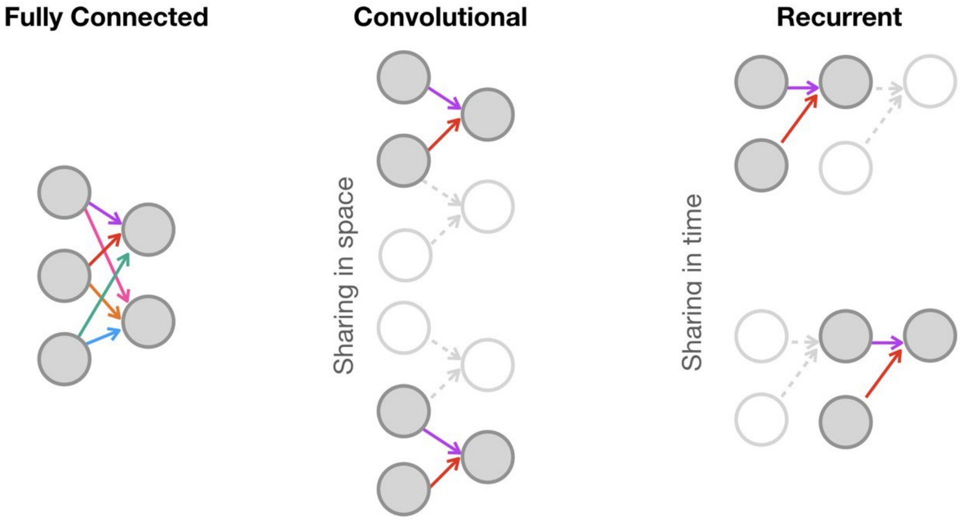


FIGURE 2.7: Visual representation of information sharing in the most common DL building blocks.

Qualitatively, GANs offer a certain amount of contextual understanding, ensuring inpainted regions are semantically integrated with the surrounding content. Their inherent adaptability allows them to manage various missing data patterns, from minor imperfections to substantial voids. Outputs are often of a quality that allows them to be used for professional and high-resolution applications ([138]).

2.4 Relational Inductive Biases

The final subject to be introduced in this thesis revolves around the concept of relational inductive biases ([12]). This concept is crucial for understanding how specific architectures are chosen based on the task at hand and the structure of the available data. In this section, will be described the relational inductive biases related to the building blocks discussed in the previous chapters. We will also briefly touch upon biases related to the other two fundamental components in the field of DL, to provide a clearer understanding of this concept. In essence, being able to select the right relational inductive biases when designing a DL architecture can greatly simplify the learning process for entities, relationships, and how they come together.

To paint a clearer picture of relational inductive biases, let's start by exploring another important idea in DL: combinatorial generalization. The main point here is breaking down complex systems into smaller parts (entities) and understanding how they work together ([139, 140]). These parts are organized like a tower, helping us ignore small differences and focus on common ways things appear and behave ([141]). This way, we can solve new problems by putting together things we already understand. This notion of fitting combinatorial generalization into AI models has been around since the very beginning of AI itself ([142]).

One of the key ways to incorporate relational reasoning into AI algorithms is by using a relational inductive bias. This bias sets specific constraints on how entities interact and form relationships during the learning process. The presence of this inductive bias enables a learning algorithm to prioritize one solution over another, regardless of the data it observes. It can be integrated into a DL process in various ways. For instance, it can take the form of a regularization term, which is added to prevent overfitting. In a Bayesian model ([143]), the bias could be embedded in the selection and setup of the prior distribution. Alternatively, it might be part of the algorithm itself.

As mentioned previously, recent years have witnessed the emergence of diverse ML architectures. These architectures adhere to a design blueprint that involves assembling fundamental building blocks to construct intricate, deep computational structures. The manner in which a relational inductive bias can be embedded via architectural choices follows a twofold trajectory: firstly, within the selection of the constituent building blocks themselves, and secondly, in the orchestration of their arrangement into layers.

To offer an intuition into the relational inductive biases inherent in diverse DL methods, it becomes essential to delineate certain components within the building blocks under scrutiny. These ingredients encompass the entities, the relationships, and the principles governing the amalgamation of said entities and relationships.

2.4.1 Relational Inductive Biases in standard Deep Learning Building Blocks

To make the idea of relational inductive biases and their integration through different DL building blocks clearer, we'll list the main building blocks and the biases they introduce during learning. These concepts are also schematized in Table 2.1.

2.4.1.1 Fully Connected Layer

The most commonly encountered building block is the fully connected layer ([85]). As we have outlined earlier, a fully connected layer can be conceived as a non-linear function with vector outputs. The output vector is generated through a combination of a weight vector and an added bias term, followed by a non-linear function. In this context, the entities correspond to the units within the network, relationships involve an all-to-all connection (each unit in layer i is linked to every unit in layer j), and the governing principles are defined by the weights and biases. The guiding principle for the rule is the complete input signal, with no reusing of information or isolation. The natural consequence is that the implicit relational inductive bias in a fully connected layer is thus very weak. All input units hold the capacity to influence the value of any output unit, with no constraints across outputs. Figure 2.7 offers a visual representation of how in the fully connected layer all weights are independent, and there is no sharing.

2.4.1.2 Convolutional layers

The next frequently encountered building block is the convolutional layer ([144, 145]). In this case, the elements making up a grid (like pixels) serve as the entities, and the relationships formed are more selective. The contrast between a fully connected layer and a convolutional layer brings about essential relational inductive biases: locality and translation invariance. Locality underscores that the inputs to the relational rule consist of entities situated closely to one another in the input signal's coordinate space, distinct from distant entities. Translation invariance, on the other hand, denotes the reuse of the same rule across different spots in the input. These biases play a significant role in handling natural image data. This is due to the fact that there's a strong correlation within local neighborhoods, which tends to decrease as we move farther away, and the statistical patterns remain mostly consistent across an image. Figure 2.7 offers a visual representation of how in the convolutional layer a local kernel function is reused multiple times across the input (Shared weights are indicated by arrows with the same color).

2.4.1.3 Recurrent Layers

A third commonly encountered foundation is the recurrent layer ([146]), commonly found in Recurrent NNs (RNN) ([147]). While we haven't delved into this in the preceding sections of this thesis, since as we are not going to use them in our research, we will outline their inherent biases to offer a clearer grasp of the concept. For more specific details about NNs, refer to [148]. A recurrent building block operates across a sequence of stages. Within this context, the inputs and hidden states at each step serve as the entities, while the relationships emerge from the dependency of the current step's hidden state on the preceding hidden state and the ongoing input. The rule governing the union of these entities employs the inputs and the hidden state of each step as parameters, with the objective of updating the hidden state. This rule remains consistent across all steps. In essence, the relational inductive bias here lies in temporal invariance—a concept analogous to a CNN's translational invariance in spatial dimensions. Figure 2.7 offers a visual representation of how the same function is reused across different processing steps in the recurrent layer (again, shared weights are indicated by arrows with the same color).

2.4.1.4 Graph Neural Networks

Graph Neural Networks (GNNs) constitute a class of DL models designed for operating on graphs ([149]). These networks structure their computations based on the graph's underlying architecture. While these networks won't be directly employed within the scope of this thesis, their mention serves the purpose of enhancing our understanding of biases. For detailed insights, readers can refer to [12]. Fundamentally, the GNN framework possesses several pronounced relational inductive biases. Graphs can capture diverse relationships among entities. Consequently, it's the GNN's input that influences how these representations interact and are demarcated, as opposed to predetermined architectural constraints. Furthermore, graphs depict entities and their relationships as sets, a quality that remains consistent irrespective of permutations. This imparts invariance to the sequence of these elements.

Chapter 3

Machine Learning for Ocean Biogeochemistry

Continuous monitoring of the ocean is crucial for a comprehensive understanding of its vast and intricate ecosystem. Over time, scientists have conducted extensive research at local and global scales to explore the ocean environment ([150]).

The ocean's significance extends beyond its vastness, playing a critical role in global environmental dynamics. As technology has evolved, so too has our ability to monitor these waters. Innovations in satellites, autonomous sensors, and buoy systems have paved the way for the collection of data at an unprecedented scale.

With such vast amounts of data available, AI (and, in particular, ML) offers a powerful tool for uncovering patterns, making predictions about the ocean's state, understanding its evolutionary trajectory, and deciphering the complex relationships between various oceanic variables.

Complementing these AI advancements, deterministic models enhance our understanding of ocean processes.

The use of ML in oceanography has not only introduced new tools but also expanded our knowledge on phenomena like mesoscale eddies, internal waves, sea ice, and marine algae. These factors are critical in determining Earth's climate, marine cycles, and ocean health ([151]).

Studies have confirmed the potential of ML in oceanography ([152–154]). However, despite this progress, the field of oceanography remains largely unexplored, particularly in restricted areas such as the Mediterranean Sea. Most of these studies are global in nature and focus on the ocean as a whole.

Oceanographic data, with its inherent non-linearity and high dimensionality, is well-suited for ML analysis. One notable advantage of ML over traditional statistics is its ability to handle complex, high-dimensional data with non-linear relationships and missing values ([155–157]). By incorporating ML, we can enhance traditional methods, leading to more significant discoveries ([158]).

The study of the ocean presents a unique set of challenges that ML can help address. Oceanographic observation datasets contain vast and intricate information that are sparse in spatial coverage, often limited to surface measurements (e.g. the measurements collected by the satellite devices), and have only a few time series spanning a handful of decades. Moreover, oceanographic phenomena operate on timescales ranging from seconds to millennia and encompass intricate physical-biological interactions across various scales. By harnessing the power of ML, it becomes possible to effectively handle these large and complex datasets, extract valuable insights from such limited time series, and understand the intricate interactions occurring across various temporal scales

Despite the inherent challenges, the integration of ML in oceanography holds the potential to revolutionize foundational theories ([47]). This integration enables better scientific understanding, larger and more detailed analyses of patterns ([48]), and improvements in environmental management, policy-making, and public engagement ([49]).

In this section, we focus on how ML techniques help us to handle the complexities of big data in biogeochemical oceanography. Firstly, in Section 3.1 we provide a clear definition of what we intend with the term big data, give an overview of the most common datasets in oceanography, and discuss the main challenges that arise when developing ML models trained on this data. Subsequently, Section 3.2 examines the principal application of ML within the domain of oceanography, dividing the discussion based on the different datasets used for the training. Finally, in Section 3.3 a class of MLP-based models is described in deeper detail, as these methods are the starting point of the first two research chapters which constitute this thesis.

3.1 Data Available in Oceanography

Historically, the study of aquatic ecosystems primarily centered on gathering information about water characteristics, encompassing both physical and chemical conditions. Later, this focus expanded to include biota and habitat assessments, taking into account environmental flows and riparian conditions as well ([159]).

In earlier approaches, data gathering involved measuring at specific sites and then combining the data for larger-scale assessments of ecosystems ([160]). The landscape of ocean observation underwent a significant transformation with the emergence of advanced data technologies. Satellite imagery and autonomous sensors, in particular, have revolutionized data acquisition. Those data, often labeled as big data, are characterized by high volume and complexity, often proving challenging for traditional analytical methods to handle ([161]).

These datasets are a result of progress in marine monitoring and offer an important opportunity to investigate environmental phenomena over larger areas and time-frames ([162]). But to fully exploit this potential, there is the need to understand the strengths and limitations of the available data sources.

This Section aims to do so, specifically: Section 3.1.1 provides an overview of the primary data sources available in oceanography; and Section 3.1.2 describes the challenges that comport modeling those oceanographic data.

3.1.1 An Overview of the Principal Data Sources

This section delves into the primary data sources commonly found in oceanography. The goal is to provide an overview of the data sources available for DL model development, emphasizing their essential characteristics.

Understanding the characteristics of these data is extremely useful, as it leads to the selection of appropriate DL architectures, i.e. an architecture that can efficiently extract maximal information from the data while minimizing computational costs during training.

As mentioned earlier, the data can be categorized into two groups: observations (further divided into in-situ observations and satellite observations) and data generated through model simulations.

Model simulations employ computer-based techniques to emulate real-world processes or systems, leveraging mathematical models. The model captures the key behaviors and characteristics of the chosen process or system, while the simulation illustrates how the model evolves under various conditions over time. Oceanographic numerical models use mathematical equations to describe the movement of fluids (water and air) across the planet's surface, biogeochemical processes, and sea ice phenomena ([163–165]).

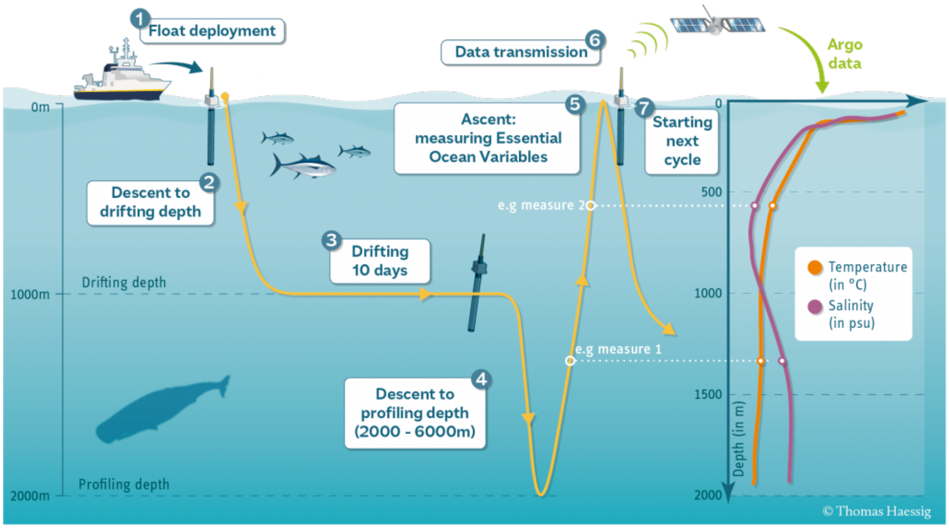


FIGURE 3.1: Argo Float's Standard 10-Day Cycle
After approximately 9 to 10 days of autonomous drift at a prescribed parking depth near 1000m, the float descends to 2000m and ascends back to the surface while conducting profiling measurements of pressure, temperature, salinity and biogeochemical variables such as nitrate, chlorophyll, oxygen and backscattering. Utilizing satellite connectivity, the float transmits collected data to a designated center for real-time processing. The processed data is swiftly made accessible online, typically within 24 hours. Image source: [166]; image author: *Thomas Haessig*

In-situ observations can be further divided into two sub-classes: data collected through ship campaigns and data collected through autonomous instruments. Ship campaigns involve collecting samples from the ocean and analyzing them using analytical methods. Autonomous instruments, such as the Argo Float and satellite, do not require human intervention during their sampling. These instruments are equipped with sensors that can measure various oceanographic parameters, such as temperature, salinity, and currents, and transmit this data to shore via satellite. In-situ observations yield high-quality data fundamental to gain insights into the ocean's physical, chemical, and biological properties.

Recent years have witnessed an important increase in the use of autonomous instruments for in-situ data collection. This is due to advances in technology that have made these instruments more reliable and cost-effective.

3.1.1.1 In-situ data

The dawn of the era of systematic ocean measurements can be traced back to the late 18th century CE. During this period, ships were primarily employed to measure physical variables, most notably temperature and salinity. Despite the emergence of

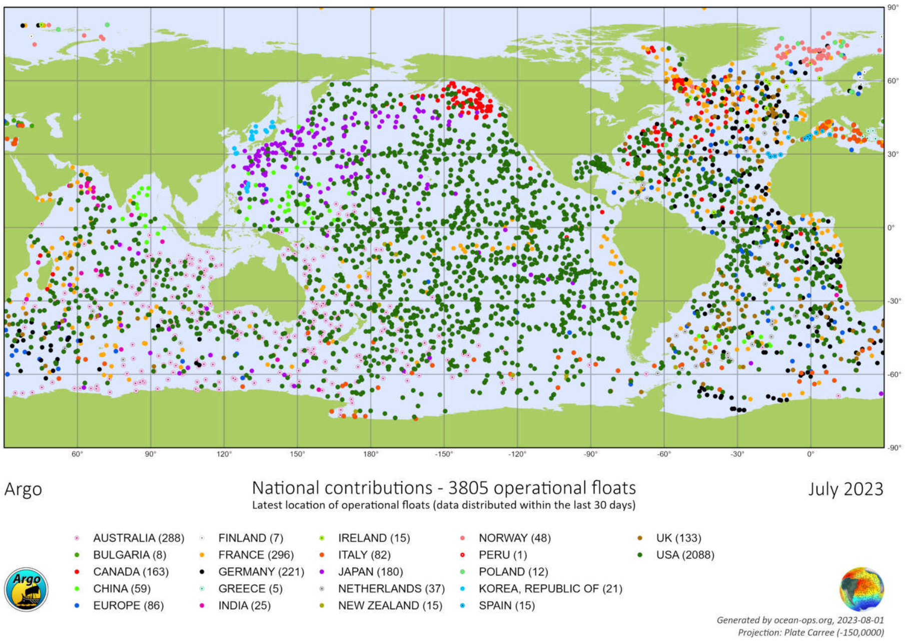


FIGURE 3.2: Location of operational floats and national contributions in July 2023. Image source: [166, 170]

theoretical tools for comprehending ocean currents in the early 19th century, oceanographic discoveries remained anchored to observations. Similarly, marine biogeochemical observations were also conducted primarily from ships, specifically water samples were collected for later laboratory analysis. These methodologies remain indispensable nowadays for their typically high accuracy.

The data gathered through these measurements have been compiled into global databases (such as the Global Ocean Data Analysis Project version 2, GLODAPv2; [167, 168]).

However, these traditional methods also exhibit certain limitations, even today: data coverage remains sparse in numerous regions, observations are scarce for the deeper layers of the ocean, the temporal resolution is low with far fewer records available for the 1970s and 1980s compared to the present time, and observations tend to be skewed towards the summer months in both hemispheres (roughly four times as many profiles in GLODAPv2 during the three summer months compared to the three winter months; [169]).

In response to the limitations inherent to traditional oceanographic methods, the last two decades have seen the emergence of advanced oceanographic instruments, notably the Argo floats, as part of the Global Ocean Observing System (GOOS). These robotic devices collect marine variable data by autonomously diving in the ocean and adjusting their depth through buoyancy changes ([32]), an example of an Argo float's standard 10-day cycle can be found in Figure 3.1.

The Argo Program, which is responsible for the management and coordination of such instruments all around the globe, holds a significant role within both the Global Ocean Observing System (GOOS, [28]) and the Global Climate Observing System (GCOS, [171]). It provides real-time data for ocean and atmospheric services, along with high-quality data for climate investigation. Initiated in 1999, the Argo Program has ensured comprehensive coverage of the upper 2000 meters of oceans globally since 2006 ([32]). The Biogeochemical (BGC)-Argo initiative started with the installation of optical ([33–35]) and oxygen ([172, 173]) sensors on profiling floats between 2000 and 2003. Each float is equipped with sensors to measure six fundamental ocean variables: chlorophyll fluorescence (Chla), particle backscatter (bbp700), oxygen (O2), nitrate, pH, and irradiance [174]. These variables are essential to comprehend crucial ocean phenomena: (i) air-sea carbon exchanges, (ii) ocean deoxygenation, oxygen minimum zones, and associated denitrification flows, (iii) ocean acidification, (iv) The biological carbon pump, (v) phytoplankton communities.

Examples regarding how the measurements provided by these instruments have been important in the oceanographic world include the characterization of ocean nitrate supply ([175]), observation of subsurface bloom dynamics ([176]), exploration of processes within oxygen minimum zones ([177]), investigation of ocean ventilation ([172]), air-sea exchange of O2 ([178]), and CO2 ([179]).

While Argo floats have revolutionized the way oceanographic data is gathered, their use is not without challenges. One of the primary issues is instrumental drift. Over extended periods, the sensors on Argo floats can exhibit drift, which can introduce minor inaccuracies in the data they collect ([180]). Additionally, despite the impressive global coverage of the Argo network, there are regions where data acquisition is sparse. Also, interruptions in satellite connectivity or equipment malfunctions can lead to occasional data loss or latency in real-time data access ([181]). In terms of durability, an average Argo float has a lifespan of about 4-5 years, after which either its battery may deplete or the device might face mechanical issues. There's the potential risk of floats becoming marine debris, or of floats inadvertently interacting with marine life or grounding in ecologically sensitive habitats, which necessitates careful planning and deployment strategies ([182]).

3.1.1.2 Satellite Data

The age of satellites for oceanographic observations began with the launch of Sputnik in 1957. This inaugural artificial satellite opened the door to observing the oceans from above the atmosphere.

In the field of oceanography, remote sensing plays an essential role. It involves estimating sea surface temperature, chlorophyll, suspended sediment, the density of colored substances, wave characteristics, and the identification of potential fishing zones (PFZ) ([36]). These sensing tools can be broadly divided into two main categories: passive and active sensors. While passive sensors solely rely on detecting energy naturally emitted or reflected from the Earth's surface, active sensors take a proactive approach by emitting their own energy source to illuminate and subsequently record reflections from the observed objects ([183]).

The ability of satellites to monitor vast oceanic expanses enables researchers to obtain comprehensive and global-scale perspectives ([184]). One of the principal benefits is the continuous temporal coverage, allowing for near-real-time monitoring and the construction of extended time-series datasets ([185]). Such extensive temporal coverage can track seasonality, interannual variability, and even decadal trends ([186]). Satellites, with their multispectral and hyperspectral sensors, can simultaneously capture information across a broad range of wavelengths ([187]). This allows for diverse applications, from mapping chlorophyll-a distributions, indicative of phytoplankton blooms ([188]), to delineating sea surface temperature patterns, which play a pivotal role in weather and climate forecasting ([189]). Furthermore, with advancements in satellite technology, the spatial resolution has improved significantly, enabling fine-scale observations crucial for coastal studies, tracking oil spills, or identifying harmful algal blooms ([190]). Lastly, in regions inaccessible or perilous for in-situ measurements, such as the polar zones or conflict zones, satellites become indispensable tools for uninterrupted data collection ([191]).

However, satellite-based oceanographic observations are not without limitations. A primary concern is that satellites primarily observe the surface or the near-surface of the ocean, providing limited insight into sub-surface processes ([192]). This surface-biased view can miss out on oceanographic phenomena like thermocline dynamics or deep-water currents ([193]). Additionally, atmospheric interference, be it clouds, rain, or aerosols, can distort or obscure satellite measurements ([194]), especially in optical and infrared bands. While certain bands, like microwaves, can penetrate cloud cover, they too face challenges in heavy rainfall conditions ([195]). Calibration and validation of satellite data remain persistent challenges, necessitating periodic ground-truthing with in-situ measurements ([196]). Lastly, satellite missions are costly, both in terms of launch and maintenance, and while the data is invaluable, it requires substantial investments ([197]).

3.1.1.3 Numerical Model Data

Ocean biogeochemical models (OBMs) are complex models designed to illustrate the ocean's dynamics, covering both physical properties and biogeochemical traits. Such models utilize a series of interconnected mathematical formulations to describe various oceanic elements like temperature, salinity, and currents influenced by external forces such as wind and internal processes like eddies and convection ([39]). Furthermore, OBMs cover the transformation mechanisms of seawater's chemical components, such as nutrients, different plankton categories, dissolved gases, non-living organic materials, and elements related to the carbon system's inorganic properties.

Through the numerical solutions to these equations, we can understand the ocean's evolution across diverse scales and resolutions. These models serve multiple purposes: they reconstruct past ocean conditions, provide insights into its current state, and predict possible future scenarios.

The heart of an OBM is the interaction between biogeochemical elements, which is mathematically represented by interconnected equations. These equations capture the rate of changes in variables that stand for nutrient concentrations, the biomass of various plankton groups, and more. The standard format of these equations is:

$$\frac{\partial C}{\partial t} = -\nu \cdot \nabla|C| + \nabla_H(K_H \nabla_H C) + \frac{\partial}{\partial z}(K_V \frac{\partial C}{\partial z}) + w_{sink}\frac{\partial C}{\partial z} + R_{bio}(T, \text{light}, \rho, C) \quad (3.1)$$

In this equation, $\frac{\partial C}{\partial t}$ represents the rate of change of a particular biogeochemical property (denoted as C) over time (t). The combined impact of ν and $\nabla|C|$ delineates the advection, which factors in the influence of ocean currents and the inherent gradients in C. The term $\nabla_H(K_H \nabla_H C)$ gives an account of horizontal diffusion, where K_H stands as the symbol for horizontal diffusivity. Vertical diffusion is represented by $\frac{\partial}{\partial z}(K_V \frac{\partial C}{\partial z})$, with K_V defining the vertical diffusivity. The component $w_{sink}\frac{\partial C}{\partial z}$ pertains to the movement of particles within the water column. Finally, the term $R_{bio}(T, \text{light}, \rho, C)$ encompasses the localized biological reactions impacting C, with influences stemming from variables like temperature, lighting conditions, density, and the concentration of C.

While the physical dynamics of the ocean can be frequently modeled using established equations, like the Navier-Stokes, the biogeochemical component, notably R_{bio}, relies on empirically derived relationships. These are based on laboratory findings, biological theories, and prevalent ecological patterns. Many marine OBMs utilize the NPZD framework as a touchstone, which encapsulates the interplay between nutrients, phytoplankton, zooplankton, and detritus ([198]).

Numerical models have considerably advanced our understanding of marine systems and processes ([199]). They offer a robust tool to study scenarios that might be inaccessible or too costly to observe directly, such as deep oceanic currents or long-term climatic changes. By permitting researchers to simulate different conditions, numerical models aid in predicting potential future states of the ocean, assisting in the crafting of mitigation and adaptation strategies. These models also offer a platform to assimilate diverse observational data. As a more cost-effective alternative compared to extensive field observations, numerical models facilitate studies over large spatial domains, from regional to global extents, and temporal scales ranging from mere hours to millennia ([199]).

However, the effectiveness of numerical models comes with certain constraints. A primary challenge is their dependency on the accuracy and completeness of input data. Inaccurate or partial data can result in misguided predictions and interpretations ([200]). Models, being a representation of reality, inevitably carry assumptions; they might leave out or inadequately portray some processes, leading to potential discrepancies between model outputs and real-world observations. The resolution of a model, both in spatial and temporal terms, can limit its capability to accurately encapsulate small-scale or swift processes. Moreover, enhancing the intricacy of a model, though allowing for a more detailed depiction of the system, also increases computational demands ([200]).

3.1.2 Challenges of Oceanographic Data

The vast realm of oceanography relies on an extensive variety of data types, each presenting its unique set of characteristics and challenges.

This data originates either from numerical models or direct observations of the ocean. Numerical models are valuable as they offer detailed spatial and temporal insights across the entire three-dimensional domain of oceans. Observational data, while limited in spatial and temporal dimensions, plays a pivotal role, providing a ground-truth perspective indispensable for validating the predictions and assumptions of numerical models.

Observations also exhibit different characteristics. Data collected by Argo float yield vertical profiles, offering variable values at different depths. Satellite measurements, while complementary, capture only the ocean's surface (which may be obscured by cloud cover). Ship-based measurements, which are considered the most reliable among the class of observational data, are the most difficult and expensive to collect, thus providing a more scarce picture of the marine framework, struggling to provide a comprehensive description of vast oceanic regions.

These structural disparities in data require the implementation of different NN architectures to fully exploit their potential.

For instance, data from ship campaigns are really sparse and scarce in both the horizontal and vertical dimensions. Thus, each sample can be handled as a point in a 3D space, which stores information related to different physical and biogeochemical variables, and an MLP results in a good choice to model them, taking scalar inputs, and generating scalar outputs. Conversely, consider satellite data portraying the Mediterranean Sea's surface. Here, a CNN proves more fitting, as it retains local relationships between neighboring measurements. This CNN structure harnesses spatial bias to improve prediction quality compared to a standard MLP.

To offer a comprehensive understanding of the data at hand and be able to select an optimal DL architecture, it is crucial to outline the primary challenges posed by big data in the realm of oceanography. While acknowledging the remarkable potential of big data as a tool to train robust ML architectures, it's equally crucial to recognize their limitations.

A primary challenge (which is not limited to the oceanographic field, but is a typical problem when dealing with big data) revolves around data management problems, such as collecting, storing, analyzing, and preserving large volumes of diverse data types ([47, 201, 202]).

While big data management challenges are prevalent across domains, oceanography introduces its own specific challenges.

Some of these have been mentioned in previous discussions. However, for the sake of completeness, a full list of oceanographic data limitations is provided below:

- *Data variety*: oceanography involves dealing with diverse data types, including information derived from various measurement methods (instrument-specific, continuous, discrete, qualitative). Different instruments or methods might produce data with varying levels of quality, precision, and accuracy. This can make it challenging to combine or compare datasets from different sources ([203]).
- *Data Integration*: often, oceanographic data is collected by different organizations, countries, or research teams, each with their own data standards and formats.

Integrating these diverse datasets can be challenging but is essential for comprehensive studies.

- *Scales and resolutions*: different tools or methods might collect data at different spatial resolutions. For instance, while satellites can provide vast coverage, they might not offer the same level of detail or resolution as in-situ measurements ([204]).
- *Increasing data dimensions*: data are continually growing in multiple dimensions, encompassing spatial (horizontal and vertical) and temporal aspects, with varying resolutions across these dimensions, thus requiring significant computational effort in order to handle them.
- *Incomplete or Sparse Data*: in some regions or for some parameters, data might be sparse or entirely lacking. This can make it challenging to draw comprehensive conclusions about those areas or parameters.

3.2 Machine Learning in Oceanography

The previous chapter introduced the primary data sources utilized in oceanography. This chapter provides a review of the latest developments in applying ML to oceanographic studies. Given the varying precision and coverage of marine basin data, the architectures and methods employed to handle these datasets vary. Consequently, the information extracted is influenced by the specific class of data under consideration.

Therefore, this literature review is organized by data class. For each class, we present the relevant ML methods proposed in literature through the years, aiming to illustrate how ML and these datasets combine to improve our understanding of the marine environment.

3.2.1 ML and Ocean Observations

As highlighted earlier, oceanography primarily depends on two main types of observational data: satellite-acquired and in-situ measurements.

3.2.1.1 ML and in-situ Observations

The integration of ML with ocean in-situ observations has seen consistent growth.

A 2014 study by Landschützer et al. [205] introduced a novel estimation of the global oceanic carbon dioxide (CO2) sink through a feed-forward NN based mapping of surface ocean CO2 readings.

In Giglio et al. [206], authors developed an Argo-based method that employs random forest regression, a tree-based supervised ML approach, to estimate oxygen levels (O2) using temperature, salinity, location, and time data.

In a study by Bushinsky et al. [207], a support vector machine (SVM) was utilized to measure the carbon dioxide (CO2) flux in the Southern Ocean. This assessment incorporated precise CO2 measurements from ships, combined with data from floating devices.

Watson et al. [208], used a gradient boosting machine (GBM), an ML model, to evaluate the uncertainties associated with historical calculations of ocean-atmosphere CO2 fluxes.

In-situ data plays a crucial role in classification tasks. Current classification endeavors, anchored in in-situ data, are in line with time-tested practices in physical oceanography. Although conventional classification techniques are still pertinent, unsupervised ML emerges as a potent tool for discerning patterns in oceanographic data.

In a study by Maze et al. [209], unsupervised Gaussian mixture modeling was employed to segment Argo sea surface temperature and height anomalies, delineating them into specific dynamical categories.

A 2019 study by Jones et al. [210] utilized unsupervised classification with Gaussian mixture modeling on Southern Ocean Argo float temperature profiles, uncovering distinct circumpolar groupings.

In 2020, Houghton and Wilson [211] delved into automated anomaly detection in climate patterns using DL autoencoder to harness the power of Argo measurements.

In a 2020 study by Rosso et al. [212], k-means clustering, an unsupervised learning method, was applied to data from Argo floats, encompassing biogeochemical metrics, to analyze spatial fluctuations in the physical and biogeochemical characteristics of both intermediate and deep waters.

On another front, in-situ measurements are adept at leveraging ML methods for tasks like dimensionality reduction and noise filtering. A classic illustration is the use of Principal Component Analysis (PCA). Its pioneering application in oceanography can be traced back to a study by Lorenz [213], wherein PCA was employed to forecast oceanic pressure patterns.

Another application of dimensionality reduction is highlighted in a study by Sonnewald et al. [214]. They employed an unsupervised NN, specifically a self-organizing map (SOM), to identify global marine ecological provinces (termed 'eco-provinces') using data on plankton community configurations and nutrient fluxes.

One of the standout applications of ML in this study is the work introduced by Sauzède et al. [215], which introduces the CANYON model, based on an MLP architecture. This foundational approach has subsequently undergone re-adaptation and enhancement in two distinct studies ([169, 216]) where substantial refinements were introduced.

These CANYON models form the base for the initial two research chapters of this thesis, providing key benchmarks for comparative analysis. Section 4.4 will offer a deeper exploration of this methodology. Before diving into these details, the next part will continue analyzing and explaining the ML methods used for other types of oceanographic data.

3.2.1.2 ML and Satellite

The advent of satellite observational technology inaugurated an epoch characterized by the proliferation of global-scale data. This exponential data augmentation necessitated advanced methodologies within the research domain to efficiently harness and analyze such a vast amount of information.

The application of ML to enhance the accuracy of satellite-derived products traces back to 1993 ([217]). These initial advancements were primarily motivated by the increased availability of data.

The applications of ML to satellite data are diverse. This review will highlight a few notable examples to shed light on the nature of research at the intersection of ML and satellite observations.

Mustapha et al. [218] proposed an NN-based clustering approach for automatically classifying anomalies in water leaving radiance spectra captured by ocean color remote sensing.

Later, Chapman and Charantonis [219] introduced a methodology using self-organizing maps (SOM), a type of unsupervised NN, to infer deep oceanic currents from surface satellite readings.

Denvil-Sommer et al. [220] presented a feed-forward NN approach, trained on satellite data, to estimate global oceanic partial pressures.

Martinez et al. [221], on the other hand, turned to Support Vector Regression (SVR), to reconstruct global chlorophyll patterns based on select oceanic and atmospheric parameters.

For processing satellite data, ML has proven to be a valuable tool for extracting geophysical information from remotely sensed data.

In an example from Castellani [222], researchers developed an artificial NN algorithm to automatically detect Mediterranean water eddies from sea surface temperature maps of the Atlantic Ocean.

In another study by Duncan et al. [223], the distinctiveness of raindrop size distributions over global oceans was investigated, examining its impact on retrievals concerning prior constraints and radiative transfer modeling.

These applications predominantly rely on instantaneous or short-term relationships and do not directly tackle the challenge of harnessing these products to enhance our understanding and prediction of the oceanic system.

Addressing missing data through satellite-based training features is another avenue explored within various ML techniques.

The Kriging method developed by Le Traon et al. [224], involves considering observations from multiple satellites with differing spatiotemporal sampling to fill gaps and estimate values for unobserved locations through linear combinations of available data.

Another successful approach is the *DINEOF* (Data Interpolating Empirical Orthogonal Functions) algorithm ([225]), which leverages a truncated basis of empirical orthogonal functions to reconstruct missing data while reducing noise and errors in geophysical datasets.

Recent advancement in this realm is offered by *DINCAE* (Data INterpolating Convolutional Auto-Encoder), proposed by Barth et al. [226]. This technique employs a CNN auto-encoder designed specifically for reconstructing missing data based on available, cloud-free pixels in satellite images.

3.2.2 ML and Numerical Modeling

The incorporation of ML has brought about significant changes in how ocean theory is represented within modeling. Even with more data available, numerical models are essential to bridge data gaps.

In building numerical models, different ML techniques can be used to tackle various challenges. These techniques can refine model detail and even replicate existing models, often with less computing power.

Among these methods, our focus lies on those that connect numerical models with available observations. The key point is that observations and theory hold more significance when combined.

These methods offer complementary strengths and weaknesses: observations have limitations due to sampling rates, whereas ocean models are constrained by finite resolution and physical coefficients. When a model covers a broad spectrum of simulations in time and space, observations can shed light on local phenomena beyond the model's reach. Despite their precision, observations alone can't cover the entire marine basin; thus, a numerical and theoretical model baseline is needed. Thus, methods are essential to extract, extrapolate, or upscale data, accounting for unresolved processes.

In Bolton and Zanna [227], ML predicts unresolved turbulent processes and subsurface flow fields by leveraging observations and model data. Specifically, CNNs are trained on degraded data from a high-resolution quasi-geostrophic ocean model.

In Zanna and Bolton [228], Relevance Vector Machines and CNNs are employed to create computationally efficient parameterizations for ocean mesoscale eddies.

By grasping the discrepancies between observational data and model states, ML could revolutionize data assimilation efficiency, ensuring model alignment with real-world data as proposed by [229].

ML tools can also represent the ocean with fewer degrees of freedom than full numerical models.

As an example, Agarwal et al. [230] presents a comprehensive inter-comparison of linear regression, stochastic, and deep-learning approaches for reduced-order statistical emulation of ocean circulation.

Additionally, ML is explored for emulating computationally costly aspects of ocean models.

An example relies on Nowack et al. [231], where authors suggest a novel approach using linear ML regression to predict ozone distributions based solely on atmospheric temperature fields.

Another significant aspect of ML ocean modeling deserves attention. In an ideal scenario with perfect data and adequate computational power, it's theoretically possible to learn ocean dynamics without knowledge of the equations of motion. Similar successes have been achieved in other fields, such as atmospheric applications ([232, 233]), and even with an idealized ocean model ([234]). In the latter case, the authors developed a simple regression model to understand if data-driven models can grasp the intricate underlying dynamics of the system. However, representing the ocean through DL is more intricate compared to the atmosphere. This is due to the scarcity of reliable three-dimensional ocean training data and the ocean's longer time scales, coupled with shorter time scales that collectively constitute its state ([158]).

3.2.3 Physical informed ML

Models are indispensable tools in oceanography, much like observations. This naturally leads us to explore how to effectively combine models with observations; constraining simulations using observations involves a process known as Data Assimilation (DA).

DA refines a theoretical representation of a system, often employing a numerical or statistical model, through a collection of observations. This process generally corrects initial conditions, boundary conditions, and model parameters to minimize discrepancies between model predictions and actual data ([214]).

Initial oceanographic DA experiments were localized ([235]), focusing on areas like the Gulf Stream meander and ring region.

However, large-scale initiatives, such as the World Ocean Circulation Experiment (WOCE, [236]) and the Global Ocean Data Assimilation Experiment (GODAE, [237]), underscored the potential of global ocean forecasting.

Before the advent of ML, DA methodologies predominantly utilized optimal interpolation techniques ([238]). However, in recent years, several studies have underscored the connection between DA and ML.

In Abarbanel et al. [239], authors establish an equivalence between ML and statistical DA, relating layer numbers in NNs to time in DA.

In Bocquet et al. [240], a model's dynamics are learned from observations, and an ordinary differential equation representation is inferred through recursive nonlinear regression.

In Geer [241], the equivalence between DA and ML is explored within a Bayesian network framework, examining how earth system models could be learned directly from observation.

A data-driven model can emulate a numerical model partially or entirely to provide forecasts. The objective then is to correct model errors or reduce computational costs. For instance, Lguensat et al. [242] presents a framework reconstructing system dynamics fully in a data-driven manner, assuming a catalog of system trajectories is available.

Furthermore, DA can extend parameterization learning to improve models directly from observations.

In Bonavita and Laloyaux [243], ANN models are used within a weak-constraint framework to enhance model error correction.

In Brajard et al. [229], ML-based parameterization is trained using direct data in a realistic scenario with noisy and sparse observations.

ML can also replace traditional methods in strongly coupled DA, which corrects coupled systems (e.g., ocean-atmosphere) comprehensively, though it's challenging due to various temporal and spatial scales ([244]).

Examples of this approach include the work of Amendola et al. [245], where a CNN is combined with an Optimal Interpolated Kalman Filter for Latent Assimilation.

Another example is the research carried out by Fablet et al. [246], where authors use automatic differentiation within a DL framework, enabling them to jointly train two key components: representing dynamic processes and solving DA challenges. By combining supervised and unsupervised strategies, this approach highlights the power of using DL to refine the understanding of complex systems in data assimilation.

In another direction, Mack et al. [247] propose a "Bi-Reduced Space" concept for addressing the complexities of 3D Variational Data Assimilation, by means of Convolutional Autoencoders. However, these approaches haven't yet been applied to realistic ocean DA scenarios.

3.3 ML for the Prediction of Low Sampled Biogeochemical Variables

Among all the applications of ML to oceanography, a specific class needs to be further investigated in this review, as it consists of the starting point for the studies developed in this thesis.

Notably, while many DL applications in oceanography focus on modeling physical variables, there's a pressing need to address the scarcity of observations related to biogeochemistry.

		NO_3^-	PO_4^{3-}	$Si(OH)_4)$	A_T	C_T	pH
CANYON	MAE	1.37	0.076	2.28	21.53	13.64	0.0181
	RMSE	1.81	0.107	2.98	32.74	19.88	0.0254
CANYON-b	MAE	0.95	0.049	0.97	11.18	12.34	0.0140
	RMSE	1.34	0.075	1.30	20.07	17.85	0.0204
CANYON-MED	MAE	0.47	0.026	0.43	6.51	6.98	1.0102
	RMSE	0.74	0.045	0.66	11.09	10.03	0.0158

TABLE 3.1: Canyon comparison of performances.
Performance indicators on the validation dataset of CANYON ([215]), CANYON-B ([169]), CANYON-MED ([216]). Results are taken from [216].

Among the various physical and biogeochemical variables collected, an important division emerges, specifically marine variable can be broadly divided into low-sampled and high-sampled variables. Although it is not a specific definition, in the following we will try to provide an intuition and a categorization of marine biogeochemical variables that we will use in this Thesis into these two groups.

3.3.1 High-frequency vs. Low-sampled Variables

In the realm of oceanographic data collection, a distinction emerges between the variables based on their sampling frequency.

Specifically, high-frequency variables refer to those that are sampled with greater regularity. These often encompass parameters that are relatively easier to measure or are of immediate significance, such as the date and geographical location of the sample. Moreover, certain physical variables like temperature, salinity, and oxygen also fall under this category due to their ubiquity and the ease of instrumentation designed to measure them. Their frequent sampling provides a dense matrix of information, offering a comprehensive spatial and temporal view of the marine environment.

Conversely, low-sampled variables denote those subjected to infrequent sampling. This category predominantly includes biogeochemical variables such as nitrate, silicate, alkalinity, chlorophyll, and ammonium. The infrequency often stems from the complexity of the measurements, the specialized equipment required, or the specific conditions needed for accurate sampling. Despite their sporadic sampling, these variables hold immense significance. They offer insights into the intricate biogeochemical processes of the ocean, which are paramount for understanding marine ecosystems, nutrient cycles, and the overall health of the oceans.

The challenge, therefore, lies in accurately predicting or estimating these low-sampled variables using the more readily available high-frequency ones, a task that the methodologies in this thesis aim to address.

3.3.2 The CANYON Method

The distinction between high-frequency and low-sampled variables is essential in order to understand the rationale behind the choice of input and output features in NN architectures.

In datasets originating from cruise measurements, such as the one used for the training of the architectures described in this chapter, a significant portion of the samples predominantly encompasses information on high-frequency variables. This frequency disparity isn't confined to ship-based observations. It's also evident in other observational systems, like the Argo float, a subject delved into in Chapter 5.

This observation underscores the importance of developing architectures that predict low-sampled variables using high-frequency ones.

In order to perform this task, the CANYON method was conceived and iteratively refined. At its core, this method leverages a feedforward NN to predict low-sampled variables starting from high-sampled ones.

3.3.2.1 CANYON

The work by Sauzède et al. [215] introduced the CArbonate system and Nutrients concentration from hYdrological properties and Oxygen using a Neural-network (CANYON) method, which estimates water-column variables, such as nitrate (NO_3^-), phosphate (PO_4^{3-}), silicate ($Si(OH)_4$), total alkalinity (A_T), dissolved inorganic carbon (C_T), pH, and partial pressure of CO2 (pCO_2) in the Global Ocean. This estimation is based on concurrent in situ measurements of temperature, salinity, hydrostatic pressure, and oxygen ($O2$), along with sampling latitude, longitude, and date. The CANYON method presents a promising approach to leverage forthcoming accurate O_2 measurements from profiling floats, enabling the derivation of properties that are challenging or costly to acquire. It offers an efficient way to fill spatial and temporal gaps in the fields of these biogeochemical variables by taking advantage of the simultaneous release of the GLODAPv2 database ([168]) and the planning of the BGC-Argo program ([248]).

3.3.2.2 CANYON-B

The subsequent work by [169] further advanced the CANYON model by introducing CANYON-B. This improved version, still based on the global GLODAPv2 dataset, employs more robust NNs, known as Bayesian NNs ([249, 250]). Unlike conventional

NNs, Bayesian NNs incorporate probability distributions for weights, regularization parameters, input, and output variables. Although this enhances model capability, it does come at a higher computational cost. In contrast to traditional data interpolation methods based on temporal or spatial interpolation (e.g., for climatologies, [251]), CANYON-B takes a variable interrelation view, establishing mappings between easily measured and accurate variables, like temperature, salinity, oxygen, pressure, location, and time, to the biogeochemical variables of interest. Furthermore, CANYON-B's performance was enhanced through modifications in the training procedure, including input feature preprocessing, network topology, and training approach. CANYON-B was rigorously validated against independent datasets, including bottle data from a GLODAPv2 subset and recent GO-SHIP cruises, as well as Argo profiles of sensor data for pCO2 and pH. The results demonstrated the superiority of CANYON-B in terms of various metrics used to assess the validity of the study.

3.3.2.3 CANYON-MED

Extending the regionalization concept, [216] developed CANYON-MED, a continuation of CANYON-B with a focus on the Mediterranean Sea. Recognizing the unique properties and events in this marine area (already discussed in Section 1.1), the authors limited the training of the model to the Mediterranean region. This approach is grounded on the understanding that ML methods yield better results when trained on datasets representative of the specific case study. To create CANYON-MED, a new quality-controlled dataset of in situ measurements was assembled specifically for the Mediterranean Sea, and NN ensembles were trained. The regional downscaling of the CANYON method resulted in improved NN performance, as anticipated. The Mediterranean Sea, characterized by its complex hydrography and dynamic biogeochemical processes, poses unique challenges for accurate modeling. CANYON-MED represents a significant advancement in regional biogeochemical modeling, providing tailored estimations that account for the Mediterranean Sea's distinct characteristics. This regionalization approach not only enhances model accuracy but also enriches our understanding of biogeochemical processes specific to this marine area.

To quantify and compare the performances of the described methods, the mean absolute error (MAE) and root mean squared error (RMSE) are presented in Table 3.1. This comparison highlights the advancements made in each iteration of the CANYON model, culminating in superior accuracy and applicability for specific oceanic regions.

Chapter 4

MLP for the Prediction of Biogeochemical Variables

Ocean observations play a crucial role in understanding the marine ecosystem and its sustainable resource utilization. However, these observations have some limitations. They are sparse, limited in time and space coverage, and collected unevenly across different variables ([60]).

Traditionally, marine variable measurements were conducted through specific cruises, where water samples were gathered and analyzed in laboratories. Although this method remains the most accurate and reliable, it is costly and suffers from temporal and spatial under-sampling issues ([61]). Consequently, our ability to quantitatively describe key processes in oceanic cycles and ecosystem changes is hindered.

Research Question

The research presented in this Chapter aims to address the following key questions:

- Can NN enable high-quality predictions of low-frequency marine biogeochemical variables based on data sampled at high frequencies?
- How can these predictions contribute to marine analysis, particularly in the context of Data Assimilation (DA) applications?

To effectively address these questions, our study will evaluate the performance of a proposed DL model by focusing on the following aspects:

- The quality of the predicted variables, is assessed using various metrics such as Root Mean Squared Error (RMSE).
- The influence of sample locations within the marine area on the model's error rates.
- The enhancement of the DA framework through the incorporation of low-frequency variables generated by our method.

The concepts of high and low-frequency sampled variables have already been introduced in Chapter 3.3.

We will develop an MLP model that predicts nitrate, phosphate, silicate, alkalinity, chlorophyll, and ammonium, using as input the sampling time, geolocation, temperature, salinity, and oxygen.

We train the model using a large in-situ data collection from cruise campaigns.

Choosing the MLP architecture was a natural decision, given that our data involves punctual input-output pairs. As such, there was no need to employ a more complex NN capable of capturing more complex structures among the input features. The simplicity and effectiveness of the MLP were well-suited to handle the specific nature of the data.

Related Works

Previous research has explored approximating nutrient concentration and the carbonate system using NNs (first in the work by Sauzède et al. [215], then in Bittig et al. [169]), but an application restricted to the Mediterranean basin was investigated only by Fourrier et al. [216]. The reader can find a more detailed description of these previous techniques in Section 3.3.

Previous results confirmed that restricting the geographical area of application leads to an improvement in predictive performance. Even if the amount of data for the training decreases, it allows for a better representation of variable relationships that characterize the peculiar biogeochemical and physical features of confined areas, such as the Mediterranean Sea. Again, the reader can refer to Section 1.1 where we discussed more in detail this concept.

Our approach differs from previous studies in several ways. Firstly, we use a regional dataset for training, unlike other studies ([215] and [169]). Additionally, we adopt a deterministic architecture instead of a Bayesian approach (unlike [169] and [216]), as it demonstrated better performance and computational efficiency during our preliminary investigation.

Contribution

We used and test novel elements such as the use of a larger in-situ dataset (*EMODnet*) for training and validation ([63]). This dataset is richer with respect to the datasets exploited in previous applications both in terms of quantity of samples and contained variables. To enhance the quality of the dataset, we introduce a novel two-step training procedure, utilizing a DL framework to remove potential incorrect data automatically. This approach semi-automates the quality check process, which is typically performed manually by oceanographic experts. Given that the EMODnet dataset comprises multiple datasets from different providers, inconsistencies can arise due to various measurement techniques and standards, transcriptions, and communication. Even if Quality Check procedures for data collection exist [63], the process of merging multiple sources is not free from generating inconsistencies among data.

The DL model developed, together with the two-step quality check routine introduced, lead to a reduction of both the fitness measures used to test the validity of the model.

Finally, we introduce a confidence interval for predictions using the concept of deep ensembles of NNs ([252]). This quantification of uncertainty is crucial when dealing with values collected over large and heterogeneous areas. Our objective is to not only provide a more accurate prediction tool for low-sampled variables but also comprehensively analyze the model's performance, including confidence intervals, prediction quality at different geolocations and depths, etc.

This chapter also proposed a practical application of the proposed model for a Data Assimilation (DA) task, to provide an example of how this reconstruction can be useful in the oceanographic field.

The Argo program is a clear example of successful international collaboration as it demonstrates the collective capacity of countries and human resources to provide global data coverage ([253]). An application of this data is the fusion of in-situ observations and numerical models within the BGC-Argo framework, which holds significant promise ([254, 255]). DA has been instrumental in advancing ocean prediction over several decades ([241]). DA consists of incorporating a wide range of observational datasets and their associated uncertainties into prediction models, addressing challenges related to the uneven distribution and scarcity of observations ([256]). The variational assimilation scheme, 3DVarBio ([257]), implemented by the Copernicus Marine

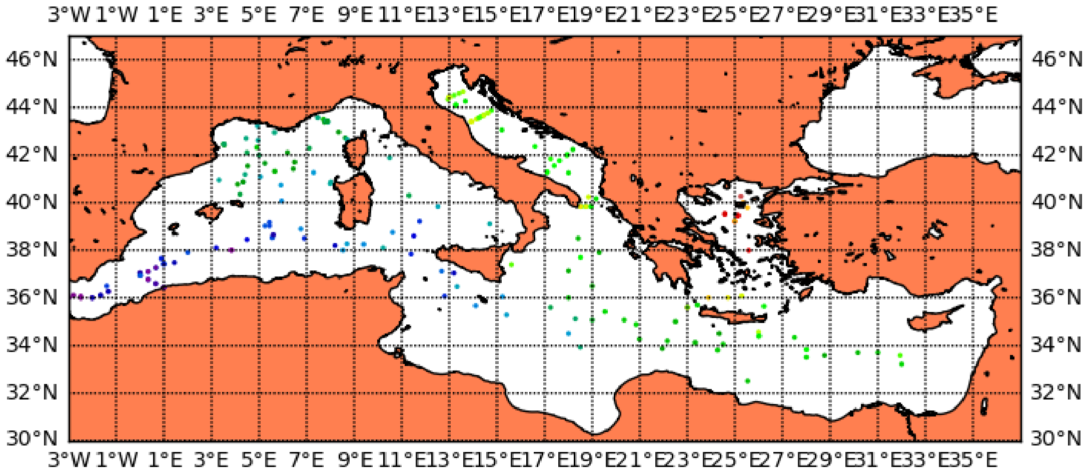


FIGURE 4.1: Alkalinity samples in the EMODnet dataset.
Solely samples in the surface are displayed, specifically the ones collected in a range between $0m$ and $20m$.

Service for the Mediterranean Sea (MedBFM), has evolved over time to encompass a more extensive array of observation types and variables. This evolution began with its initial release ([257]) and has progressively integrated coastal ocean color (OC) observations ([258]), as well as chlorophyll and nitrate profiles from BGC-Argo ([254, 255]). With the increasing availability of oxygen (O2) data from BGC-Argo, we have undertaken an additional upgrade of the MedBFM, exploring the integration of NN reconstructed profiles into the assimilation scheme.

Structure of the Chapter

The chapter is structured as follows: Section 4.1 provides details about the dataset used for training and testing the model. Section 4.2 describes the DL architecture and its implementation details are listed in Section 6.2. In Section 4.2.2, we introduce the two-step quality check routine to identify and remove cruises with anomalous data. Section 4.2.3 discusses the method for estimating prediction uncertainty. We present experimental results and their analysis in Section 4.3, also investigating the quality of the model prediction over argo nitrate profiles. Section 4.4 analyzes the result obtained by inserting nitrate profiles reconstructed using the proposed MLP into the assimilation scheme. Finally, in Section 4.5, we draw conclusions from the study.

4.1 The EMODnet Dataset

The EMODnet (European Marine Observation and Data Network) is a comprehensive marine data initiative initiated by DG MARE in 2009 ([259]). Its primary objective is to provide easy access to marine data, ensure interoperability, and remove restrictions on data usage. The EMODnet Chemistry portal specifically focuses on marine data up to the year 2018, gathered from research cruises and monitoring activities in European marine waters and global oceans. Each cruise or monitoring activity represents a subset of marine measurements obtained from specific locations or temporal periods, often with distinct sampling and analytical methodologies. Standard Quality Check procedures are applied to harmonize and validate the dataset, ensuring its reliability.

The selected Mediterranean Sea EMODnet dataset comprises a substantial collection of $101,526$ samples contributed by 74 data providers across 18 countries. It offers broad geographical coverage of the entire Mediterranean region, with longitude ranging from -5.92W to 36.19E and latitude from 31.19N to 45.77N. This extensive coverage ensures the inclusion of diverse marine environments within the Mediterranean Sea.

The spatial distribution of profile data presents a good data coverage in the Mediterranean basin. More specifically, in northern Mediterranean there are coastal areas such as Italy, Spain, France and Turkey where observations are more concentrated. Unlike southern Mediterranean where some areas are characterized by few and sparse data, like the coastal areas of Tunisia and Libya.

Each sampling entry in the dataset contains crucial information, including the date of data collection, geolocation (longitude and latitude), as well as measurements of fundamental parameters such as temperature, salinity, and oxygen. Furthermore, when available, the dataset may contain valuable information on macronutrients such as Nitrates (NO^{3-}), Phosphates ($PO4^{3-}$), and Silicates ($Si(OH)4$), essential for understanding nutrient dynamics in the marine ecosystem. Additionally, the a part of the samples of the dataset includes other biogeochemical variables such as Total Alkalinity (A_T) and Chlorophyll-a, a key indicator of phytoplankton biomass.

4.2 The Deep Learning architecture

The selected model architecture for this study is an MLP–the general characteristics of this class of NN are discussed in Section 2.2.1.

Specifically, we consider a two-hidden layer MLP with $tanh(x)$ as the chosen nonlinear activation function. We opted for an MLP with just two hidden layers, as our initial

Topology	1	2	3	4	5	6	7	8	9	10
# neurons in hidden layer 1	31	50	40	15	27	25	19	41	33	22
# neurons in hidden layer 2	23	30	20	8	23	25	11	9	29	15

TABLE 4.1: MLP topologies selected.

investigation into optimizing the architecture's hyperparameters indicated that adding more than two layers did not yield any performance improvement. Additionally, we introduce a Scaled Exponential Linear Unit function (SELU) after the output layer. The SELU function is defined as follows:

$$\mathrm{SELU}(x) = \lambda \begin{cases} x & x > 0 \\ \alpha e^{x} - \alpha & x \leq 0 \end{cases} \tag{4.1}$$

where α and γ are two fixed constants. The introduction of the SELU non-linear function significantly improved the model's performance. Its automatic regularization of network parameters and normalizing properties contribute to robust learning ([260]).

For training, we employ the backpropagation algorithm, which updates the weights and biases of the model during each epoch ([261]). To ensure a comprehensive evaluation of the model's predictive capabilities, we introduce ten different MLP topologies. Each topology maintains the same number of neurons in the input and output layers but varies the number of neurons in the two hidden layers. The selection of these parameters was guided by an initial phase of hyperparameter tuning, during which we evaluated various topologies and chose those that demonstrated the best performance. The specific topology configurations are listed in Table 4.1.

In addition to producing accurate predictions, we also aim to provide a confidence interval for each prediction. To achieve this, we exploit the properties of deep ensemble networks ([252]). By training multiple topologies, we obtain ten different results, and the final output of the model consists of the average of these predictions. The uncertainty associated with each prediction is calculated based on the differences among these results.

This comprehensive approach enables us to not only improve the predictive performance of the MLP but also to quantify the uncertainty in our predictions, making our model more reliable and informative for the given oceanographic dataset.

4.2.1 Experimental Setting

As stated above, to train and validate the model, measurements from the EMODnet dataset are used. The inputs chosen are:

- date (year, month, day)
- geolocation (latitude, longitude, depth)
- temperature
- salinity
- oxygen

The outputs we aim to predict consist of:

- nitrate (NO_3^-)
- phosphate (PO_4^{3-})
- silicate ($Si(OH)_4$)
- total alkalinity (A_T)
- chlorophyll-a
- ammonium (NH_4+).

The primary objective of our experimental campaign is to evaluate whether our model can effectively predict variables with low sampling frequency using variables that are sampled more frequently. To this end, we will measure the accuracy of our neural network in reproducing these low-sampled variables. The effectiveness of our model will be gauged through quantitative metrics like the RMSE, as well as by visually comparing the actual values with those predicted by our model. Additionally, to gain a comprehensive understanding of our model's performance and how it is influenced by input features such as geolocation or the timing of the sample, we will conduct an in-depth analysis of the results obtained.

Before training, the data is randomly shuffled, and subsequently, the dataset is divided into two sets: the training set, utilized for optimizing the model's weights, and the test set, employed to assess the performance of the proposed model. This partition is achieved by allocating 80% of the data to the training set and 20% to the test set.

To enhance the network's performance, a data preprocessing phase is undertaken, incorporating the most effective operations introduced by [216], [169], and [215].

Initially, the latitude input is normalized by dividing it by 90. This transformation ensures that latitude values, which typically range from -90 to 90, are scaled within the range of $[-1, 1]$.

Regarding the longitude input (denoted as lon), adjustments are made to account for its periodic nature. The following expressions are utilized:

- $|1 - \text{mod}(lon - 110.360)/180|$
- $|1 - \text{mod}(lon - 20.360)/180|$

Furthermore, the depth input is transformed to address the network's complexities in deep waters. The transformation involves a combination of linear and non-linear functions, as depicted in Equation 4.2:

$$D_{new} = \frac{D}{20000} + \frac{1}{(1 + \exp(-\frac{D}{300}))^3} \tag{4.2}$$

Where D denotes the original depth, and D_{new} represents the new preprocessed input depth.

The input and output data undergo a normalization process, which serves two main purposes: expediting the training process and reducing the likelihood of converging to local optima. This normalization is achieved by subtracting the mean $\bar{x}$ from the original variable x and then dividing the result by the standard deviation σ, as shown in Equation 4.3:

$$x_n = \frac{\gamma(x - \bar{x})}{\sigma} \tag{4.3}$$

In this equation, both $\bar{x}$ and σ are computed based on the data within the training set. The constant γ is introduced to expand the number of data points that fall within the range of $[-1, 1]$ after normalization. This normalization process is beneficial as it helps the training process converge more efficiently and ensures that the model is not heavily influenced by the scale of the input and output data, making it more robust and effective in capturing patterns and generalizing to unseen data.

The hyperparameters controlling the Adam algorithm ([97]), which serves as the gradient-based optimizer for minimizing the loss function, have been individually tuned for each variable. Through a preliminary study, various combinations of hyperparameter values were tested, and the optimal settings were determined. The chosen optimal values are documented in Table 4.2.

	NO_3^-	PO_4^{3-}	Si(OH)$_4$	A_T	Chl-a	NH$_4$+
Epochs	50.000	50.000	50.000	50.000	25.000	50.000
lr	0.005	0.005	0.005	0.001	0.005	0.005

TABLE 4.2: Epoch and Learning Rate used for the training phase.

The performance evaluation of the models employs two metrics: the *Mean Absolute Error* (MAE) and the *Root Mean Square Error* (RMSE). These metrics are defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\|Y_{P_i} - Y_{S_i}\| = \frac{1}{n}\sum_{i=1}^{n}\|e_i\|$$

and the *Root Mean Square Error* (RMSE), defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{P_i} - Y_{S_i})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}$$

where n is the dimension of the dataset; Y_{S_i} is the set of in situ values of the considered output, and Y_{P_i} the corresponding set of predictions. Both MAE and RMSE are widely used performance metrics in regression tasks. MAE represents the average absolute difference between the predicted and actual values, while RMSE calculates the square root of the average squared difference between the predicted and actual values. These metrics provide valuable insights into the accuracy and precision of the model's predictions for the given dataset.

4.2.2 The Two-Steps Quality Check Routine

EMODnet, being the product of an extensive data collection effort, poses the possibility of containing incorrect, noisy, or unreliable samples. Indeed, when applying our model to the original dataset, some outlier outputs occur, leading to significant deficiencies in our predictions. While the number of these anomalous values is relatively small, accounting for approximately 4.6% of the data, their impact on the overall error within the test set was notable. This raised concerns about potential biases introduced by these outliers on the model's predictive capabilities. The presence of such anomalous data in the EMODnet dataset may have led the model to learn from erroneous measurements, thus adversely affecting the relationships inferred from valid measurements.

To address this potential issue, we propose a two-step quality check routine aimed at identifying and removing incorrect data from the dataset. This process is summarized through a flowchart, as depicted in Figure 4.2. In the first step, the model is initially trained using the complete dataset, including both good-quality measurements

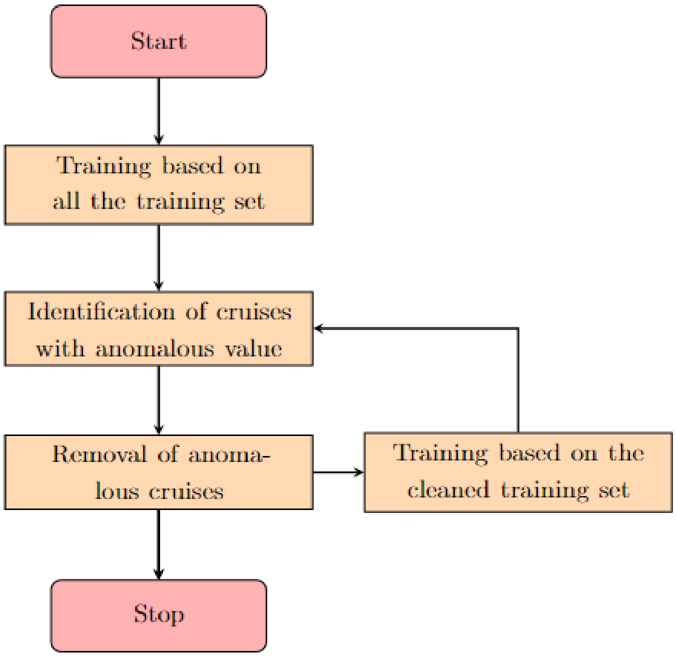

Start
Training based on all the training set
Identification of cruises with anomalous value
Removal of anomalous cruises
Training based on the cleaned training set
Stop

FIGURE 4.2: Two-step quality check routine flow chart.

and outliers. Subsequently, we proceed to identify subsets within the data that contain anomalous samples, which contribute to the highest 0.1% of error predictions. These subsets are scrutinized to ascertain whether they correspond to specific and well-defined categories, such as originating from the same sampling cruise, date, or provider. If so, they are removed from both the training and testing sets. Additionally, the criterion for subset removal is only applied if at least 25% of the samples within the subset are classified as outliers, ensuring a robust and meaningful removal process.

Having executed the initial cleaning, we retrain the model using the updated, cleaned dataset. This process of identifying and eliminating outliers is then iterated, with the model trained anew in each iteration until no further predictions are classified as unreliable. This iterative approach helps ensure a thorough and gradual improvement in data quality.

By implementing this rigorous quality check routine, we aim to mitigate the impact of anomalous data on our model's performance and enhance its predictive capabilities. This iterative process allows us to gradually fine-tune the training data, thereby producing more accurate and reliable predictions. The final dimensions of the training datasets, post-cleaning, are presented in Table 4.3.

This rigorous data quality assurance approach not only helps enhance the reliability of our model but also contributes to the credibility and accuracy of our findings in the context of oceanographic data analysis and prediction.

NO_3^-	PO_4^{3-}	$Si(OH)_4$	A_T	Chl-a	NH_4+
20.686	25.335	16.187	1.292	6.040	9.900

TABLE 4.3: Dimension of the training set for each variable, after the removal of noisy and unreliable data.

4.2.3 Prediction Confidence Interval

In addition to generating predictions, it is crucial to quantify the corresponding uncertainty, particularly when dealing with values collected over large and heterogeneous areas. To accomplish this task, we employ the concept of a *confidence interval* ([262]), a widely used statistical measure that quantifies prediction uncertainty.

For estimating predictive uncertainty, we utilize NN ensembles, commonly referred to as *deep ensembles*. This approach has proven to be highly effective in improving predictive performance and, importantly, provides a practical and scalable method for uncertainty estimation [263]. For a more comprehensive understanding of the topic, we refer readers to [264].

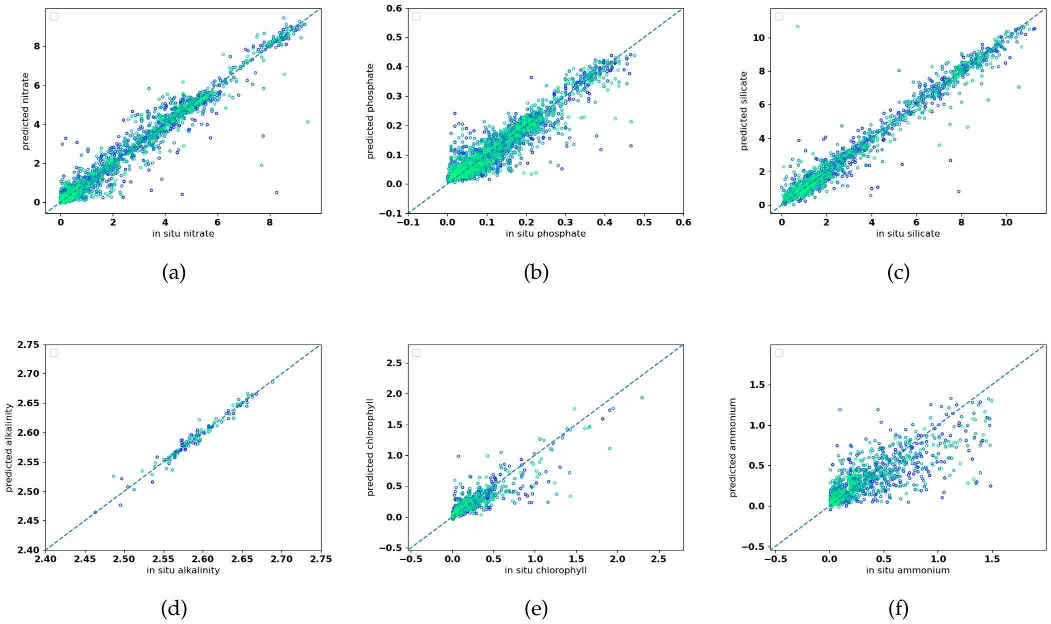
To calculate the uncertainty of our model's prediction, we obtain ten outputs for a given input by introducing changes in the network topologies, such as varying the number of neurons in each layer. This ensemble of ten NNs will be trained, ensuring a balanced trade-off between the number of outputs and computational resources. The confidence interval, representing the uncertainty, is then computed as the difference between the third quartile and the first quartile of the set comprising the ten different predictions.

The range between these two quartiles has been demonstrated to serve as a robust indicator of the reliability of ensemble DL predictions [264]. By providing a confidence interval for our predictions, we offer valuable insights into the uncertainty associated with our model's forecasts. This information is especially relevant when working with oceanographic data collected over diverse and extensive regions, allowing stakeholders to make well-informed decisions based on the level of confidence in the model's predictions.

Incorporating this measure of uncertainty enhances the credibility and usability of our predictive models, ensuring they are better suited to address real-world challenges in the field of oceanography and data science.

		NO_3^-	PO_4^{3-}	$Si(OH)_4$	A_T	Chl-a	NH_4+
MAE	canyon-med	0.47	0.026	0.40	6.5		
	our model	**0.26**	**0.019**	**0.31**	**5.6**	0.09	0.13
RMSE	canyon-med	0.73	0.045	0.70	11.1		
	our model	**0.50**	**0.031**	**0.58**	**8.2**	0.017	0.21

TABLE 4.4: MLP fitness values (MAE and RMSE).
A comparison between the current state of the art, *Canyon-Med* ([216]), and our method is also reported. The best results are highlighted in bold.



(a) (b) (c)

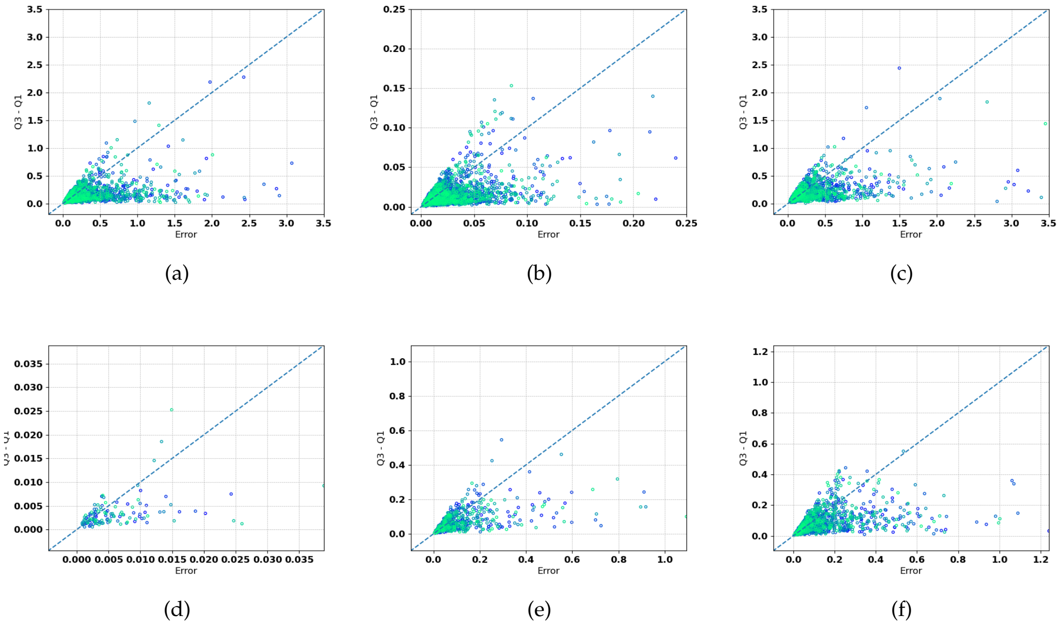
(d) (e) (f)

FIGURE 4.3: Comparison of in situ samples and MLP outputs.
(a) Nitrate (NO_3^-), (b) Phosphate (PO_4^{3-}), (c) Silicate ($Si(OH)_4$), (d) Alkalinity (A_T), (e) Chlorophyll-a, (f) Ammonium (NH_4+). The color denotes the depth of the samples, from green (shallow) to blue (deep).

4.3 Results

For each of the six variables, individual training and testing fitness, measured by both MAE and RMSE, are computed. However, in Table 4.4, we only present the testing fitness results, along with a comparison to the findings achieved by [216], which currently represents the state-of-the-art for the Mediterranean Sea application. The results reveal a general decrease in both fitness metrics compared to the current state-of-the-art. The most substantial decreases in skill metrics, ranging from 30% to 45%, are observed for NO_3^-, PO_4^{3-}, and $Si(OH)_4$. On the other hand, Alkalinity exhibited the least improvement, with a decrease in skill of around 15%.

Figure 4.3 presents a scatter plot that offers a visual comparison between the real values of the in situ measurements (on the x-axis) and the corresponding predictions made



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 4.4: Comparison between the prediction's RMSE and the corresponding interquartile range
(a) Nitrate (NO_3^-), (b) Phosphate (PO_4^{3-}), (c) Silicate ($Si(OH)_4$), (d) Alkalinity (A_T), (e) Chlorophyll-a, (f) Ammonium (NH_4+). The color denotes the depth of the samples, from green (shallow) to blue (deep).

by our model (on the y-axis). In the scatter plot, points that lie close to the diagonal indicate accurate predictions, implying that the model performed well in those instances. Upon examining the plots, we observe that for NO_3^-, PO_4^{3-}, and $Si(OH)_4$, the points are well-distributed along the diagonal, indicating a good agreement between the model's predictions and the actual measurements for these variables. However, for Chlorophyll-a and NH_4^+, the points appear more scattered, suggesting that the model's performance is comparatively less precise for these variables. The spread of points away from the diagonal may indicate greater variability and complexity in predicting these particular variables accurately. Furthermore, it is essential to consider that the number of data points for A_T is lower than for the other variables. This limitation in the data availability for A_T could potentially impact the robustness of the results for this variable.

Figure 4.4 displays scatter plots that illustrate the relationship between the error obtained by our model (on the x-axis) and the range between the third quantile and first quantile (on the y-axis), which represents the confidence interval. These scatter plots effectively demonstrate that a small difference in quantile values corresponds to small errors in the model's prediction, validating the reliability of this metric for computing the confidence interval. When the quantile difference is small, it indicates that the ensemble predictions exhibit low dispersion and are more consistent, reinforcing the

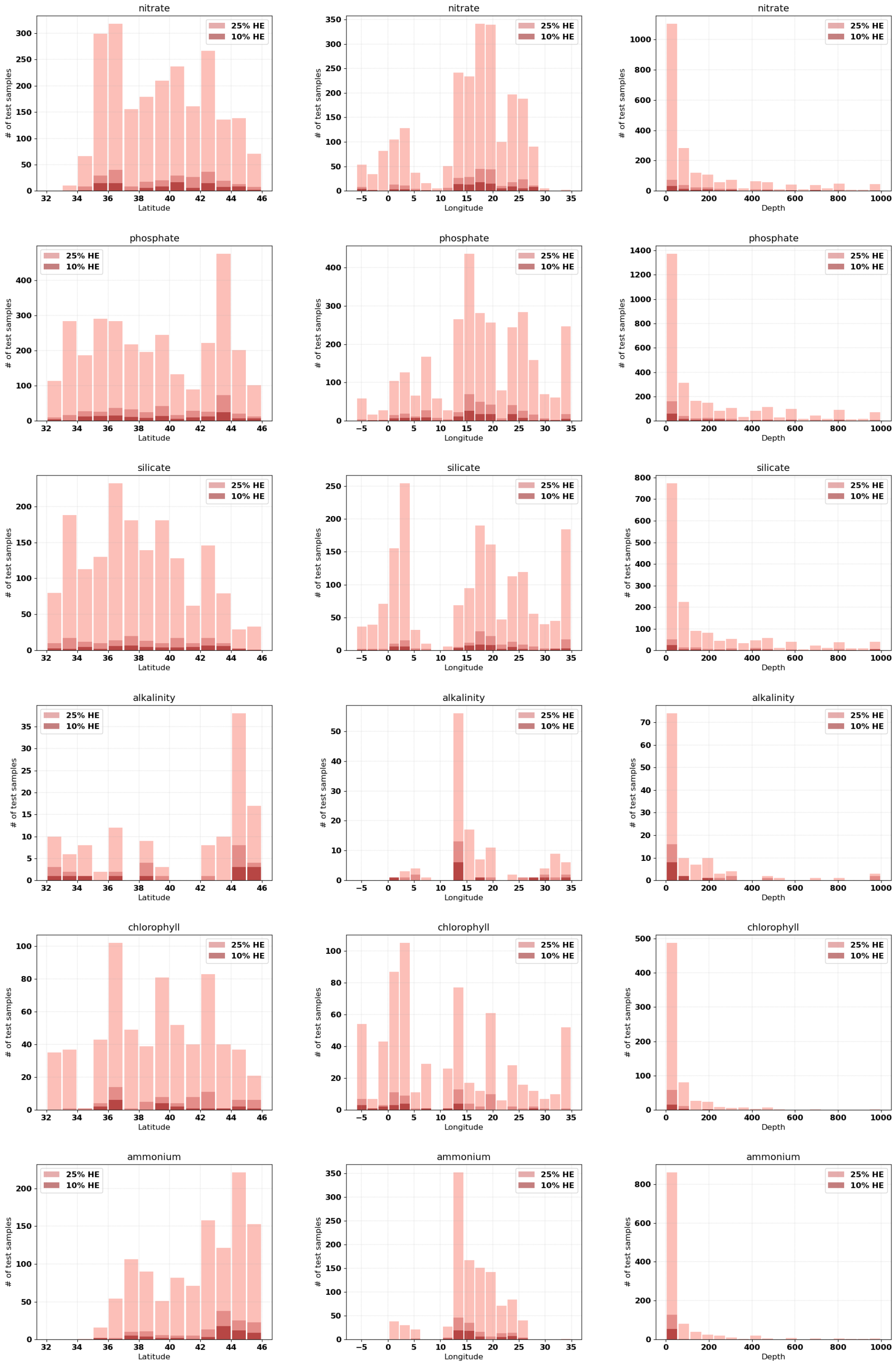


FIGURE 4.5: Test value distribution for the trained MLP model.
Variable represented are, from top to down: NO_3^-, PO_4^{3-}, $Si(OH)_4$, A_T, Chlorophyll-a, and NH_4+. Histograms in the first, second, and third columns represent the distribution of the results according to, respectively, their latitude, their longitude, and their depth. both 25% (25% HE) and 10% (10% HE) of the predictions leading to a higher error are highlighted in different colors.

robustness of the confidence interval as a measure of predictive uncertainty. Moreover, the plots also reveal an interesting pattern. As the errors in the predictions decrease, the dispersion of the ensemble predictions also tends to decrease, suggesting that the model's performance is less sensitive to the specific topology of the network when making more accurate predictions. This observation highlights the effectiveness of the ensemble approach in enhancing the model's stability and generalization capabilities. However, it is worth noting that a few points lie on the bottom-right of the plot, representing cases where there is low dispersion in the ensemble predictions despite poor model performance. This phenomenon may indicate the presence of outliers that were not identified by our two-step analysis for data cleaning. These outliers might have a significant impact on the model's predictions, leading to poor performance for certain variables.

Among the variables, NO_3^-, PO_4^{3-}, and $Si(OH)_4$ are those with a larger number of poor predictions with low dispersion.

Figure 4.5 presents a series of histograms depicting the distribution of test data as a function of latitude, longitude, and depth. For each distribution bin, the number of predictions with the highest 25% and 10% errors (25%HE and 10%HE, respectively) is indicated in darker colors. This visualization facilitates an investigation into potential inhomogeneity between the data distribution and the error distribution. The histograms for NO_3^-, PO_4^{3-}, $Si(OH)_4$, and Chlorophyll-a show good and fairly homogeneous coverage in terms of latitude. However, the distribution of latitude is biased due to the presence of the three major basins in the Mediterranean Sea (western, Adriatic, and Levantine basins). On the other hand, Alkalinity exhibits a biased distribution, with the largest number of observations concentrated in the Northern Adriatic Sea. Similarly, NH_4^+ displays a biased coverage, with observations primarily sampled in the Adriatic Sea. Examining the distribution of 10%HE predictions along the horizontal dimension (both latitude and longitude) reveals uniform patterns for all variables, except for PO_4^{3-} and $Si(OH)_4$. These two variables demonstrate a higher frequency of 10%HE predictions for latitudes ranging between $41° - 42°$N, with 30% and 20% of predictions falling into the 10%HE category for PO_4^{3-} and $Si(OH)_4$, respectively. Regarding the depth dimension (third column of the plots), the distribution of 10%HE predictions shows that the model tends to be more accurate on the surface for all variables. However, the percentage of samples characterized by 10%HE increases with depth. Specifically, the ratio between 10%HE samples and total samples available exceeds 20% for depths beyond 100 meters. These plots offer additional indicators for assessing the reliability of model predictions. For instance, if the model is applied to samples belonging to geographical areas that have been predicted with higher precision in the past, the corresponding results can be labeled as reliable with higher confidence. Conversely, geographical areas exhibiting higher occurrences of 25%HE or 10%HE predictions should be subject to more careful investigation, as they may indicate peculiarities that deviate from the mean model performance.

The diagnostic metrics presented in Figures 4.3 to 4.5 collectively provide an overall assessment of the goodness of the reconstruction for each of the six variables. Regarding Nitrate (Figure 4.3(a)), the model demonstrates satisfactory performance, accurately approximating the observations. A similar behavior is observed for Phosphate (Figure 4.3(b)), where the model skillfully predicts values across both the lower and higher ranges. For Silicate (Figure 4.3(c)), the results are particularly satisfactory, as the model achieves accurate predictions across various value ranges. Alkalinity (Figure 4.3(d)) also shows positive results, although the bias distribution of observations (Figure 4.5) in certain zones may impact our ability to draw definitive conclusions about error distribution in those regions. In the case of Chlorophyll-a (Figure 4.3(e)), a clear distinction emerges in the quality of predictions for different value ranges. Lower values are predicted more accurately than higher ones. This discrepancy is attributed to the difference in the quantity of high and low-value samples available in the training set. The scarcity of high-value samples might have affected the model's ability to accurately predict such values. Regarding Ammonium (Figure 4.3(f)), this marine variable exhibits the least accurate predictions among the six variables. The complexity of ammonium distribution, which depends on numerous interacting biological and chemical processes, is a plausible reason for the model's limitations in reconstructing this variable. The explicative variables used in our models may not have been sufficient to capture the complexities involved in ammonium variability in the Mediterranean Sea. Additionally, the biased spatial distribution of observations, with a concentration of data in the Adriatic Sea (latitude $43° - 46°$N, longitude $13° - 18°$W), might have hindered the model's ability to generalize and map ammonium variations accurately. The accumulation of poor predictions in this marginal sea further highlights the challenges in predicting ammonium accurately.

4.4 MLP for Data Assimilation

In this section, is reported the impact of incorporating the proposed MLP model into the framework of DA. Specifically, we will evaluate the effect of the integration of BGC-Argo and MLP reconstructed profiles into the biogeochemical simulations of the Mediterranean Sea conducted by the Operational System for Short-Term Forecasting

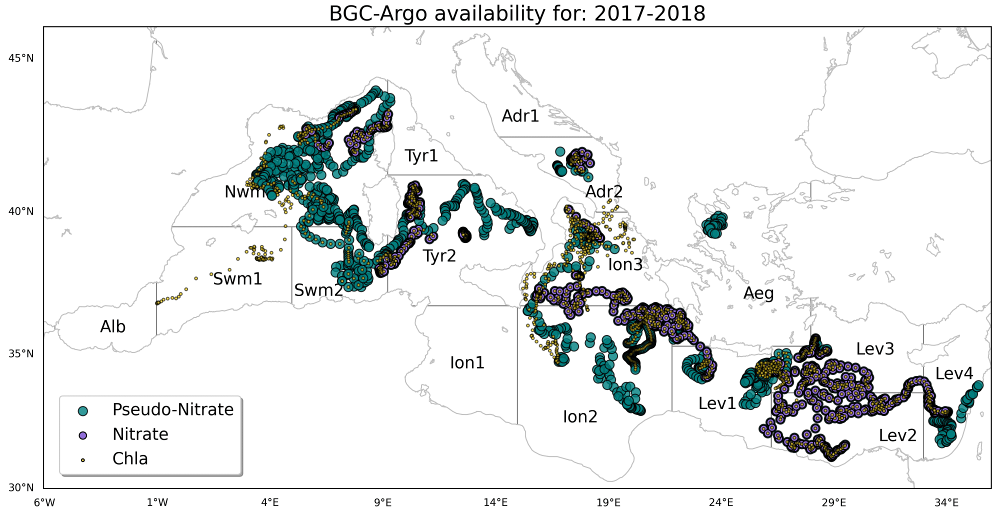


FIGURE 4.6: BGC-profiles of chlorophyll-a (red), Nitrate in-situ (orange), and reconstructed Nitrate (grey) assimilated in the Mediterranean Sea (2017-2018).

of the Biogeochemistry of the Mediterranean (MedBFM) [38]. We will prove that integrating reconstructed profiles can lead to enhanced modeling outcomes and improved performance. The whole theoretical framework and the investigation of different aspects of the combination of NN and DA is reported in Appendix A, while here solely the principal results which highlight the potential of the proposed MLP model are summarized.

To perform this task are evaluated the model's outputs and compared with the results obtained through the addition of the DA routine. By analyzing these outputs and conducting a comparative analysis, we aim to determine which approach yields more precise results: the plain model, the model integrated with solely BCG-Argo profiles or the model integrated with both BCG-Argo profiles and reconstructed profiles. To assess precision, we will compare this variation of the model against observational data sources, such as float measurements and satellite data.

4.4.1 The Assimilation Scheme

Test Case	Chl	O2	NO3	Updated Variables
HIND	–	–	–	–
DAfl	1773	1924	938	phyto biomass, NO3 , O2 and PO4
DAnn	1773	1924	2146	phyto biomass, NO3 , O2 and PO4

TABLE 4.5: Summary of the numerical experiments and assimilated BGC-profiles

The assimilation scheme employed in this study is founded on the 3DVarBio method, as described in previous works [254, 255, 257, 258]. To provide more detailed insight, the assimilation scheme employed incorporates oxygen, chlorophyll, nitrate, all assimilated variables, and various phytoplankton biomasses, including phosphate, in its updates.

The assimilated observations comprise the QC BGC-Argo dataset detailed in Table 4.5. Specifically, the assimilation process incorporates oxygen and nitrate profiles within the $0 - 600m$ layer, while chlorophyll data are assimilated within the $0 - 200m$ layer. The nitrate dataset utilized as input for assimilation comprises 938 BGC-Argo profiles and an additional 2146 reconstructed nitrate profiles. The number of reconstructed nitrate profiles is more than double with respect to the amount of nitrate profiles collected in-situ by the Argo float devices, confirming the utility of the developed MLP methods in augmenting the information about a fundamental variable which allow the improving of the knowledge of the Mediterranean Sea. Notably, these reconstructed nitrate profiles are distributed across the western (61%) and eastern (39%) Mediterranean Sea, resulting in a more extensive and spatially consistent coverage. This distribution is visually represented in Figures 4.6, highlighting the improved spatial representation achieved through the inclusion of these reconstructed profiles.

To evaluate the effects of various assimilation configurations, we conducted three distinct numerical experiments. These experiments covered from January 1, 2017, to December 31, 2018. The setup of the MedBFM closely aligns with the standard configuration utilized in the Mediterranean Analysis and Forecast biogeochemical system of the Marine Copernicus Service. This setup encompasses several critical components, including open boundary conditions in the Atlantic, the utilization of climatological data for nutrients, carbon, and alkalinity from 39 rivers, as well as data from the Dardanelles Straits. Additionally, the initial conditions for the model were derived from the EMODnet dataset, with further details provided [265]. This comprehensive setup forms the basis for our numerical experiments, enabling us to assess the impact of different assimilation configurations effectively.

Figure 4.6 also illustrates the segmentation of the Mediterranean domain into sub-basins, which will serve as the basis for validation in subsequent chapters. These sub-basins adhere to a common division widely employed in oceanography to characterize the Mediterranean Sea. To ensure the robustness and consistency of our validation metrics, we may employ different combinations of these sub-basins or subsets based on data availability. The following abbreviations represent these combinations:

- Levantine (**lev**) comprises lev1, lev2, lev3, and lev4.

- Ionian (**ion**) consists of ion1, ion2, and ion3.
- Tyrrhenian (**tyr**) includes tyr1 and tyr2.
- Adriatic (**adr**) encompasses adr1 and adr2.
- South Western Mediterranean (**swm**) is constituted by swm1 and swm2.
- North Western Mediterranean (**nwm**).

To test the impact of the additional observations reconstructed with MLP the model described in [2] has been run under three scenarios:

1. Control Run (HIND): This serves as the baseline simulation without any assimilation.
2. Assimilation of BGC-Argo Chlorophyll, Nitrate, and Oxygen (DAfl): In this simulation, we incorporate data assimilation for BGC-Argo chlorophyll, nitrate, and oxygen.
3. Enhanced Data Assimilation with Additional Reconstructed Nitrate Profiles (DAnn): This simulation builds upon DAfl by including additional reconstructed nitrate profiles, further enhancing the assimilative setup.

During the data assimilation process, profiles may be excluded if the model-observation misfit exceeds predetermined thresholds. This highlight the fact that the MLP reconstructions can be affected by local biases. This ensures that only high-quality data is incorporated into the assimilation process, improving the overall accuracy of the simulations.

4.4.2 Results

The performance of the simulations outlined in Table 4.5 is assessed by comparing model outputs with both satellite ocean color (OC) chlorophyll data and BGC-Argo profiles. For the satellite comparison, have been consider daily averages generated by the model. However, when evaluating metrics based on BGC-Argo profiles, have been utilize the model's first guess, representing its state before the assimilation process. To test the model's effectiveness in reproducing specific bloom and stratification conditions, is employed the RMSE metric. This evaluation is conducted for two distinct seasons: winter (from February to April, FMA) and summer (from June to August, JJA). We assess the model's performance across 16 Mediterranean Sea sub-basins or in

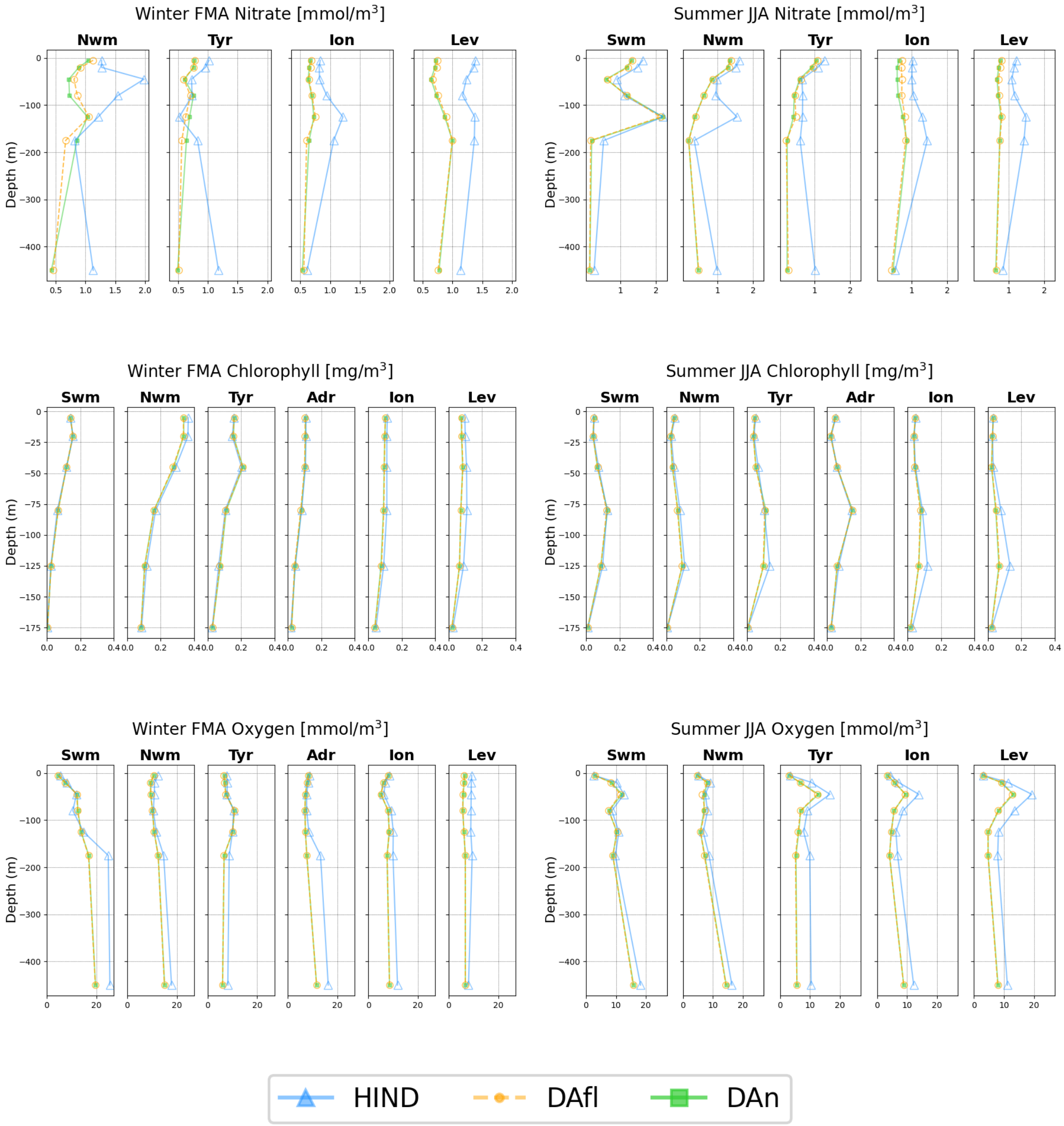


FIGURE 4.7: Seasonal Nitrate, Chlorophyll and Oxygen RMSE (top, middle, bottom): Bloom (left) and Stratification (right) seasons in the Mediterranean Sea aggregated sub-basins for the HIND run (pale blue), DAfl run (orange) and DAnn (green).

aggregated combinations of these sub-basins, as indicated in Figure 4.6. This approach provides insights into the model's ability to capture the unique characteristics of each season in these specific regions.

The validation of the model is demonstrated in [2], here we focused on the impact of the MLP reconstructed data on simulation dynamics.

In terms of the Winter RMSE concerning OC chlorophyll levels in the HIND simulation a reduction in RMSE (up to 10%) compared to HIND is achieved with the DAnn simulation, demonstrating that expanding the nitrate float network contributes to improved representation of surface phytoplankton dynamics. In general, all Mediterranean sub-basins show a reduction in RMSE, except for Alb, Swm, and Nwm. However, it's important to note that during the summer stratification period, a slight deterioration in the assimilated runs is observed. The RMSE values with respect to OC chlorophyll increase in all sub-basins during this season, albeit fairly similarly in the two assimilation runs: approximately 6% in DAfl and 7.5% in DAnn, respectively. It's worth mentioning that the RMSE values in summer are notably lower than in winter, underscoring the seasonal variability in chlorophyll levels in the Mediterranean Sea, which includes very low chlorophyll values at the surface.

The assimilation of in-situ BGC-Argo data leads to a substantial enhancement in the model's representation of nitrate in to the HIND run (Figure 4.7). Winter RMSE reduction goes from 40% (DAfl) to approximately 46%, when also reconstructed profiles are assimilated, while the reduction of summer RMSE increases from 59% in DAfl to 63% in DAnn. The most significant RMSE reduction between DAnn and DAfl is observed in the Nwm and Tyr sub-basins during winter and in the Ion sub-basins during summer. These findings underscore the effectiveness of assimilating both BGC-Argo and reconstructed profiles in enhancing the model's representation of nitrate levels.

The integration of the MLP and DA has led to multiple significant impacts. The Figures related to the effects of this integration are reported in Appendix A. Firstly, regarding the impact on biogeochemical vertical dynamics, the assimilation of nitrate effectively rectifies a prevalent positive bias inherent in the model across all Mediterranean regions, illustrated by the blue pattern in Figure A.7. The introduction of reconstructed profiles amplifies these corrections, making them more robust and precise. At the Mediterranean scale, there is an 8% reduction in nitrate concentration below the nitracline in the DAfl run and an 11% reduction in the DAnn run. The nitracline depth undergoes changes, characterized by deeper values following the DAfl assimilation, with variations ranging from a few meters (e.g., Nwm) to tens of meters (e.g., Ion2). This deepening phenomenon becomes more pronounced with the inclusion of reconstructed profiles in DAnn. Remarkably similar patterns are observed in the Hovmöller diagrams of phosphate (Figure A.9). These patterns can be attributed to the high positive values in the covariance matrix between nitrate and phosphate, as elaborated in [255].

4.5 Discussion and Final Considerations

The DL architecture proposed in this Chapter relies on the availability of high-frequency sampled variables, including essential data such as temperature, salinity, and oxygen, complemented by additional ancillary information like the temporal aspect and geospatial coordinates of the sampling locations, in order to predict low-frequency sampled variables, such as nitrate, phosphate, silicate, alkalinity, chlorophyll and ammonium. The core objective is to bridge the gap between high-frequency and low-frequency variables, allowing us to make informed predictions of the latter using the former as inputs.

Building upon previous research ([169]-[215] and [216]), specific improvements are introduced, tailored to the Mediterranean Sea context. A larger and higher-quality training dataset from [63] is used, accompanied by a new two-step quality check routine to enhance dataset reliability. Additionally, a confidence interval estimation is incorporated, leveraging deep ensembles of NNs. Accurate validation demonstrated significant improvements over existing methods, with enhanced prediction performance for all variables and fitness metrics. These advances were attributed to the larger and improved dataset, as well as the refined quality check process. The inclusion of two additional variables, Chlorophyll-a and Ammonium (NH_4+), further showcased the model's potential.

The proposed DL framework offers valuable applications that can contribute significantly to oceanographic research and monitoring efforts.

Ship-based sampling is known to have its inherent limitations. These limitations often stem from practical constraints, such as the availability of ship time and human resources, as well as environmental factors like adverse weather conditions that can restrict sampling in specific regions or during certain periods, such as the winter season. These challenges have historically hindered the collection of comprehensive and continuous oceanographic data. The introduction of the MLP model in this context offers a promising solution. This model exhibits the capability to effectively fill gaps in observations, making it possible to harness the full potential of datasets like EMODnet. It becomes feasible to generate data where conventional sampling methods fall short.

Moreover, the proposed MLP extends its utility beyond data completion. It can be used as a valuable tool for identifying periods and areas within the Mediterranean Sea where data density remains insufficient, both spatially and temporally. Such limitations can hinder our ability to comprehend complex oceanic processes and assess long-term variability accurately. As we demonstrated in this study, areas where the DL model exhibits less satisfactory results often coincide with deficiencies in the training database, which may fail to adequately capture the full spectrum of spatial and temporal variability.

Additionally, this approach can also applied to the vast network of BGC-Argo profiling floats, each equipped with advanced sensors capable of measuring the input oceanographic variables (temperature, salinity and oxygen). This extension can significantly enhance the flow of biogeochemical data, contributing to a more comprehensive understanding of the marine environment. Furthermore, the proposed MLP model, as part of this approach, can play a crucial role in quality control. It can assist in refining parameters like NO_{-}^{3} and pH obtained from these autonomous platforms, ensuring data accuracy by correcting sensor drift during deployments and adjusting deep-sea values as required.

As a demonstration of the potential of the described MLP model, the results of the nitrate reconstruction have been integrated in a big model-DA numerical experiments of the Copernicus marine Biogeochemical system of the Mediterranean Sea. The assimilation of nitrate data from BGC-Argo profiling floats effectively mitigates a prevailing positive bias observed in the model across various Mediterranean regions. Moreover, the inclusion of reconstructed profiles further enhances this corrective effect. Concurrently, the adjustment of phosphate concentrations based on error covariances contributes to spatial and multivariate alterations that hold the potential to rectify pivotal biogeochemical phenomena, such as the nitracline and deep chlorophyll maximum. This application underscores the practicality and efficacy of developing predictive marine models. It is important to note that the utility of the model we have constructed extends beyond this specific case, as it can be applied to a multitude of scenarios, exemplifying its versatility and broader applicability.

The heterogeneity of the Mediterranean Sea, characterized by significant physical and biogeochemical gradients, poses challenges for the model's predictions, especially in regions with uneven data coverage. Investigating the benefits of restricting the investigated areas for more precise results, while maintaining sufficient training data, warrants further exploration.

In conclusion, this DL framework significantly advances data-driven oceanography and offers practical solutions for marine variable prediction and quality control.

Chapter 5

PPCon: 1D CNN for Predicting Biogeochemical Variables

Access to reliable and extensive oceanic measurements remains restricted due to the challenges of collecting comprehensive observations on multiple temporal and spatial scales, as well as variability in the availability of observations across different biogeochemical variables ([17, 26, 60]). The introduction of autonomous oceanographic instruments such as Biogeochemical (BGC) Argo floats have notably expanded our ability to obtain subsurface and deep ocean measurements ([253]). BGC-Argo floats are autonomous profiling platforms that incorporate physical and biogeochemical sensors, allowing to collect time-series of vertical profiles across various sea conditions and throughout the complete annual cycle ([266, 267]). Strong efforts have been devoted toward the improvement of the long-term reliability and accuracy of autonomous measurements in recent years ([215]).

However, the measurement of biogeochemical variables such as nutrient concentration and carbonate system variables (e.g., nitrate, chlorophyll, and pH) remains more demanding and expensive compared to physical variables (e.g. temperature, salinity) and oxygen. In fact, among the BCG sensors, oxygen is the most commonly measured variable: there have been approximately $250,000$ oxygen profiles collected worldwide, which is twice the number of profiles for chlorophyll, and more than four times the number of profiles for nitrate and bbp700 (`https://biogeochemical-argo.org`).

By being able to predict low-sampled variable profiles from high-sampled ones, we can unlock the full potential of the Argo observing system. This enhances our understanding of the behavior of biogeochemical variables by augmenting the original dataset with synthetic (yet reliable) profiles.

5.0.1 Research Question

One of the primary messages this Thesis aims to provide is that different types of oceanographic data require distinct architectures to effectively extract knowledge from them. When the task involves predicting a punctual output, an MLP architecture is a suitable choice. However, for predicting vertical profiles using a vertical profile dataset, an architecture that directly infers the complete vertical profile proves to be beneficial. This approach takes advantage of architectures like the CNN that operate on vector inputs instead of individual points.

In this chapter, our investigation centers on a key question:

- Does employing a CNN architecture enhance the prediction accuracy for low-frequency sampled variables derived from high-frequency data, when the training dataset comprises vertical profiles?

To address this question, we have developed a CNN-based architecture and concentrated on the following aspects:

- The accuracy of low-frequency variable predictions by a CNN model trained on BCG-Argo float vertical profiles.
- The performance variability of the CNN across different geographic regions and seasons.
- A comparative analysis of prediction quality between the new CNN model and the previously used Multi-Layer Perceptron (MLP) model for variables that both methods can predict.
- The extent to which our method increases the number of low-frequency sampled vertical profiles in the BCG-Argo dataset.

Related Works

Existing ANN-based techniques to infer low-sampled variables starting from high-sampled ones are based on MLP architecture ([1, 169, 215, 216]), as described in Section 4.4. This kind of DL method is also the one investigated in Chapter 5, which is trained used a dataset originating from cruise measurements. Despite their widespread use, applications based on MLPs currently lack awareness of the typical shape of biogeochemical variable profiles they aim to infer. When these methods are employed to forecast profiles from Argo float measurements (which is the objective of this Chapter), they may generate jumps and irregularities, in the reconstruction. This originates from the fact that MLPs are trained on individual data points and provide pointwise outputs, which makes the generation of regular profiles challenging as the NN is not aware of the vertical neighbors of predicted variables.

Contribution

In this study, we evaluate the effectiveness of a one-dimensional (1D) CNN model ([65]) for predicting nutrient vertical profiles from input data such as sampling time, geolocation, and profiles of temperature, salinity, and oxygen, using Argo float measurements as the training dataset. This approach, called PPCon (Predict Profile Convolutional) is applied to generate synthetic profiles of nitrate, chlorophyll, and backscattering (bbp700). Thanks to the intrinsic spatial-aware nature of its CNN architecture, PPCon can leverage the typical shape of vertical profiles of a variable as a prior constraint during training. PPCon predictions are characterized by lower error with respect to the one obtained with MLP while showing also smoother predictions and the disappearing of phenomena such as gaps and irregularities in the generation of vertical profiles.

Structure of the Chapter

This Chapter is organized as follows: Section 5.1 presents the dataset utilized for training the DL (DL) architecture, including its key characteristics. Subsequently, Section 5.2 provides a detailed overview of the PPCon approach, encompassing the architecture, preprocessing techniques applied to input data, and the specialized loss function employed for network training. In Section 5.3, we outline the specific experimental settings employed to enable complete reproducibility of the PPCon architecture. Section 5.4 presents a summary of the key results obtained during the experimental campaign we conducted to validate our proposed techniques, and Section 5.5 discusses them. Finally, Section 5.6 presents the conclusions drawn from our work and directions for future research.

5.1 Dataset: the Argo GDACs

The data used to train and test the architecture discussed in this paper comes from the Array for Real-time Geostrophic Oceanography (Argo) program [64], specifically the Argo float collecting also biogeochemical variables (BCG Argo float). This program is an important part of the Global Ocean Observing System (GOOS) (https://www.goosocean.org/) and is dedicated to monitoring changes in the temperature and salinity of the upper ocean. The Argo program was primarily designed to observe pressure, temperature, and salinity (conductivity) within the upper 2000 meters of the ocean. However, due to advancements in float and sensor technologies, newly developed sensors now enable profiling floats to accurately observe biogeochemical properties. Over the past decade, there has been a consistent and substantial increase in the number of biogeochemical profiles obtained through BGC-Argo float platforms. For instance, by 2011, the global ocean had accumulated approximately 45, 000 BGC-Argo profiles across all parameters, while, by 2017, this number had risen to almost 390, 000 profiles. Thus, the BGC-Argo floats network has become a crucial component of the ocean observing system, enabling the monitoring, understanding, and prediction of changes in the ocean ecosystem. As of now, the Argo program has amassed over 2 million data profiles, and analysis of this data has made significant contributions to basic research as well as national and international climate studies [268].

Our investigation specifically centers on the dataset made available by the Argo Global Data Assembly Centers (GDACs) (e.g., Coriolis, NOAA, among others), which disseminate Delayed Mode (DM) data procured from 11 data centers situated across 9 countries. DM data undergoes a more exacting quality control process and is typically released a few months later to their sampling [269]. GDACs also supply Real-Time Mode (RT) data, however, given the lower quality of RT data, our analysis is based only on available DM data.

Our model was trained using data from the Mediterranean basin collected between 2015 and 2020. To ensure the data's reliability, we only selected profiles that were marked with quality flags (QFs) of 1, 2, or 8 for variables such as temperature, salinity, nitrate, and chlorophyll. Furthermore, we applied a preprocessing step on the bbp700 variable based on the study by [270], which introduced a new set of real-time quality-control tests for this variable, as so far no procedures have been agreed upon to quality control bbp700 data in real-time. Specifically, we applied three of the procedures introduced in this work to bbp700 profiles: the missing-data test, which detects and flags profiles that contain a substantial amount of missing data; the high-deep-value test, which flags profiles with unusually high bbp700 values at depth; and the negative-BBP test, which flags data points or profiles with negative bbp700 values.

5.2 PPCOn: Profile Prediction Convolutional NN

CNNs have emerged as a prominent architecture ([102]). While the two-dimensional (2D) CNN architecture is designed to extract spatial features from two-dimensional data such as images ([101]), the 1D CNN is specifically designed to extract temporal features from one-dimensional sequential data such as signals or time series data ([271]). Due to their streamlined and efficient configuration that employs only 1D convolutions (scalar multiplications and additions), 1D CNNs offer advantages in terms of real-time processing and cost-effectiveness for hardware implementation. ([65]).

This section introduces the PPCon architecture, which is primarily a 1D CNN with additional MLPs employed to transform punctual data into a vectorial shape - necessary for the training of the convolutional component. The input variables for PPCon include sampling data, geolocation, temperature, salinity, and oxygen, while the output variables comprise vertical profiles for nitrate, chlorophyll, and BBP. Despite using the same architecture, a separate model is trained for each output variable, and different hyperparameters (number of epochs, weights of the loss function, and so on) are set for each of them. This separate tuning is necessary due to some intrinsic differences such as the numerosity of the training set and the variable ranges. The hyperparameters are tuned manually by comparing performance on the test set composed of unseen data, based on a fitness metric to be introduced later. A specific loss function is specifically designed to promote good performances, generalization capabilities, and smooth predictions.

5.2.1 Input Preprocessing

TABLE 5.1: The MLP component of the PPCon model illustrated in diagram form. All the 4 MLPs used in the PPCon architecture share the same architecture.

Layer	Output Size	Activation Function
Input	$[32, 1]$	-
Linear	$[32, 80]$	SELU
Linear	$[32, 140]$	SELU
Linear	$[32, 200]$	SELU
Output	$[32, 200]$	-

The data considered for feeding the DL architecture comprises a collection of measurements, where each input-output pair consist of the information collected by a singular float profile. The inputs consist of two distinct data categories, punctual and vectorial. Punctual data encompasses temporal and geospatial parameters, such as the sampling date (specifically year and day), geolocation, and geographic coordinates (latitude and

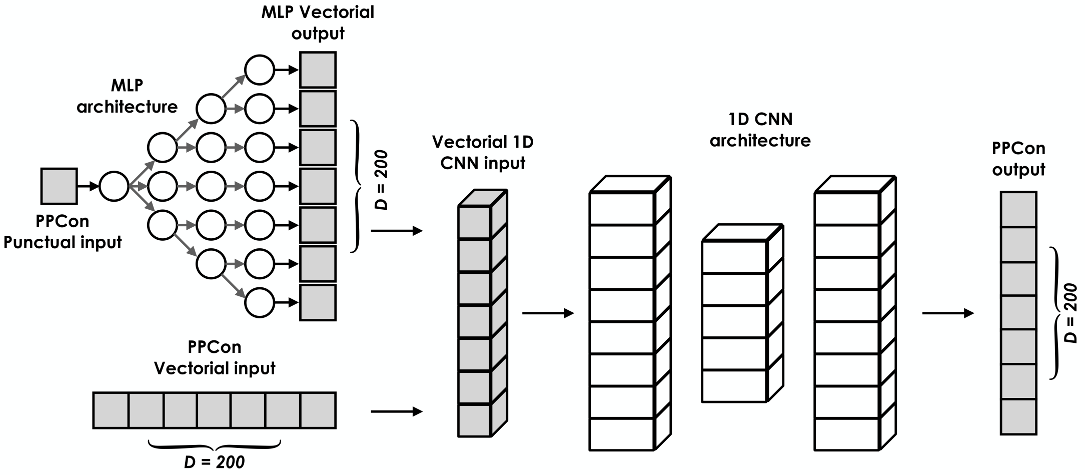


FIGURE 5.1: Illustration of the PPCon architecture.

longitude), while vectorial data encapsulates profiles of temperature, salinity, and oxygen, as recorded by the float instruments. Given that the 1D CNN architecture operates only on vectorial input data, a coherent transformation of punctual features into vectorial ones is required.

In this regard, we leverage an MLP architecture that accepts punctual input and transforms it into vectorial form. MLPs are employed in order to enable the NN to automatically learn how to weigh differently the importance of such punctual input features in correspondence of different levels of depth. A separate MLP is trained for each of the four pointwise inputs. The MLP architectures have the same number of layers and neurons contained in these layers (Table 5.1), since there are no a priori reasons to make them different.

During training, the weights of the MLP are optimized along with the weights of the 1D CNN architecture. As the MLP operates as a non-linear function, this training approach enables the creation of a mapping between punctual input and its vectorial equivalent. This enables PPCon to effectively exploit punctual information and achieve optimal learning outcomes. The output vectors generated by the MLP are concatenated with the remaining vectorial input, yielding a seven-channel tensor that serves the input of the PPCon architecture.

5.2.2 PPCon Architecture

The convolutional component of the PPCon architecture, summarized in Table 5.2, is a DL model comprising multiple 1D convolutional and deconvolutional layers.

TABLE 5.2: The convolutional component of the PPCon model illustrated in diagram form.

The key attributes of the NN are outlined, encompassing parameters, output size (represented as [batch size, number of channels, input length]), as well as any additional layers. More specifically, ”BN” denotes the batch normalization layer, ”SELU” represents the non-linear selu() activation layer, and ”DP” indicates the presence of a dropout layer (the dropout rate is the same for each layer and is 0.2).)

Layer	**Kernel**	**Stride**	**Padding**	**Output Size**	**Add. Details**
Input	-	-	-	$[32, 7, 200]$	-
Conv. 1D	2	1	2	$[32, 64, 203]$	BN, SELU, DP
Conv. 1D	2	2	1	$[32, 128, 102]$	BN, SELU, DP
Conv. 1D	4	1	1	$[32, 128, 101]$	BN, SELU, DP
Conv. 1D	4	1	2	$[32, 128, 102]$	BN, SELU, DP
Deconv. 1D	2	2	2	$[32, 128, 200]$	BN, SELU, DP
Conv. 1D	3	1	1	$[32, 128, 200]$	BN, SELU, DP
Deconv. 1D	2	2	1	$[32, 64, 398]$	BN, SELU, DP
Conv. 1D	2	2	1	$[32, 32, 200]$	BN, SELU, DP
Conv. 1D	3	1	1	$[32, 1, 200]$	BN, SELU
Output	-	-	-	$[32, 1, 200]$	-

The input to the PPCon architecture consists of four point-wise inputs — latitude, longitude, day, and year — which are transformed into a vectorial input using an MLP architecture. In addition, the architecture uses for the training three 1x200 input vectors representing the profiles of temperature, salinity, and oxygen.

The input tensor has a 1-dimensional shape, with a total number of channels equal to 7, one for each of the three variables to reconstruct.

The architecture includes a total of nine layers, each of which applies a set of filters to the input tensor. These filters are designed to detect specific features or patterns, with the number and size of the filter kernels specified by the parameters of each layer. To enable effective feature extraction across different scales, various stride parameters are employed to specify the step size at which the filters are applied to the input tensor. To ensure that the output tensor has the same shape as the input tensor, padding parameters are incorporated, adding zero padding to the borders of the input tensor. The output tensor is then normalized through a batch normalization ([272]) layer after each convolutional layer. The normalization process ensures that the output tensor has a mean of zero and a unit variance, thereby minimizing the effect of covariate shifts and enhancing the stability of the training process. Following normalization, the output tensor is passed through a scaled exponential linear unit (SELU) activation function

([273]), which is defined as:

$$f(x) = \begin{cases} \lambda x & \text{if } x \geq 0 \\ \lambda \alpha (e^x) & \text{if } x < 0 \end{cases} \tag{5.1}$$

where and $\lambda \approx 1.0507$ and $\alpha \approx 1.6732$. SELU has been selected as an activation function as it induces self-normalization properties. Dropout layers ([274]) are also incorporated to prevent overfitting during training, promoting robust generalization and enhancing the NN ability to learn diverse features from the input data. These layers randomly drop out some of the network neurons, with the specific probability of dropout (d_r) specified for each layer in the architecture's hyperparameters.

The final convolutional layer produces a 1-channel output tensor, which represents the final prediction of the model.

5.2.3 Loss Function

The choice and design of a loss function is a crucial step in the development of DL models, as it determines the objective to be optimized during training and can have a significant impact on the model's ability to generalize to new data. Besides the ability to reproduce skilfully output variable profiles, we want the PPCon architecture to mitigate overfitting and produce a smooth prediction curve.

To fulfill these objectives, we define a loss function comprising three components: first, the Root Mean Square Error (RMSE) between the target output and the PPCon architecture's prediction, to assess prediction quality. Second, to mitigate overfitting phenomena, a regularization term known as λ-regularization is employed, which penalizes complex curves in proportion to the square of the model's weights ([275]). By promoting smaller weight values, this technique encourages the generation of more general predictions. The severity of this penalty is determined by a multiplicative factor λ, which is a hyperparameter of the model. The final component of the loss function is incorporated to promote the generation of a smoother output curve. This term, controlled by a hyperparameter α, serves as a regularization technique that penalizes sharp variations in the output. The final loss formula is as follows:

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{N} |\theta_i| + \sum_{i=1}^{n-1} (\hat{y}_{i+1} - \hat{y}_i)^2 \tag{5.2}$$

where y represents the target value, $\hat{y}$ the output of PPCon model, n the length of both y and $\hat{y}$ and N the total number of weights of the DL architecture.

5.3 Experimental Study

This section presents the experimental settings for the PPCon architecture, which are defined for each predicted variable under consideration. The complete code for the reproducibility of the results presented in this paper is available at the following URL: https://github.com/gpietrop/PPCon.

Our experimental campaign aims to evaluate our model's ability to recreate the vertical profiles of variables with low sampling frequency by feeding high-frequency sampled data into the proposed architecture. We will examine the proficiency of our neural network through two key approaches: first, by calculating the RMSE between the predicted and original profiles, and second, by providing a visual depiction of the generated profiles to observe how accurately the network reproduces these vertical profiles. Additionally, to further validate the robustness of our model, we will investigate its performance on an external validation dataset. This dataset will include float cycles that the network has not previously encountered either in the training or in the testing phase, offering an assessment of its predictive capability in novel scenarios.

5.3.1 Training

We divided the dataset into three subsets: training, testing, and validation. The training set was used for model training and parameter optimization. The testing set was utilized to evaluate the model's performance on unseen data and assess its generalization ability. Finally, the validation set was employed for hyperparameter tuning and model selection. The dataset was randomly partitioned, ensuring that each subset contained a representative distribution of the overall data characteristics. The sizes of the training, testing, and validation sets were chosen as 80, 10, and 10. This approach enabled us to assess and validate the performance of our model effectively. Moreover, before operating this partition, a few float instruments have been selected and all of their measurements have been excluded from both the training, test, and validation set. These samples will be used as an external validation dataset. The metrics and the performances over this external validation dataset are a more effective indicator of the generalization capabilities of the PPCon model with respect to the metrics on the test set.

In order to train the NN model efficiently, the input dataset is partitioned into minibatches, where each minibatch contains 32 samples. The batch size, which determines the number of samples processed before updating the model weights, is a hyperparameter that must be set prior to training. By processing multiple samples in a minibatch,

	#samples	#epochs	batchsize	dropout rate	λ	α
nitrate	2337	50	32	0.2	0.001	0.001
chlorophyll	3189	150	32	0.2	0.0001	0.0001
BBP	3942	100	32	0.2	$1e^{-7}$	$1e^{-7}$

TABLE 5.3: Summary of PPCon hyperparameters and dataset sizes for all inferred variables.

the model can update its parameters more frequently, which can lead to faster convergence and improved generalization performance ([276]). Once all the mini-batches have been processed by the optimization algorithm, the model has completed an epoch of training.

Adadelta ([277]) is the algorithm that is selected as the optimizer for training the network due to its ability to dynamically adapt over time using only first-order information. This method eliminates the need for manual tuning of the learning rate and has been found to exhibit robustness in the face of noisy gradient information, various data modalities, different model architecture choices, and hyperparameter selection.

Let us recall that the PPCon architecture includes a 1D CNN and four MLPs, which convert point-wise input into a vector form suitable for use by the CNN. The MLPs and the CNN component of PPCon were trained using the same optimizer, with concurrent weight updates across all networks. This approach enables the joint learning of optimal information transfer from point-wise input to vector form, as well as the accurate generation of predicted profiles based on the input tensor.

To accelerate the training process, the model was trained using a GPU (graphics processing unit), which allowed for parallelized computation of the forward and backward passes.

The model's performance was evaluated once every 25 epoch by assessing its ability to predict outcomes on the test set, which consists of previously unseen data. To prevent overfitting and minimize computational burden, we introduced an early stopping routine. Specifically, the training was interrupted if the error metrics on the validation set increased for two consecutive evaluations (i.e., after 50 epochs of training). The final model selected was the one trained prior to the two consecutive test loss increases.

5.3.2 Experimental Settings

Since each output variable has intrinsic differences in training set size, range of values, and profile shapes and variabilities, a separate hyperparameter tuning step is performed for each of them. These hyperparameters were tuned using a systematic search over a range of values, guided by the performance of the model on a held-out validation set. To avoid overfitting in the test set, we employed cross-validation techniques to estimate the generalization performance of the model and selected the hyperparameters that yielded the best performance.

The hyperparameters used for training the three PPCon architectures are summarized in Table 5.3, together with the size of the dataset, the total number of epochs performed, and the batch size dimension which have been already discussed in previous sections.

In our experiments, we applied a dropout rate of 0.2, which was consistent across all trained models. This means that during training, each neuron in the NN has a 20% chance of being randomly excluded from the computation. Dropout regularization is a technique used to prevent overfitting by encouraging each neuron to encode information independently, thereby inhibiting co-dependencies among neurons.

Table 5.3 also reports the multiplicative factors that determine the relative contributions of different elements that compose the loss function defined in Section 5.2.3. The values of these hyperparameters vary depending on the variable being inferred, as these variables have different orders of magnitude and result in RMSE values that vary in magnitude as well. It is crucial to accurately balance the regularization term, governed by λ, and the smoothness term, governed by α, to prevent them from dominating the loss function's RMSE component. The optimal values reported in Table 5.3 guarantee a good and smooth prediction of the vertical profile.

The last implementation detail to be addressed concerns the creation of vectors used to feed the PPCon architecture. As previously discussed, vectorial inputs of different natures are fed into the CNN component of PPCon: firstly, the outputs of an MLP architecture; secondly, vectors representing input variables (temperature, salinity, oxygen) at different depths. To ensure that all input vectors have the same length, we adopted the following strategy: (i) the output and input variables have been interpolated on a regular grid of size 200, and (ii) the output of MLPs have the same length and discretization of the input variable vectors. For nitrate, we considered a depth range of 0 to 1000 meters with an interpolation interval of 5 meters, whereas, for chlorophyll and BBP, we considered a depth range of 0 to 200 meters with an interpolation interval of 1 meter. Then, we set the output layer dimension of the MLP to 200 to ensure that all input vectors have the same length. As a result, the final dimension of the input tensor is 7 (the number of inputs) x 200 (the length of the input vector) x the number of dimensions in the training set.

	North West Med.	South West Med.	Tyrrhenian	Ionian	Levantine
Latitude	$40°N - 45°N$	$32°N - 40°N$	$37°N - 45°N$	$30°N - 45°N$	$30°N - 37°N$
Longitude	$-2°E - 9.5°E$	$-2°E - 9.5°E$	$9.5°E - 15°E$	$14°E - 22°E$	$22°E - 36°E$

TABLE 5.4: Geographical limits of the five areas in which is divided the Mediterranean for the posterior analysis.

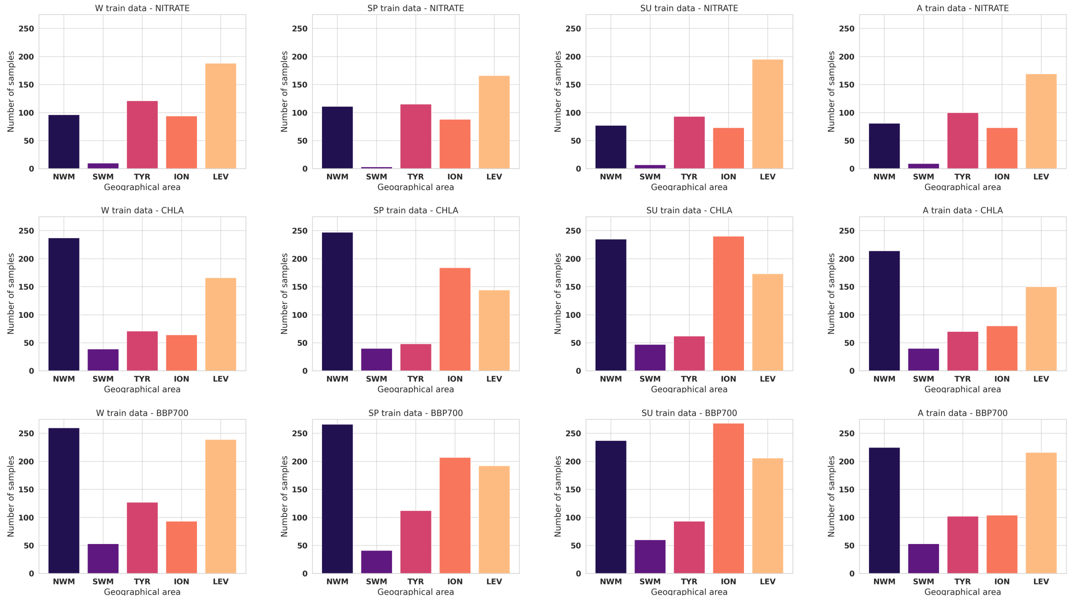


FIGURE 5.2: Histograms representing the distribution of the training samples over different season and different geographic areas.

5.3.3 A Posterior Validation Analysis of PPCon

To validate the PPCon architecture, we will conduct a thorough analysis of its average performance in different geographic areas and across various seasons. The choice underlying this investigation originates from the fact that diverse geographic areas and seasons are known to possess distinct profile properties, such as the shape of the vertical profiles (e.g., depth and slope of the nitracline, or depth and intensity of DCM) and the values at the surface and in deep water. Our goal is to evaluate the model's ability to accurately capture these variations. We want to stress again that the model is trained on the entire dataset, and this division is purely for a posteriori analysis of the performances. This a posteriori analysis of the performances has the objective to identify possible influence and bias of the uneven spatial and temporal distribution of the profiles on the performance of the PPcon model.

Five different geographic areas are considered, namely: Northern West Mediterranean, Southern West Mediterranean, Tyrrhenian, Ionian, and Levantine. The geographical limits related to these variables are reported in Table 5.4. Moreover, Figure 4.6 reports a visual representation of those limits, displayed directly on the Mediterranean Sea area.

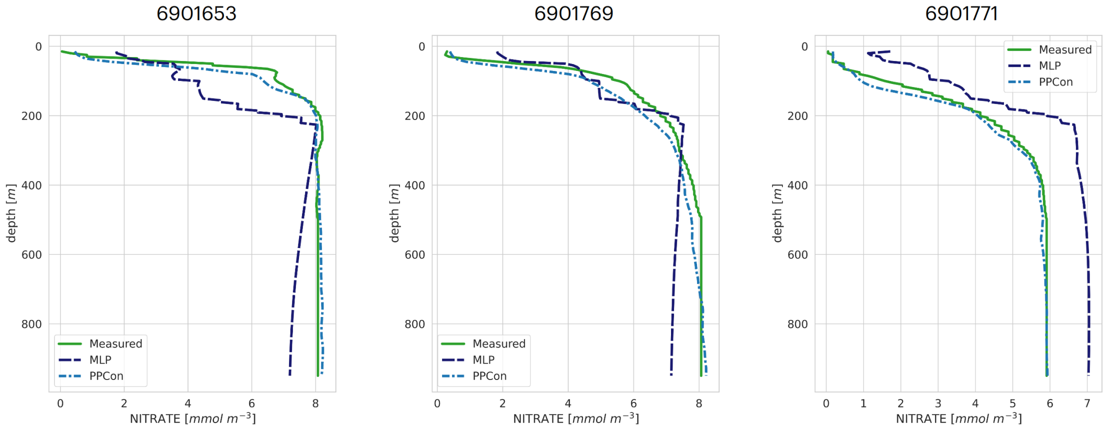


FIGURE 5.3: Profiles of nitrate for some selected floats (WMO numbers in the title) and dates.
Comparison between measured profile (green lines), MLP reconstruction (azure dashed lines), and PPCon reconstruction (blue dash-dotted lines). Profiles are from the subset used for the test.

	WMO	Date	Latitude	Longitude
Nitrate	6901653	24/07/2014	41.85	4.55
	6901769	20/12/2016	39.11	10.96
	6901771	18/10/2015	36.01	20.12
Chlorophyll	6902901	10/10/2019	42.91	7.62
	6901776	13/04/2014	42.79	7.01
	6901773	19/09/2017	32.93	31.24
bbp700	6901657	13/02/2019	39.80	7.23
	6901776	20/09/2014	42.76	7.25
	6903240	20/05/2018	42.20	28.57

TABLE 5.5: WMO, date, and geolocation of the float profiles reported in Figure $1-3$.

To gain a comprehensive understanding of how the model performs across different seasons, we conducted an analysis of four distinct time periods: winter (January to March), spring (April to June), summer (July to September), and autumn (October to December).

Figure 5.2 offers valuable insight into the distribution and variability of the training samples specific to each geographic region and season. This information can be particularly useful for post-analysis purposes, as it sheds light on the amount of information that the PPCon architecture possesses for a given region and season and such insights are essential for justifying the model's predictive capabilities.

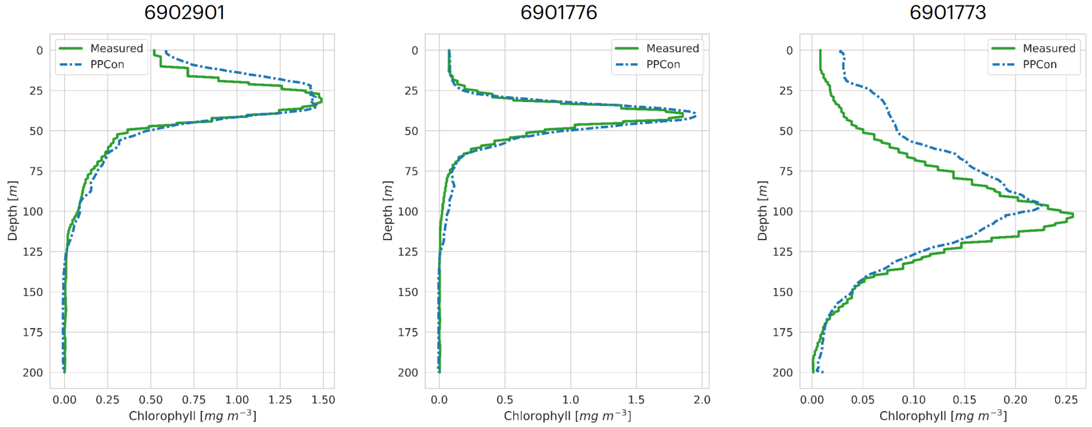


FIGURE 5.4: Profiles of chlorophyll for some selected floats (WMO numbers in the title) and dates.
Comparison between measured profile (green lines) and PPCon reconstruction (blue dashed lines). Profiles are from the subset used for the test.

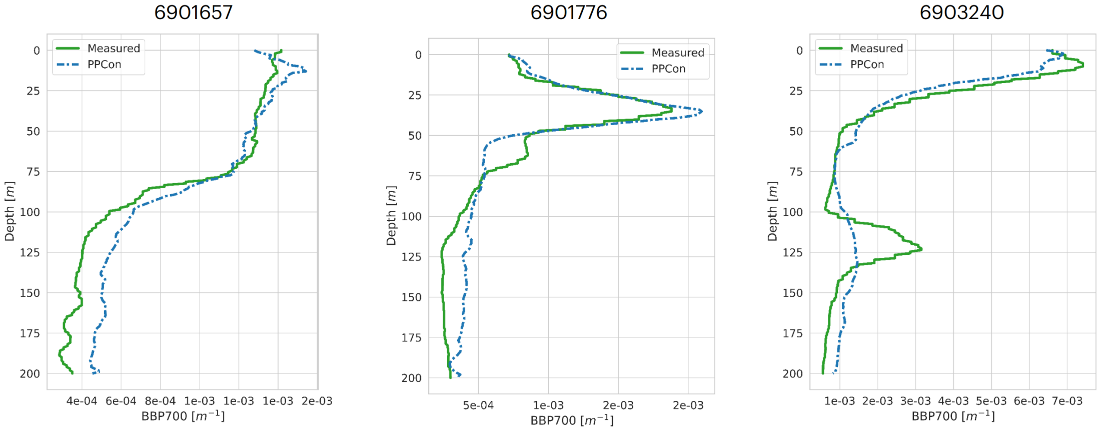


FIGURE 5.5: Profiles of bbp700 for some selected floats (WMO numbers in the title) and dates.
Comparison between measured profile (green lines) and PPCon reconstruction (blue dashed lines). Profiles are from the subset used for the test.

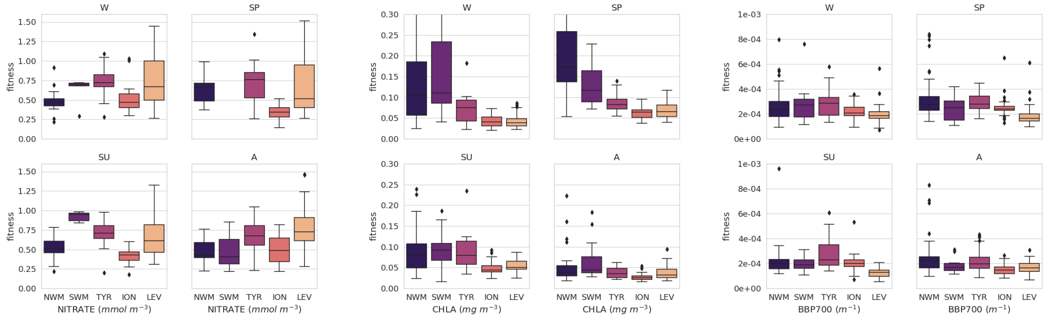


FIGURE 5.6: Boxplot representing the RMSE of PPCon prediction in the test set.
The box represents the RMSE distribution computed above all the profiles belonging to a fixed geographic area and a fixed season

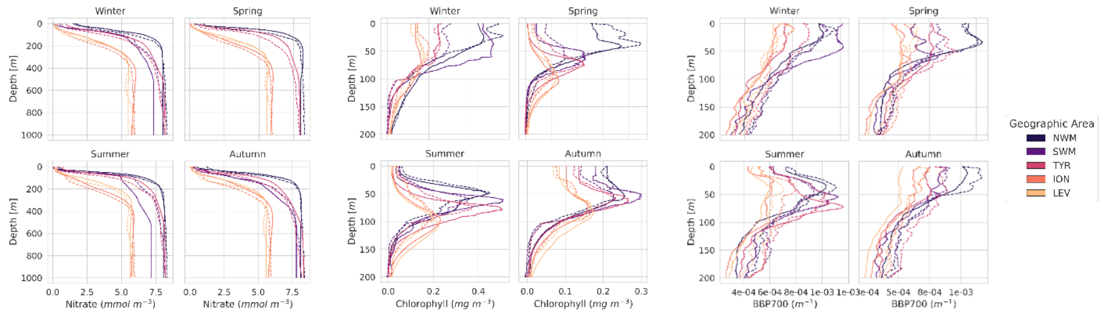


FIGURE 5.7: Comparison of the mean of PPCon predicted profiles with the mean measured by the float instruments in the test set.
The mean is computed over all the profiles belonging to a fixed geographic area and a fixed season

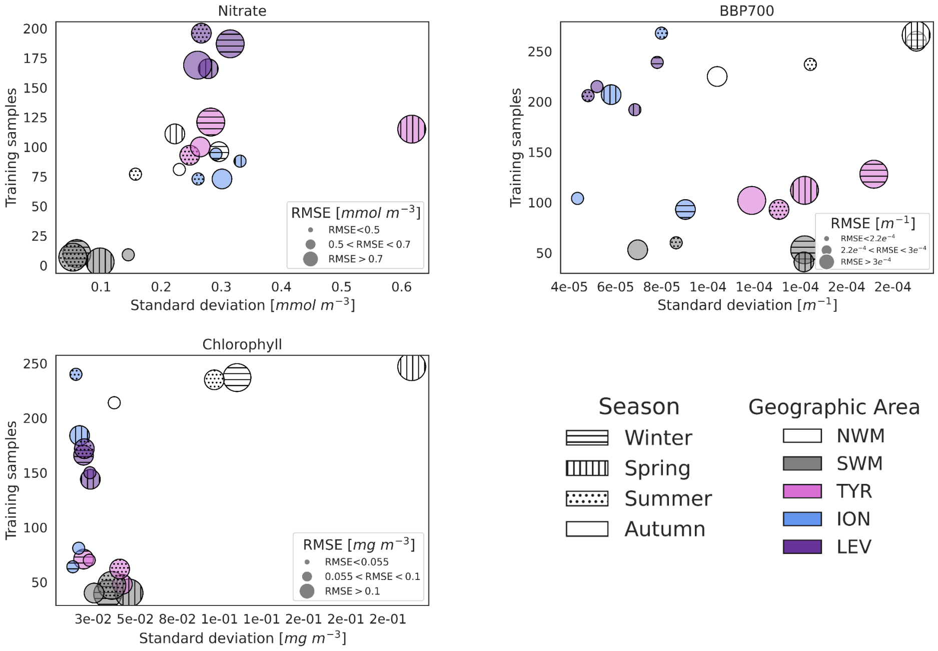


FIGURE 5.8: Plot of the RMSE distribution with respect to the data variability and training dataset size
The different sub-areas are displayed through different colors of the symbols and different seasons are displayed through different symbol fill patterns. RMSE values are categorized by the size of the symbols.

5.4 Results

This section presents the results of the PPCon model in predicting nitrate, chlorophyll, and bbp700 profiles. The effectiveness of the model is evaluated by presenting both quantitative skill metrics (i.e., RMSE) and visual representations of the predicted profiles based on the test set.

Specifically, we will assess the PPCon performance over different seasonal variations (Table 5.6), and different geographic areas (Table 5.7). The absence of overfitting is

		Nitrate	Chlorophyll	bbp700
Winter	Train	0.640	0.082	$2.561e^{-4}$
	Test	0.665	0.098	$2.571e^{-4}$
Spring	Train	0.591	0.129	$3.101e^{-4}$
	Test	0.610	0.134	$2.691e^{-4}$
Summer	Train	0.570	0.069	$1.871e^{-4}$
	Test	0.615	0.067	$1.941e^{-4}$
Autumn	Train	0.631	0.045	$2.211e^{-4}$
	Test	0.640	0.045	$1.971e^{-4}$

Table 5.6: RMSE calculated between the float measurements and the reconstructed values obtained from the PPCon architecture for all variables inferred. This metric is evaluated individually for the train and test sets. The RMSE is computed for different seasons of the year (described in Section 5.1).

		Nitrate	Chlorophyll	bbp700
North Western Med	Train	0.514	0.119	$2.661e^{-4}$
	Test	0.531	0.139	$2.641e^{-4}$
South Western Med	Train	0.637	0.103	$2.501e^{-4}$
	Test	0.643	0.098	$2.581e^{-4}$
Tyrrhenian	Train	0.706	0.074	$3.871e^{-4}$
	Test	0.704	0.075	$2.831e^{-4}$
Ionian	Train	0.426	0.051	$2.051e^{-4}$
	Test	0.445	0.047	$2.081e^{-4}$
Levantine	Train	0.680	0.055	$1.731e^{-4}$
	Test	0.718	0.051	$1.791e^{-4}$

Table 5.7: RMSE calculated between the float measurements and the reconstructed values obtained from the PPCon architecture for all variables inferred. This metric is evaluated individually for the train and test sets. The RMSE is computed for different geographic areas (described in Section 5.1).

supported by reporting the RMSE for both the training and test sets, which exhibit not-dissimilar values.

In terms of performances across different geographic areas (Table 5.7), it can be seen that the lowest RMSE values for chlorophyll and bbp700 are in the eastern sub-basins, while for nitrate the lowest and highest values are in the two eastern sub-basins. Notably, the prediction accuracy for nitrate is significantly higher in the Ionian Basin, with RMSE values below 0.5. Considering the temporal evolution of RMSE values (Table 5.6), the highest values of chlorophyll and bbp700 are in spring and winter, which appears reasonable given the higher variability of vertical pattern during these seasons ([254, 255]). Errors for nitrate are pretty homogeneous among the seasons, which are the highest during winter (e.g., vertical mixing season) and the lowest values during the stratification seasons (i.e., spring and summer). As for chlorophyll, the western basin of the Mediterranean shows higher RMSE values. This can be attributed to the naturally elevated chlorophyll levels observed in that specific area, which consequently lead to higher RMSE values.

For each variable investigated, we present three instances of vertical profile reconstruction using the PPCon architecture, compared to the profile measured by the float instrument whose corresponding identification number is indicated above each profile. To ensure geographic and seasonal diversity, we have selected profiles representing different regions, including at least one from the West Mediterranean and one from the East Mediterranean. Figure 5.3-5.5 displays examples of, respectively, reconstructed nitrate, chlorophyll, and bbp700 profiles. For the nitrate variable, is also reported the reconstruction performed by the MLP model ([1]). The information related to these profiles, such as the date and geolocation of sampling, are reported in Table 5.5.

The obtained results confirm the quality of the profiles generated by the PPCon architecture, which appears to better reconstruct the shape and smoothness of the profiles than the previous MLP architecture. Indeed, PPCon is able to capture different profile shapes associated with different geographic and seasonal conditions, as clearly demonstrated by the predicted nitrate and chlorophyll profiles. Higher quality in the prediction is achieved for the nitrate variable, followed by chlorophyll, and last by bbp700. This outcome is expected, as the nitrate variable exhibits lower variability in the values and profile shapes than the other two variables. For a more detailed analysis of the behavior of the PPCon architecture quality of the predicted profiles, Figure 5.7 reports a comparison between the mean of PPCon predicted profiles and the mean of profiles measured by the float instruments (in the test set) for all investigated variables, providing a more specific insight on the PPCon performances in different geographic areas and seasons.

Figure 5.6 displays boxplots of the RMSE obtained by the PPCon model when predicting profiles belonging to the test set. Each individual box within the figure signifies the distribution of RMSE values computed across profiles within a consistent geographic area and a fixed season. These results confirm what is already observed with Tables 5.6-5.7. The quality of the prediction is lower in the Southern West Mediterranean, and this is coherent with the dataset training samples distributions displayed in Figure 5.2. One interesting observation that arises from these plots regards the uniformity observed among these boxplots across distinct seasons. This uniformity extends to both the relative positioning of different boxes—indicating which area's predictions outperform the others—and the breadth of these boxes—indicating the distribution of RMSE values and the extent to which prediction quality varies across different areas.

Figure 5.7 presents a comparison between the mean values of the PPCon predicted profiles and the mean values of the sampled measurements obtained from the float

Nitrate		Chlorophyll		bbp700	
6901648	0,4288	6901648	0,1739	6901649	2,6231e^{-4}
6901764	0,4583	6901496	0,0878	6900807	1,5071e^{-3}

TABLE 5.8: RMSE calculated between the float measurements and the reconstructed values obtained from the PPCon architecture over the external validation dataset.

instruments in the test set. This comparison encompasses all the variables examined. The mean values are computed based on the profiles within a specific geographic area and season. These profiles serve as additional indicators to assess the reliability of predictions within different frameworks, providing valuable insights into the precision of predictions at various depth levels. These results confirm the previous observations, particularly the finding that the prediction quality is superior for the nitrate variables, followed by chlorophyll, and lastly, bbp700. Additionally, an interesting characteristic of the PPCon prediction is its higher quality in deep water compared to surface water. This can be attributed to the higher variability of profiles in the surface water, making it more challenging for the NN to accurately capture the diverse shapes.

In order to understand the impact of the training set numerosity and of the variability of profiles on the quality of the PPCon predictions we investigated the relation between these quantities and the PPCon error. Specifically, Figure 5.8 reports points whose size corresponds to the prediction RMSE (divided according to the four seasons and the five geographic areas) and their relation with the variability of the training set (on the x-axis), quantified by the standard deviation, and the numerosity of the training set (on the y-axis). This figure offers also valuable insight into the geographical and seasonal distribution of the training d The analysis of the nitrate plot reveals pretty homogeneous errors across all sub-areas and seasons, suggesting a lack of a strong relationship between errors and variability or data availability. In terms of chlorophyll and bbp700, both variables exhibit similar behavior. In particular, data availability appears to have no significant impact on the error, whereas RMSE tends to increase proportionally with the variability.

5.4.1 PPCOn performance over an external validation dataset

For each inferred variable, Figures 5.9-5.11 display Hovmöller diagrams of measured and reconstructed float instruments belonging to the external validation set, and Table 5.8 reports the correspondent RMSE values. This represents a particularly stringent validation test since no one of the profiles measured by these floats have been encountered by the PPCon model during the training or validation phases. The figures compare the in situ float measurements (upper diagram) and the predictions generated by

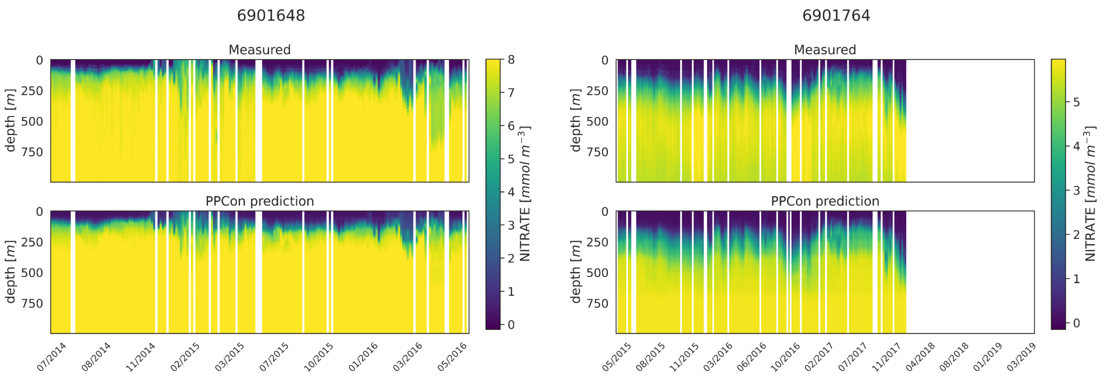


FIGURE 5.9: Hovmöller diagrams for the nitrate of two selected floats (WMO name in the title) belonging to the external validation set.
BGC-Argo measurements (upper panels) and PPCon prediction (lower panels) are compared. WMO 6901648 sampled the $40°N - 42°N$ and $2°E - 6°E$ area during $2014 - 2016$, whereas WMO 691764 sampled the $31°N - 34°N$ and $26°E - 40°E$ area during $2015 - 2017$.

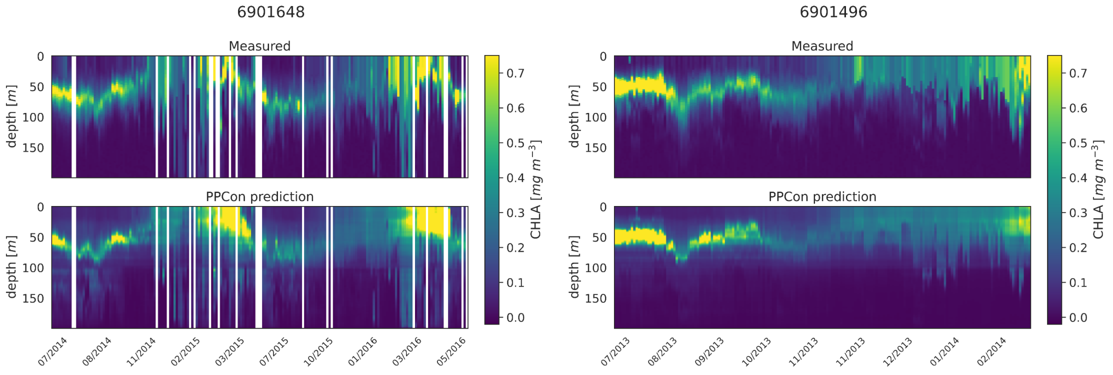


FIGURE 5.10: Hovmöller diagrams for the chlorophyll of two selected floats (WMO name in the title) belonging to the external validation set.
BGC-Argo measurements (upper panels) and PPCon prediction (lower panels) are compared. WMO 6901648 sampled the $40N - 42N$ and $2E - 6E$ area during $2014 - 2016$, whereas WMO 6901496 sampled the $42N - 43N$ and $7E - 12E$ area during $2013 - 2014$.

the PPCon architecture (lower diagram) for floats that have been specifically selected to cover different geographical regions of the Mediterranean Sea (e.g., one in eastern and one in the western Mediterranean Sea) White lines in the diagram indicate float measurements that cannot be compared due to various reasons such as the sensor's momentary inability to measure the specific variable inferred or the absence of one of the inputs necessary for the PPCon architecture (e.g. at least one between temperature, salinity, and oxygen). This could be attributed to limitations in the sensor or unacceptable quality flags associated with the collected data. Nevertheless, the number of profiles that can not be calculated by PPCon is pretty low and it does not degrade the very good capability of the reconstructed profiles to reproduce the temporal evolution of the vertical dynamics shown by the measured floats.

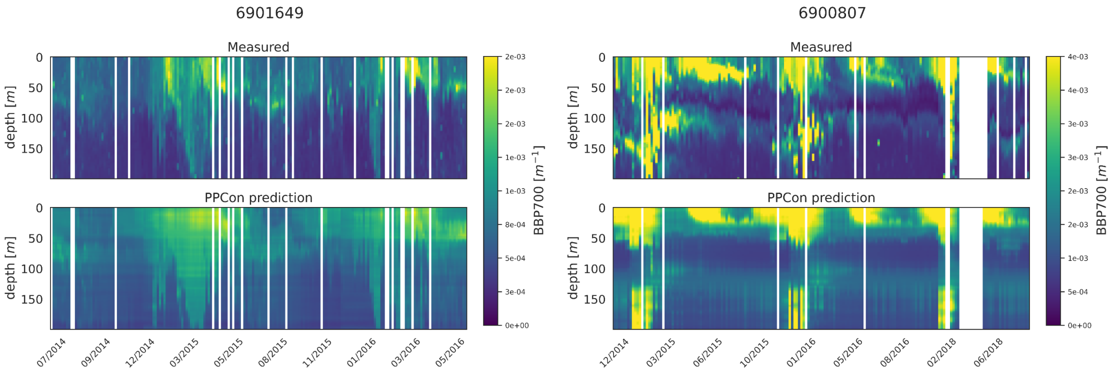


FIGURE 5.11: Hovmöller diagrams for the bbp700 of two selected floats (WMO name in the title) belonging to the external validation set.
BGC-Argo measurements (upper panels) and PPCon prediction (lower panels) are compared. WMO 6901649 sampled the $39°N - 41°N$ and $3°E - 7°E$ area during $2014 - 2016$, whereas WMO 6900807 sampled the $41°N - 44°N$ and $29°E - 35°E$ area during $2014 - 2018$.

These plots confirm the PPCon capability of performing accurate prediction also regarding float devices which are totally unseen by the model. The nitrate (Figure 5.9) reconstructions exhibit a very good performance of PPCon in predicting the vertical dynamics associated with the temporal evolution of the nutricline depth (i.e., the depth at which the sharp increase of the nitrate values is observed), the values in the deep layers (which are different in the sub-areas sampled by the two floats), and the occurrence of deep vertical mixing when surface concentration increase to values higher than $3mmol/m^3$. Particularly impressive is the capability of PPCon to reconstruct the temporal dynamics of chlorophyll (Figure 5.10). The reconstruction effectively captures the evolution of the chlorophyll surface peaks during winter and the formation of the deep chlorophyll maximum during summer in both floats representing the two areas of the Mediterranean Sea. Among the three variables, the bbp700 (Figure 5.11) shows the least accurate predictions. However, the model still displays the ability to infer the key characteristics of the variable's temporal behavior. Nonetheless, the generated predictions for bbp700 appear slightly less detailed compared to the original sampling, indicating a partial limitation of the model in capturing small-scale variations.

Quantitatively, the prediction quality of the PPCon architecture (RMSE values in Table 5.8 are pretty aligned with the metrics calculated over the test set, as indicated in Table 5.7. In particular, nitrate errors of the two floats are pretty homogeneous and 30% lower than the RMSE values of the test set. The errors in chlorophyll and bbp700 predictions exhibit greater variability, with values almost double for the floats in the western Mediterranean with respect to the eastern ones. This is, however, in line with results reported in Table 5.7, and Figure 5.8 where higher errors are associated with higher variability.

5.5 Discussion

To our knowledge, the PPCon architecture is the first attempt to predict vertical BGC-Argo profiles by means of a convolutional architecture. Its primary objective is the incorporation of typical profile shapes during the training phase, in contrast with previous architectures which all relied on MLP architectures and point-wise strategy.

Moreover, the posterior study that we conducted shows that is no significant variation in the error across different geographic areas and seasons of the year (Table 5.6-5.7), confirming that PPCon can be successfully applied to all the float profiles collected in the Mediterranean basin.

Specifically, the PPCon architecture serves as a valuable tool for significantly enriching the BGC-Argo dataset. This became useful as ocean observing systems - while essential for the monitoring of the health of the marine ecosystem ([26]) - have large limitations given their sparse and scarce space-temporal coverage. Surface satellite observations are limited by cloud covers and incomplete swaps of satellite sensors ([278]) while profiling the ocean interior is limited by the capacity of deploying and retrieving sensors and measurements with sufficient coverage. Gap filling and interpolation of satellite observations ([279], [280, 281]) are nowadays consolidated practices to provide gap-free and high-level products ([226, 282]). Our PPCon architecture presents a valuable approach to harness the potential of the Argo and BGC-Argo network by enabling the synthetic generation of essential variables (chlorophyll, nitrate, and bbp700) even when these costly sensors are not present in the deployed floats. The application of PPCon on the GDACs BCG-Argo float dataset (spanning from 2015 to 2020) enabled the generation of 5234 synthetic nitrate profiles, 3879 chlorophyll profiles, and 3307 bbp700 profiles, which means doubling the chlorophyll and bbp700 BGC-Argo profiles and more than triple those of nitrate. Enhancing the float dataset through the inclusion of reconstructed nutrient profiles (and possibly other biogeochemical variables) has been demonstrated successful in observing system simulation experiments ([283, 284]) and in real assimilation numerical experiments (Amadio et al., (2023)). In particular, the assimilation of reconstructed profiles effectively corrects a widespread positive bias observed in the Operational System for Short-Term Forecasting of the Biogeochemistry of the Mediterranean (MedBFM), and the addition of the reconstructed profiles increase the spatial impact of the BCG-argo network from 20% to 45% (Amadio et al., (2023))

5.6 Final Considerations

This Chapter presents a novel approach for reconstructing low-sampled variables, namely nitrate, chlorophyll, and bbp700, using high-sampled variables such as date, geolocation, temperature, salinity, and oxygen. The introduced model, named PPCon, utilizes a spatial-aware 1D CNN architecture that effectively learns the characteristic shape of the vertical profile, enabling precise and smooth reconstructions. PPCon represents a notable advancement over previous techniques relying on MLPs, which operate on point-to-point input and output, making it challenging to generate continuous curves when forecasting complete vertical profiles. The training dataset consists of a collection of BGC-Argo float measurements in the Mediterranean basin. The proposed architecture has been specifically designed to handle both punctual and vectorial input, with careful tuning of the architecture and loss function for the task. An extensive hyperparameter tuning phase has been conducted to ensure the best architecture for each variable.

To evaluate the accuracy of the profiles generated by the PPCon architecture, both quantitative metrics and visual representations of the results have been provided. Additionally, the method has been validated on an external dataset to verify its generability. The results confirm the model's ability to predict high-quality synthetic profiles, with particularly accurate predictions for the nitrate variable, followed by chlorophyll, and lastly, bbp700. The RMSE for nitrate reconstruction is reduced from $RMSE_{MLP} = 0.87$ to $RMSE_{PPCon} = 0.61$. PPCon demonstrates its capability to capture and learn distinct typical shapes in the profiles, which characterize the inferred variables across different seasons and geographic areas. Detailed error analysis confirms the model's robust performance, accounting for seasonal and regional variations, suggesting that PPCon's ability to learn these differences can make it successful for broader-scale training beyond the Mediterranean basin. Furthermore, the model exhibits accurate performance on an external validation dataset, confirming its capability for generalization. The PPCon architecture serves as a valuable tool for significantly enhancing the BGC-Argo dataset, thereby improving our capacity to observe and comprehend the ocean and its changes effectively.

Chapter 6

GANs for Integrating Deterministic Model and Observations.

Improving the capability of monitoring and forecasting the status of the marine ecosystem has important implications (e.g. sustainable approaches to fishing and aquaculture, mitigation of pollution, and eutrophication), especially considering the changes caused by human activities ([26]).

An unprecedented improvement in monitoring the oceans has arisen from satellite sensors in the 90s and in situ autonomous oceanographic instruments. Satellites cover with high resolution the whole marine domain but only at the surface and they suffer from cloud cover. The marine domain can be monitored also through automatic devices, introduced in Chapter 3, i.e. Argo-floats. Float instruments provide profiles of the water column but their number is not sufficient to cover exhaustively the ocean surface. As an example, figure 6.1 shows the spatial distribution of the float profiles collected during 2015 over the entire Mediterranean sea. Hence, observational data available are spatially sparse, and with a scarcity of series spanning more than a few decades.

Deterministic physical-biogeochemical models have been exploited to simulate the marine environment, as they can provide reanalyses and predictions for the whole 3D domain. However, uncertainties in parameterization and input data and high computational costs can impact their reliability and applicability.

The current state-of-the-art deterministic marine ecosystem modeling integrates observations (e.g. satellite ocean color, BGC Argo float) with ocean model through data assimilation methods ([285]). The incorporation of ML techniques offers alternative and stimulant opportunities for advancing the capacity of integrating theory, knowledge, and observations to simulate the marine environment ([158]).

Research Question

The principal research questions addressed in this chapter are as follows:

- Can a DL model integrate diverse data types, such as those provided by deterministic models, satellite, and in-situ observations, effectively leveraging the strengths of each method?
- Is the proposed DL model capable of accurately replicating 3D marine fields? Specifically, does it replicate the patterns of horizontal sections and the typical shapes of vertical profiles, accounting for variations across different geographic areas and seasons?
- Can the DL model learn from observed variables and, through the designed architecture, transfer this knowledge to variables that are not measured?

To evaluate these questions, we will:

- Conduct an experimental analysis to determine if the deep learning approach can effectively assess the spatial and temporal variability of physical and biogeochemical variables.
- Analyze non-in-situ measured variables to see if the architecture can transfer insights from measured variables to them.

Contributions

The DL method proposed in this work is based on the approach of filling missing pixels of a considered image, which is a well-known and extensively studied computer vision task, often referred to as *image inpainting* [286]. Since this method has been created specifically to synthesize visually realistic, coherent, and semantically plausible pixels for missing regions, the idea is to exploit its architecture to assemble a model capable of skilfully reconstructing the physical and biogeochemical variables and also to fill the information gap typical of the inhomogeneity of in-situ and satellite observations. This approach is tested in the Mediterranean Sea, where a rich collection of model, satellite, and in-situ data are already available: a validated model [287], 1km-resolution satellite data from Copernicus [288] and in-situ BGC-Argo floats [255].

Specifically, our method is composed of two DL models. The first DL model, referred to as *EmuMed*, harnesses Generative Adversarial Networks (GAN) ([111]) and is structured upon an inpainting architecture ([133]). *EmuMed* derives spatial and temporal relationships among marine ecosystem variables from the output of the deterministic model *MedBFM*, exploiting the inherent architecture of the DL model. The second DL model, denoted as *InpMed*, extends *EmuMed* by incorporating additional observations while retaining the same architectural framework as *EmuMed*.

We will represent the 3D marine domain by dividing it into horizontal maps at various depths: essentially, the 3D domain is modeled using 2D maps. Therefore, it's crucial to choose an architecture that capitalizes on the spatial horizontal relationships present in variables from the same horizontal field

Based on these considerations a 2D CNN-based architecture seems the more suitable for dealing with this kind of spatial data since these networks conserve the spatial structure of the input.

The central concept underlying our approach was to transform horizontal maps of the designated domain into imagery that mirrors the marine environment, akin to photography. In this context, traditional RGB channels are replaced by channels that represent various marine variables. In standard imagery, the three color channels exhibit a strong interdependence, as they work in tandem to generate a broad spectrum of colors. Analogously, we endeavor to establish an inherent and robust interrelation among marine ecosystem variables, given their natural correlations and dependent on each other.

Then, considering that dealing with in-situ measurements leads also to the aforementioned problem of the insufficient spatial coverage of information, an architecture capable of filling areas where measurements are missing becomes essential. These additional considerations lead us to choose a convolutional architecture specifically constructed to deal with inpainting tasks above horizontal sections of the marine domain.

Challenges

Modeling marine ecosystem variables using DL presents several challenges, which have already been introduced in Chapter 3. Here the ones that influence mostly the choice adopted for the learning architecture and the training are mentioned again, for a more in-depth discussion the reader is remanded to Chapter 3.

- Marine datasets encompass four dimensions, namely: temporal (time), vertical (depth), and two horizontal (latitude and longitude). These dimensions are distinguished by varying scales and units, with horizontal and vertical spatial dimensions measured in kilometers and meters, respectively.

- The number of variables that our model must be able to predict can become extremely large considering the large amount of physical and biogeochemical components that are necessary for a satisfactory understanding of the marine ecosystem dynamics.
- The DL architecture is chosen and designed to be suitable for merging two types of information of different nature, i.e. the output of a deterministic model and datasets of observations. This architecture should be also able to spread local information provided by the in-situ measurement, to update the model not only in the precise location where the measurement has been performed but also around its neighborhood.
- Unlike other ML applications, we cannot rely on ground-truth data. Indeed, the deterministic model is just an approximation of the marine ecosystems, with the observations providing only a very sparse and incomplete picture of it.

Structure of the Chapter

The chapter is structured as follows: in Section 6.1, the particulars of the DL architecture are introduced, encompassing the network structure, the employed loss function, and other essential details aimed at ensuring the complete reproducibility of the proposed model. Section 6.2 provides an overview of the experimental context. This encompasses defining the specific boundaries of the domain under study, its discretization, and the selection of hyperparameters which led to the convergence of the DL model. Proceeding to Section 6.3, the obtained results are presented, both in terms of visual representation of the reproduced domain and through the application of statistical measures. Section 6.4 houses the principal discussions arising from the results obtained. Lastly, Section 6.5 concludes and outlines potential avenues for improvement in future research.

6.1 Material and Method

In this section, we describe our DL model, divided into two parts: *EmuMed* and *InpMed*. The model's architecture, along with details like the training process and loss function, is presented in Section 6.1.1. Section 6.1.2 focuses on the *EmuMed* component. Lastly, we discuss the *InpMed* component in Section 6.1.3.

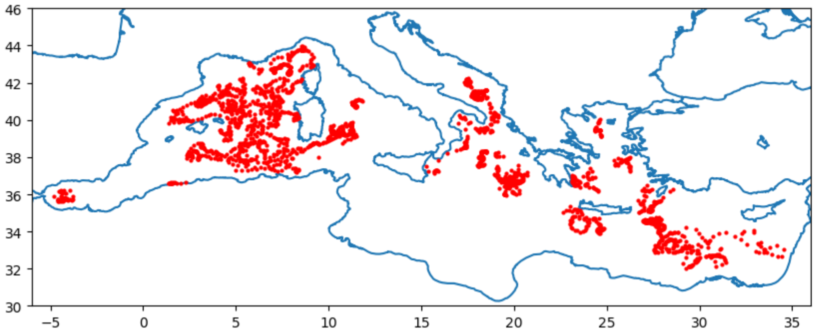


FIGURE 6.1: Map of the *float* measurements over the Mediterranean Sea collected during the year 2015.

6.1.1 DL Architecture

The models introduced in this work are based on a convolutional inpainting architecture [133], which is in turn based on GANs. In this paper, we will consider an inpainting model composed of three interacting convolutional NNs: the *completion network* used to complete the image; the *global discriminator*, and the *local discriminator*, which are two auxiliary networks. The completion and the discriminators compete in a two-player game, where simultaneous improvements are made to both of them during the training phase. Thus, while the completion network learns how to fill the holes in a realistic and coherent way, discriminators are trained to understand whether or not the provided input has been completed. The improvement of the completion implies a betterment of the discriminators' performance; and vice-versa, the improvement of the discriminators' capability to recognize completed input implies a rise in the completion performance, to fool the discriminators.

Completion Network. The completion network is a convolutional NN, consisting of 17 layers, as detailed in [133]. The architecture exploits an encoder-decoder technique that initially decreases the resolution of the input features to reduce the computational effort and then restores the original resolution. Like in the image generation task, the input of the completion network is an RGB image with binary channels, where 1 indicates that a mask is applied to the input pixel, and the output is an RGB image, properly completed.

The core characteristics of this architecture are summarized in Table 6.1. This table shows that the architecture consists of convolutional, deconvolutional, and dilated convolutional layers ([289]). Dilated layers are favored because they expand the area each layer considers as input without adding more trainable parameters, which keeps the computational cost manageable. Dilated convolutions introduce a new parameter

TABLE 6.1: Completion architecture details illustrated in diagram form.

The key attributes of the NN are outlined, encompassing parameters, output size (i.e. number of output channels), as well as any additional layers. More specifically, "BN" denotes the batch normalization layer and "SELU" represents the non-linear selu() activation layer.)

Layer	**Kernel**	**Stride**	**Dilation**	**Output Size**	**Add. Details**
Conv. 2D	5	1	1	64	BN, SELU
Conv. 2D	3	2	1	128	BN, SELU
Conv. 2D	3	1	1	128	BN, SELU
Conv. 2D	3	2	1	256	BN, SELU
Conv. 2D	3	1	1	256	BN, SELU
Conv. 2D	3	1	1	256	BN, SELU
Dilated Conv. 2D	3	1	2	256	BN, SELU
Dilated Conv. 2D	3	1	4	256	BN, SELU
Dilated Conv. 2D	3	1	8	256	BN, SELU
Dilated Conv. 2D	3	1	16	256	BN, SELU
Conv. 2D	3	1	1	256	BN, SELU
Conv. 2D	3	1	1	256	BN, SELU
Deconv. 2D	4	1	2	128	BN, SELU
Conv. 2D	3	1	1	128	BN, SELU
Deconv. 2D	4	1	2	64	BN, SELU
Conv. 2D	3	1	1	32	BN, SELU
Output. 2D	3	1	1	3	SELU

called the dilation rate, which specifies the gap between values in a kernel. From a mathematical point of view, a 2-D dilated convolution can be defined for each pixel as follows ([106]):

$$y(m,n) = \sum_{i=1}^{M}\sum_{j=1}^{N} x(m + r \times i, n + r \times j)w(i,j) \tag{6.1}$$

where $y(m,n)$ is the pixel component of the output of dilated convolution from the pixel component of input $x(m,n)$ and a filter $w(i,j)$ with the length and the width of M and N respectively. The parameter r is the dilation rate. If $r = 1$, a dilated convolution turns into a normal convolution. In dilated convolution, a small kernel with dimensions $k \times k$ is expanded to a larger size, specifically $k + (k-1)(r-1)$, using a dilated stride of r. This expansion permits the flexible gathering of multi-scale contextual information while maintaining the original resolution of the input.

In contrast to other architectures that employ multiple pooling layers to reduce resolution, our network model reduces resolution only twice. It achieves this by utilizing strided convolutions, reducing the image size to $\frac{1}{4}$ of the original. This approach is crucial for preserving sharp textures in the missing regions, avoiding the issue of blurriness ([290]).

TABLE 6.2: Local Discriminator architecture details illustrated in diagram form.

The key attributes of the NN are outlined, encompassing parameters, output size (i.e. number of output channels), as well as any additional layers. More specifically, ”BN” denotes the batch normalization layer and ”RELU” represents the non-linear ReLU() activation layer.

Layer	Kernel	Stride	Padding	Output Size	Add. Details
Conv. 2D	5	2	2	64	BN, RELU
Conv. 2D	5	2	2	128	BN, RELU
Conv. 2D	5	2	2	256	BN, RELU
Conv. 2D	5	2	2	512	BN, RELU
Conv. 2D	5	2	2	512	BN, RELU
Fully Conn.	-	-	-	1024	RELU

TABLE 6.3: Global Discriminator architecture details illustrated in diagram form.

The key attributes of the NN are outlined, encompassing parameters, output size (i.e. number of output channels), as well as any additional layers. More specifically, ”BN” denotes the batch normalization layer and ”RELU” represents the non-linear ReLU() activation layer.

Layer	Kernel	Stride	Padding	Output Size	Add. Details
Conv. 2D	5	2	2	64	BN, RELU
Conv. 2D	5	2	2	128	BN, RELU
Conv. 2D	5	2	2	256	BN, RELU
Conv. 2D	5	2	2	512	BN, RELU
Conv. 2D	5	2	2	512	BN, RELU
Conv. 2D	5	2	2	512	BN, RELU
Fully Conn.	-	-	-	1024	RELU

Discriminator Networks. Two discriminators play against the completion network introduced above: the global discriminator and the local discriminator. The former tests the reliability of the input in its entirety, while the latter focuses on smaller areas, thus paying more attention to details. The discriminators take as input the complete image (adequately re-scaled), both of them are implemented using convolutional NNs followed by a fully connected layer producing a real-valued vector as output. The detailed architecture selected for the discriminators is reported in Table 6.2-6.3.

The global context discriminator takes the entire image as input. It’s composed of six convolutional layers and one fully-connected layer, which generates a single 1024-dimensional vector. The local context discriminator follows a similar pattern, but it takes in a smaller image patch, specifically the initial input resolution is half of that for the global discriminator, and thus the first layer used in the global discriminator isn’t needed here. In both cases, the output is a 1024-dimensional vector representing the context information either globally or locally around the completed region.

Finally, the two resulting vectors are concatenated and passed again as input of a fully-convolutional layer, that returns a continuous value indicating the probability that the provided input is real or fake.

Training. The completion network accepts both an image and a mask as input, with the mask using binary values to indicate which pixels need to be filled in. Prior to commencing the pixel completion task, the first step of the completion network involves replacing the region in the training image that requires completion with a constant color. This color corresponds to the mean pixel values found within the entire dataset. This preliminary step significantly enhances the NN performance.

The loss function employed to train the completion network, introduced in [135], is the weighted MSE defined as follows:

$$L(x, M_c) = ||M_c \odot (C(x, M_c) - x)||^2 \tag{6.2}$$

where M_c is the completion region mask, x is the input image, $C(x, M_c)$ denote the completion network, $\odot$ stands for the pixel-wise multiplication and $|| \cdot ||$ is the Euclidean norm. Furthermore, the GAN loss [111] is used for training together completion and discriminators network. While discriminators aim to maximize the average of the log probability of real images and the log of the inverse probability for fake images, the generator aims to minimize the log of the inverse probability predicted by the discriminator for fake images. Therefore, the generator tries to minimize the following function while the discriminator tries to maximize it:

$$\min_{C} \max_{D} \mathbf{E}[\log D(x, M_d) + log(1 - D(C(x, M_c), M_c))] \tag{6.3}$$

where M_d is a random mask, $D(x, M_d)$ is the discriminator's estimate of the probability for the real input x with mask M_d to be real, $D(C(x, M_c))$ is the discriminator's estimate of the probability for the fake input x to be real, and $\mathbf{E}$ indicate the average over the training input. Finally, taking in account both Equation 6.2 and Equation 6.3, the resulting loss function is:

$$\min_{C} \max_{D} \mathbf{E}[L(x, M_c) + \log D(x, M_d) + \alpha log(1 - D(C(x, M_c), M_c))] \tag{6.4}$$

where α is a fixed hyperparameter. The training of the algorithm, which is schematized in Algorithm 2, consists of three main phases: during **phase 1** the completion network is trained among all the features of the training set for T_C epochs; then, during **phase 2**, the completion network is fixed and the discriminator network is trained for T_D epochs;

Algorithm 2 Pseudocode of the training steps for the DL architecture underlying *EmuMed*, and, consequently *InpMed*.

```
for $t = 0 \ldots T_C$ do                                              ▷ phase 1
    for all $x$ in the training set do
        Generate masks $M_c$ with random holes.
        Compute $C(x, M_c)$.
        Update completion network weights through Equation 6.2.
    end for
end for
for $t = 0 \ldots T_D$ do                                              ▷ phase 2
    for all $x$ in the training set do
        Generate masks $M_c$ with random holes.
        Compute $D(x, M_d)$ and $D(C(x, M_c), M_c)$.
        Update discriminator network weights through binary cross entropy loss
    end for
end for
for $t = 0 \ldots T_{CD}$ do                                           ▷ phase 3
    for all $x$ in the training set do
        Generate masks $M_c$ with random holes.
        Generate masks $M_d$ with random holes.
        Compute $D(x, M_d)$ and $D(C(x, M_c), M_c)$.
        Update discriminator network weights through binary cross entropy loss.
        Update completion network weights through Equation 6.4.
    end for
end for
```

finally, during **phase 3** both the completion network and the discriminators are trained at the same time for T_{CD} epochs.

6.1.2 EmuMed

The *EmuMed* is the first-step model that we present in this paper, named after the fact that it behaves as an emulator (meaning that it is learning information from) of the deterministic biogeochemical marine model *MedBFM* [38, 287]. The architecture underlying *EmuMed* is the one presented in Section 6.1.1: a generative convolutional NN trained through adversarial loss. The input (tensors) employed for the training are obtained from a discretization of data generated through a simulation of the deterministic model. These tensors represent 2-dimensional maps of a fixed region of the Mediterranean Sea (here, with the term map we denote a horizontal section of the region at a fixed depth). The role that pixels accomplished for the image completion task is carried out by rectangles that represent a discretization area of the Mediterranean Sea, while the standard RGB channels are substituted with channels that contain values representing the oceanographic physical and the biogeochemical variables that we aim to reproduce. Thus, *EmuMed* consists of a generative model capable of reconstructing the relationship between biogeochemical variables for the Mediterranean Sea domain considered. Among the strengths of this DL model with respect to the classic deterministic one, we annoverate, firstly, that the convolutional intrinsic structure combined with the choice of assigning a channel to each marine variable, allows the creation of an intrinsic link between them, that leads to a capability of capture biological and chemical interactions. Notice that *EmuMed* is significantly less computationally expensive compared to the deterministic model, as obtaining a prediction for a certain variable does not require a complete simulation.

6.1.3 InpMed

InpMed is the second step of the model presented in this chapter, obtained starting from *EmuMed* and then performing a further training phase adding both in-situ measurements collected by the float devices and by satellite sensors. This additional training phase is performed according, again, to phase 1 of the Algorithm 2 described in Section 6.1.1. The weights of *EmuMed* are updated to fit these new data.

In this model, the loss is computed once more employing Equation 6.2. However, in this instance, the mask is applied in all regions except where observational data is present. As a result, the model undergoes an automatic enhancement in its performance solely based on the available observational data. The reduction in the loss signifies the model's increasing capability to simulate observations effectively.

InpMed ensures an improvement in the simulating capability of the model, as the convolutional structure guarantees the distribution of information provided by observation.

Another crucial point is that there are certain marine indicators that are not measured either through in-situ devices or via satellite sensors, such as the primary production, which prediction can be improved anyway by taking advantage of a combination of the ML architecture and of the observed data. In fact, thanks to the convolutional structure, observations concerning measured variables (e.g. temperature, oxygen, and so on) do not only update the model weights among their channel but improve the prediction with respect to all the variables considered including the unobserved ones. Relations between variables are learned through the training of *EmuMed*; subsequently, *InpMed* exploits the information provided by the measured variables and the relations learned from the deterministic model to improve the prediction for both the measured variables and the ones that cannot be measured.

TABLE 6.4: Experimental settings values concerning the DL architecture. Horizontal lines separate: common parameters, *EmuMed* parameters and *InpMed* parameters.

Parameter	Value
Completion input size	$30 \times 65 \times 75$
Local Discriminator input size	$20 \times 50 \times 50$
lr_c	0.01
lr_d	0.01
T_c	5000
T_d	200
T_{cd}	1000
α	4×10^{-4}
lr_{float}	0.001
T_{float}	200

TABLE 6.5: Experimental settings values concerning the definition of the input structure.

Parameter	Value
time interval	weekly
latitude interval	$36° - 44°$
longitude interval	$2° - 9°$
depth interval	$0 - 600$ m
time resolution	weekly
latitude resolution	12 km
longitude resolution	12 km
depth resolution	20 m

6.2 Experimental Setting

The geographical area considered in this work is the western Mediterranean portion, specifically, the one with latitude ranging between 36 and 44 and longitude varying between 2 and 9 and the vertical dimension covers a depth ranging from 0 to 600 meters. It consists of the portion between southern France and northern Africa, delimited on the east limit by Corsica and Sardinia and on the west limit by Balearic islands. The spatial resolution is 12 km in both latitude and longitude axes and 20 meters in the vertical one. The time period covers the year 2015 which is discretized on a weekly basis. Therefore, the 4-dimensional input tensor consists of a horizontal map, of the central-western Mediterranean Sea area, whose dimensions are: length, height, width, and channel. Each channel of the tensor, in turn, collects one marine ecosystem variable. Namely, the variables considered are temperature, salinity, oxygen, chlorophyll-a, and primary production. All the variables can be obtained via the deterministic model *MedBFM*, while only the first four are collected via float measurement (Anywayas it is not possible to measure primary production via any sensor), and only chlorophyll-a

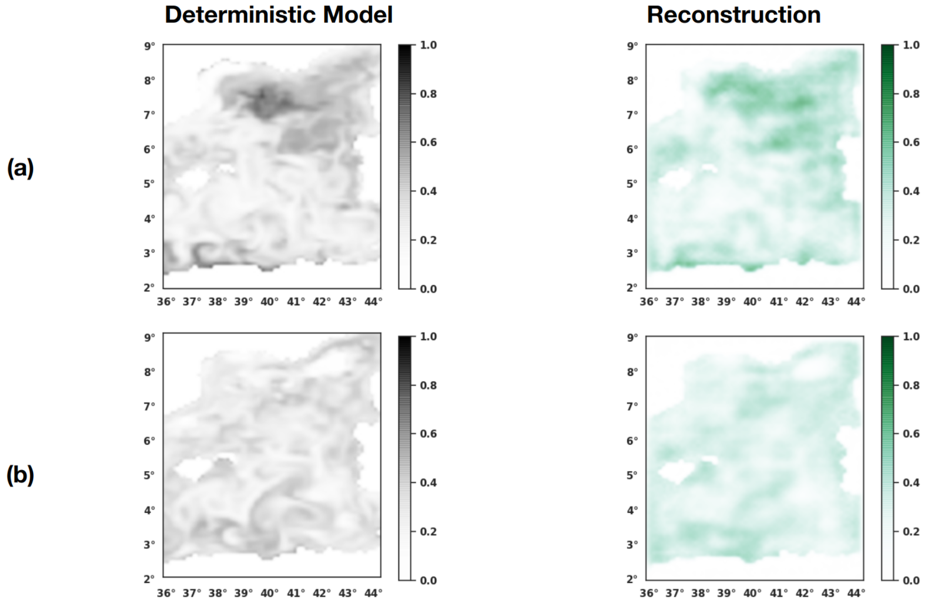


FIGURE 6.2: Surface chlorophyll maps for weeks 21 and 33: (a) generated by *MedBFM* and (b) reconstructed via DL model.

can be inferred through satellite. Each location in the 3D field has 5 variables associated and a time resolution of one week is used (thus, 52 weekly observations are available). Due to their nature, each float provides a 1D profile, where latitude and longitude are fixed and only the depth can change. Finally, since satellites can observe only the surface of the water, they provide 2D data (with spatial gaps due to cloud cover). Thereby, the 5 variables are positioned in a 3D field and 52 weekly snapshots are available, float observations are a given number of 1D profiles where latitude and longitude are fixed and only depth changes and satellite observations are 2D maps of the surface with holes due to the cloud covers. For the training of *EmuMed*, we used the following hyper-parameters (summarized in Table 6.4): the completion network is trained for 5000 epochs, the discriminator network is trained for 200 epochs, and finally the two networks are trained simultaneously for 1000 epochs. These hyperparameters have been fixed after an appropriate preliminary study. The optimizer chosen is ADADELTA [277], which sets the learning rate for each weight in the network automatically. The learning rate initial value for both the completion and the discriminator is set to 0.01. Subsequently, *InpMed* is trained for 200 epochs with a learning rate value set initially to 0.01.

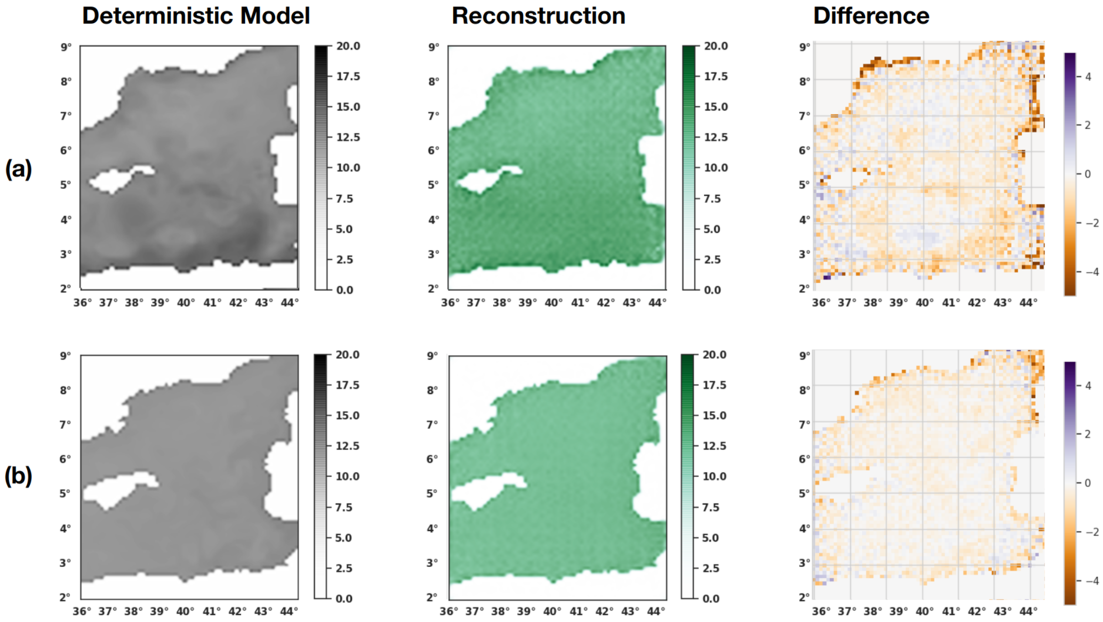


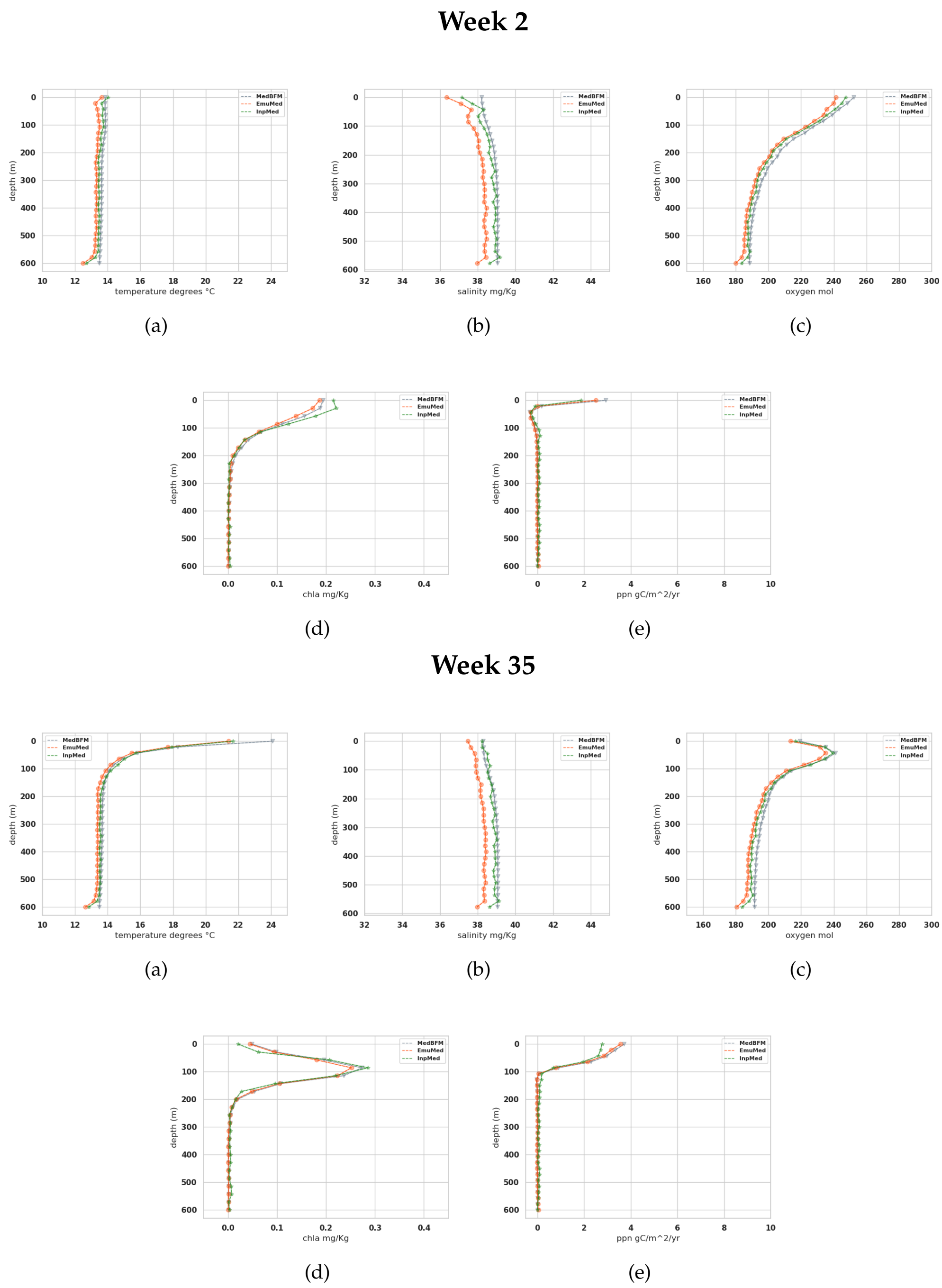
FIGURE 6.3: Temperature maps at 60 meters and 400 meters depths: (a) produced by *MedBFM*, (b) reconstructed using the DL model, and the difference between them

6.3 Experimental Results

To assess the goodness of the proposed reconstruction, the analysis focuses on the quality of the reconstruction of the map of some of the inferred variables and on some statistical properties (averages and variances) of the simulated fields.

In particular, Figure 6.2 reports maps of the surface chlorophyll demonstrating *InpMed* capability to emulate the intense spatial variability of surface chlorophyll fields. Given the significant variability observed in horizontal surface chlorophyll layers—a variability that surpasses other variables studied—we selected chlorophyll to evaluate the DL model's proficiency in mimicking spatial distribution. To assess the DL model's efficacy in capturing the diverse patterns of the chlorophyll variable across seasons, we present two maps corresponding to distinct times of the year. One map depicts the chlorophyll surface during week 21 (May) and the other during week 33 (August). The May map reveals pronounced values in the Gulf of Lion and along the Algerian current in the southern region. In contrast, the August map illustrates the typical summer conditions, marked by subdued values and smaller patches. The DL model (as seen in the right column) effectively replicates these characteristics.

Figure 6.3 reports another visual representation of the results, depicting a comparison of temperature maps generated by the deterministic model and the ones generated by the DL model proposed. To further investigate the model's behavior at different depth levels, the comparison includes maps illustrating temperature distribution in the



(a) (b) (c)

(d) (e)

(a) (b) (c)

(d) (e)

FIGURE 6.4: Vertical profiles of the spatial averages over the considered domain. Variables represented are: (a) temperature, (b) salinity, (c) oxygen, (d) chlorophyll, (e) primary production.
The gray line represents the *MedBFM* model predictions, the orange line represents the *EmuMed* predictions, and the green represents the *InpMed* predictions. First row reports results relative to week 2 (winter), second row is relative to week 35 (summer).

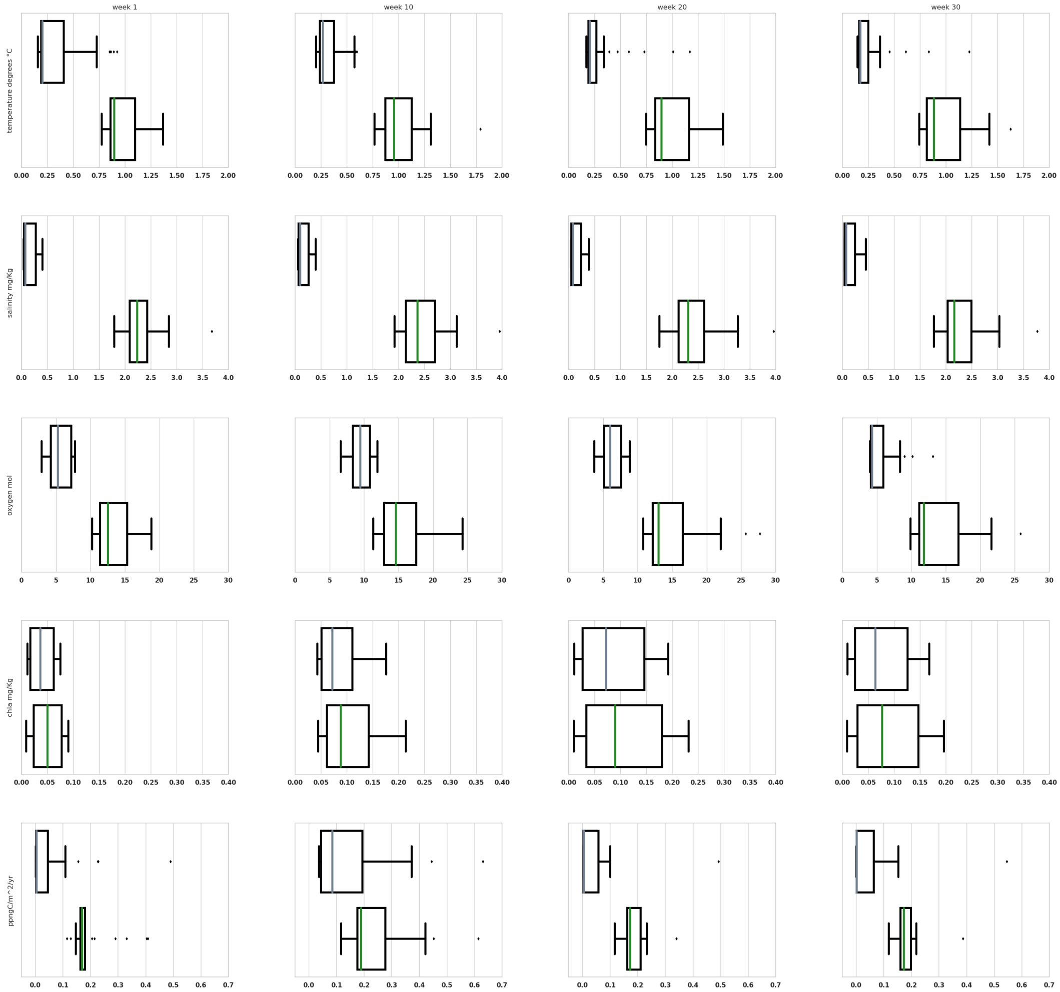


FIGURE 6.5: Box-plots showing the distributions of the standard deviation of the marine variables computed from the spatial maps at different depths.
From top to bottom are shown, respectively, temperature, salinity, oxygen, chlorophyll, and primary production. For the first three, the mean is computed over 600*m*, and for the last two is computed over the first 200*m*. The box plot with the median gray line represents *MedBFM*, while the box plot with the green median line represents *InpMed*. From left to right are shown, respectively, week 1, week 10, week 20, and week 30.

uppermost layer (approximately 60 meters) and in a deeper layer (around 400 meters). To provide a clearer view of the distinctions between the deterministic predictions and the DL model's outputs, the figure also presents a map highlighting the differences between the two. The higher the difference, the higher the uncertainty of the DL model in emulating the deterministic one.

Figure 6.4 shows the vertical profiles of the spatial averages among two given weeks of the year (e.g., one in winter and one in summer), assessing the capability of the technique to simulate different periods for all modeled variables. These plots compare the original deterministic model *MedBFM* profile with the *EmuMed* and *InpMed* reconstructions, showing the benefits provided by the different architectural components. The benefit can be judged by the fact that *InpMed* distances *EmuMed* showing that information from observations was integrated.

Finally, Figure 6.5 compares, via box-plot, the distributions of the standard deviation of the spatial variability of *MedBFM* and *InpMed*. Four weeks of the year are displayed in order to show how the spatial heterogeneity of the marine proprieties varies throughout the year and how it is successfully handled by *InpMed*. In particular the differences show the impact of the integration of new information from observations.

6.4 Discussion

Regarding the model's ability to infer distinct horizontal surface layers, performance was evaluated using a variable (chlorophyll) characterized by complex and high horizontal variations. The results affirm that the DL model effectively captures the distinct horizontal distributions typical of different periods of the year. The DL model accurately discerns the elevated presence of chlorophyll during spring compared to autumn, and it also identifies the regions where chlorophyll concentration is higher in both seasons. However, it's worth noting that the DL model appears to generate images with smoother patches than the deterministic model. This observation may suggest the need for enhancing the model's capacity to infer fine details in the horizontal component.

As for the model's capacity to replicate the horizontal field of variables at different depth levels, the results obtained with temperature first indicate that the model successfully reproduces variations in domain shape at varying depths. An analysis of difference distribution highlights that the DL model is more adept at replicating the distribution of variables with depth than that observed at the surface. This is likely attributed to the fact that, in the Mediterranean Sea, the behavior of water temperature changes with depth, adhering to a pattern commonly observed in many oceans and seas. The uppermost layer of the Mediterranean Sea, often referred to as the mixed layer, experiences the most pronounced temperature variations. This layer is directly affected by solar radiation and is subject to seasonal fluctuations. In contrast, the deeper layers of the Mediterranean Sea exhibit a higher level of temperature stability. The rate of temperature change with depth becomes significantly slower in this region. This observation applies not only to temperature but also extends to all the predicted variables. Another observation arising from these plots is the DL model's relatively weaker performance near the coastlines: in fact, even if the overall structure looks quite realistic, the difference between the deterministic model and the one proposed increases significantly in the coastal areas, where the variability is higher. This effect is a consequence of the employed architecture, as learning near the image borders in CNNs poses more significant challenges due to reduced receptive fields. Again, this observation applies to all the predicted variables.

Results show that the *EmuMed* has learned well to reproduce the typical mean vertical profiles, simulated by *MedBFM*, for all variables but salinity (Figure 6.4). Deviations of *InpMed* from *EmuMed* profiles (e.g., orange and green lines in oxygen, chlorophyll, and primary production in Figure 6.4) highlight how the inclusion of the observations in the ML architecture introduced possible corrections (i.e., new information) to the *MedBFM* simulated fields. Temperature shows that the inclusion of observations had a marginal effect while salinity shows that observations bring *InpMed* profiles closer to *MedBFM* highlighting a possible inaccurate reconstruction of *EmuMed*, that anyway is corrected in the second phase of the training, confirming the added values of the two-step architecture implemented in *InpMed*. Regarding the spatial variability of horizontal fields of marine variables, maps of Figure 6.2 show qualitatively the good performance of the ML reconstruction. From a quantitative point of view, the comparison of the standard deviation boxplots (Figure 6.5) shows that the spatial variability of *InpMed* is generally higher than *MedBFM* for all variables in all selected weeks. This highlights that, when observations are included in the reconstruction, the ML model *InpMed* simulates horizontal fields characterized by more spatial structures, possibly due to the effect of the integration of observations.

A separate comment can be done for primary production (i.e., a variable that is not observed). Despite the mean vertical profiles are not substantially changed by ML architecture (Figure 6.4), it is possible to notice that the *InpMed* model differs from the *EmuMed* even if, during the training, no observed data have been provided for primary production. The fact that the variability introduced by the observed variable is clearly propagated by *InpMed* can be also observed in Figure 6.5. This evidence confirms that information provided by observed data has an effect on the *InpMed* capability to simulate the unobserved variable, thanks to the relations among variables learned from the output of the deterministic model by the DL architecture exploited.

In terms of the overall structure, employing multiple channels in CNNs allows for the simultaneous prediction of various biogeochemical variables. This approach enhances the capacity of the DL model to gather a broader range of information and to distribute that knowledge among different variables. To the best of our understanding, this represents the first endeavor to model multiple biogeochemical variables simultaneously, while also merging the results of a deterministic model and observations concurrently.

In the current phase, we have opted for a 2D CNN instead of a 3D CNN mainly for computational efficiency, though this choice doesn't permit the most effective dissemination of information across different depth levels. To fully exploit the potential for learning and make the most of the available information, it would be worthwhile to consider the use of a 3D architecture.

6.5 Final Considerations

We investigated the integration of an existing ecosystem deterministic model with in-situ and satellite information through a convolutional generative DL architecture. This case study limited to a restricted area of the Mediterranean aims to demonstrate the reliability of ML reconstruction for marine ecosystem variables and to show how the different components of the ML architecture contribute to the reconstruction quality. Merging these two different kinds of information allows us to combine their strengths and exploit them to lessen each other's limits. We remark that a DL model can also be less computationally expensive, once trained, with respect to a deterministic one, as it does not require an entire simulation in order to get specific variable estimations (e.g., primary production). Such a comparison will be one of the aspects that will be investigated in future studies. Moreover, exploiting the intrinsic structure of the architecture, the learning framework makes possible the spread of information provided from the observed variables (temperature, salinity, oxygen, chlorophyll) also to variables that are not possible to directly collect via in-situ or satellite measurements (e.g., primary production). Experimental results on both *EmuMed* and *ImpMed* have confirmed the validity of the proposed approach, showing that our models can infer correctly information from the deterministic model and, in the case of *ImpMed*, also from observations. This work represents the first step to exploiting DL architecture aimed at merging large deterministic model output with observations to reconstruct the marine ecosystem's temporal and spatial variability. Our main goal will be to extend this architecture by inserting a larger number of channels, exploiting this architecture to model unobserved variables (as we did for primary production). The extension of the present ML model to the entire Mediterranean Sea represents another important computational challenge given the significant increase in data volume to handle.

Chapter 7

Discussion and Conclusion

In recent years, our understanding of ocean state and variability has witnessed significant improvements. This progression is attributed to three factors: an unprecedented capability of getting observations, novel data-driven methods (particularly NN), and a substantial boost in computing resources.

The oceanography field today benefits from an abundance of data sources: deterministic model-generated outputs, satellite-derived measurements, and in-situ oceanographic instrument readings. These vast data sources naturally lend themselves to the application of DL methods. The significance of data in training DL models is evident – a richer dataset typically leads to an improvement in the model's accuracy.

However, the vastness of oceanographic data also brings its set of challenges. Deterministic models provide a comprehensive view of physical and biogeochemical variables across the marine domain. Yet, these models, being simulations, can sometimes be inaccurate due to uncertainties in their formulation and parameterization. On the other hand, while observational data is often more accurate, it is constrained in terms of spatial and temporal coverage.

Dealing with this complex set of oceanographic data requires innovative solutions. This thesis proposes and evaluates various NN architectures to address these challenges, focsed on the biogeochemistry of the Mediterranean Sea.

In Chapter 4, is presented a method that enables the prediction of low-sampled biogeochemical variables using scalar point-based inputs. In contrast, Chapter 5 proposes a transition to vectorial data, resolving the same challenge of the previous chapter but directly predicting vertical profiles. Finally, Chapter 6 expands the dimensionality by employing a 2D CNN to predict outcomes in a 3D domain.

7.1 Final Discussion

The ability to predict low-sampled variables from high-sampled ones offers a significant opportunity. It not only enhances our understanding of biogeochemical variables within the Mediterranean Sea but also unlocks the full potential of the observing systems.

For instance, within cruise-based measurements and datasets derived from Argo floats, variables like temperature, oxygen, and salinity are more frequently observed. Predicting additional marine variables based on these frequent observations can provide comprehensive datasets, substantially enriching our knowledge of the Mediterranean's biogeochemistry.

However, the structure of datasets containing observations varies depending on how observations are collected. Given this variability, it's crucial to design architectures that optimally extract the information within the input features. The goal is to maximize the potential of their characteristics and provide meaningful predictions for low-sampled variables.

When it comes to predicting low-sampled variables using input from a cruise measurements dataset (such as EMODnet, described in Chapter 4), an MLP approach is essential due to the dataset's scalar nature.

This MLP methodology can also be effectively applied to Argo-float datasets. By leveraging individual entries from the vertical profiles as scalar input features, we can predict an entire vertical profile. Starting with Argo-floats equipped with sensors for temperature, salinity, and oxygen, it becomes feasible to predict profiles of additional variables such as nitrate, phosphate, silicate, alkalinity, chlorophyll, and ammonium.

A visual representation of profiles predicted by the MLP model described in Chapter 4 is illustrated in Figure 5.3, which shows a comparison between measured profiles, MLP reconstruction and PPCon reconstruction.

The MLP model's predictions exhibit overall good performance, closely capturing the characteristic shape of nitrate profiles. Moreover, the model reproduces the expected nitrate values within the deeper layers in 2 out of 3 examples.

In addition to the visual comparison, an evaluation of the skill metrics for the reconstructed BGC-Argo nitrate profiles confirms the MLP model's capability in inferring vertical profiles. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) calculated for approximately 2200 nitrate profiles spanning the period from 2015 to

2020 amount to 0.75 and 0.87, respectively. These values closely resemble those computed on the test set data, underscoring the model's quality and its capacity to generalize effectively to an independent and more extensive dataset.

However, upon closer inspection, a potential weakness of the model is evident: the reconstructed profiles are not as smooth as expected.

This irregularity may result from the model being trained on punctual data, which could lead to a lack of awareness of the typical shape of biogeochemical variable profiles that it aims to infer.

In light of this limitation, an alternative prediction model has been developed, the PPCon introduced in Chapter 5, explicitly tailored for Argo float data. This novel model naturally incorporates information concerning profile shapes in the learning process.

This approach harnesses Convolutional DL architectures, specifically 1D CNN, which are inherently spatially-aware structures adept at processing structured data. By capitalizing on the spatial awareness embedded within CNNs, the model becomes better equipped to capture the distinctive shapes and seamless variations present in marine variable profiles.

There are notable distinctions between the two approaches: MLPs were trained on cruise data, which are known to be more precise in collecting variables than autonomous sensors such as the BGC-Argo ([248, 291]). However, while MLP architectures can provide good training and test errors ([1, 169, 215, 216]), they have been found to exhibit higher errors when predicting BGC-Argo profiles ([1]). In contrast, the PPCon architecture, which relies directly on BGC-Argo float measurements for the training, showed very good test and external validation performances.

However, it must be noted that an intrinsic measurement error is introduced by the higher uncertainty of the variables measured throughout the autonomous sensors. We alleviated this limitation by using only DT and high-quality checked Argo and BGC-Argo floats data, however, the use of the present PPCon in operational oceanography ([254, 292]) should be considered cautiously given the lower reliability of Adjusted or near real-time (NRT) Argo data. According to the analysis conducted by [293], the BGC-Argo float data for oxygen, nitrate, and chlorophyll concentrations exhibit RMSE values evaluated at $5.1 \pm 0.8 \mu mol/kg$, $0.25 \pm 0.07 \mu mol/kg$, and $0.03 \pm 0.01 mg/m^3$, respectively. We have demonstrated that the RMSE for the PPCon architecture is 0.61. Therefore, a research question pertains to understanding how the measurement error of the float instrument impacts the performance of the PPCon architecture, and how to estimate an overall error that combines the contribution of the instrument error and the error associated with the PPCon.

Although both MLPs and PPCon employ similar input information (date, geolocation, temperature, oxygen, and salinity), their treatment of this data differs significantly. MLPs process the input and output as discrete data points, while PPCon utilizes vector representations of the vertical profiles. This approach was necessitated to effectively exploit the potential of a 1D CNN, which intrinsically preserves the characteristic profile shape of the input and output variables [65]. When comparing the predictive performance of these techniques in generating vertical profiles from float data, distinct differences emerge. MLPs tend to produce profiles characterized by apparently artificial discontinuity and jumps, while the profiles generated by PPCon exhibit a smoother and more realistic appearance (Figure 5.3). This improvement is confirmed also by the RMSE, which is lower when using the PPCon model ($RMSE_{PPCon} = 0.61$) compared to the state of the art of MLP architectures ($RMSE_{MLP} = 0.87$ according to [1]).

Additional significant observations regarding these two methods become evident. While, on one hand, the MLP model can be harnessed to predict biogeochemical variables using input data derived from cruise measurements (i.e., discrete inputs) and from the Argo dataset, the PPCon model allows for the prediction of variables exclusively utilizing vertical profiles from the Argo dataset as input. The PPCon can not predict starting from punctual input, thus it can not be applied to datasets such as EMODNet.

While the PPCon model uses variable profile inputs to predict a variable profile output, the inpainting model introduced in Chapter 6 expands this concept to the 3D component. In fact, profiles of BCG-Argo are integrated with a previously trained 3D emulator to predict a 3D field of biogeochemical variables.

Both architectures leverage CNNs: the former uses a 1D structure, while the latter incorporates multiple 2D configurations to represent the horizontal maps of the 3D domain.

This design choice stems from the mutual goal of integrating information from neighboring variables through convolutional operators. In the first architecture, the CNN is tailored to model the vertical relationship (depth). Conversely, in the second, the emphasis is on the horizontal relationship (latitude and longitude), with vertical dimension consistency ensured by the input, namely the deterministic model and the profiles.

7.2 Conclusion and Future Works

The use of DL in oceanography has witnessed significant growth in recent years. However, most applications have concentrated on the physical variables of the problem rather than biogeochemistry, leaving this scientific area relatively underexplored.

The methodologies presented in this thesis cater to a diverse range of data types and structures. By adopting a flexible approach, we can harness the full potential of the available oceanographic data. The results obtained demonstrate not only the feasibility but also the effectiveness of using NN in this domain. The predictions generated are both accurate and reliable, making them invaluable for further oceanographic research and applications.

The thesis seeks to convey an insight that different tasks, datasets, and dimensions of the problem necessitate distinct DL architectures. It provides empirical evidence of the potential of DL models to address various oceanographic challenges, such as bridging the gap between sparsely and densely sampled variables and discovering innovative ways to integrate observations with deterministic models.

Throughout, we highlight the importance of multivariate relationships among marine biogeochemical variables. Harnessing these relationships has enhanced predictive capabilities. This is evident in the first two chapters, where certain variables were used to predict others. It is also demonstrated in the latter chapters, where these relationships facilitated the transfer of information from observations to other variables.

Despite the advancements and successes highlighted in this thesis, challenges persist and there is space for improvements in the design of the models.

The PPCOn architecture, while adept at predicting smoother vertical profiles, is constrained by the number of predictable variables due to the limitations of float instruments. On the other hand, the MLP model allows for a broader biogeochemical indicator prediction spectrum. A comprehensive model should expand its predictive scope beyond the Argo dataset, possibly harnessing theoretical relationships or extracting properties from more expansive ship-based datasets, integrating them within convolutional paradigms.

Moreover, the reiteration of the DA procedure with the PPCOn dataset is necessary. The only reason why we performed the DA study using the MLP-generated dataset was chronological. Given PPCOn's enhanced synthetic profile quality, we believe that DA using the PPCOn-generated dataset will provide better results.

The inpainting method's architecture, which predicts the 3D domain via 2D horizontal maps at variable depths, might impede vertical information propagation. A possibility that needs to be investigated in the future is the use of a 3D model, which consequently can take as input the whole 3D marine domain, avoiding the loss of the information provided by the vertical link between variables. While computationally demanding, the potential model enhancements could be significant.

Additionally, an embedding of the PPCOn and inpainting models is worth exploring. The PPCOn model, by enhancing the dataset with vertical biogeochemical profiles, could serve as a valuable training augmenter for the InpMed model. This amalgamation could refine deterministic model corrections and amplify efficacy.

In closing, while this thesis provides a comprehensive exploration of NNs in oceanographic data analysis, the journey doesn't end here. The vastness and complexity of the oceans demand continuous exploration and innovation.

Appendix A

Combining Neural Networks and Data Assimilation

The Array for Real-time Geostrophic Oceanography (Argo) programme appears to be one of the better examples of countries and human resource capacities in working together to provide global data coverage [253] that supports the investigation of the present (analysis), future (forecast) and past years (reanalysis) ocean state conditions. In the last 10 years, the increase of in-situ observations from autonomous platforms (Johnson et al. 2013 and Johnson and Claustre 2016) has opened up new perspectives for biogeochemical oceanographers. Indeed, BGC-Argo (biogeochemical Argo, 2023) has yielded new insights in describing the interior of the global ocean [295] and key processes such as the deep chlorophyll maximum (Mignot et al. 2014, Barbieux et al. 2019, D'ortenzio et al. 2020, Ricour et al. 2021, and Barbieux et al. 2022), oxygen and [300], nutrients vertical fluxes (Taillandier et al. 2020 and Wang et al. 2021) and carbon exports (Dall'Olmo and Mork 2014 and Wang and Fennel 2023).

Among the BGC sensors, oxygen (O2) is currently the most common measured variable, with approximately 270,000 profiles worldwide until now, which is double that of suspended particles and chlorophyll and more than four times those of nitrate, downwelling irradiance, and pH (https://biogeochemical-argo.org). Currently, the Global Data Assembly Centres (GDACs, e.g., Coriolis, NOAA) distributes oxygen data in Real Time (RT) Adjusted (AM) and Delayed Mode (DM). The quality of AM data is controlled within 24 hours using internationally agreed and automatic quality-control procedures performed at the surface, along the entire vertical profiles and along the trajectory [305]. DM data are generally distributed a few months later (nearly six months) in more rigorous form [269].

Major efforts have been devoted to improve the long-term reliability and accuracy of autonomous O2 measurements [215]. As described in [306] and in [307], raw oxygen data from floats can have errors of up to 20% in terms of oxygen saturation (at the surface) due to sensor drift during storage out of the water. By improving the accuracy up to 5-10%, 1st-order correction methods can correct storage drift by multiplying the oxygen concentrations with a gain factor derived from a reference dataset [308]. Additionally, oxygen measurements are calibrated using values of saturation at the surface (in water for the older ones and in air for floats with new sensors, Bushinsky et al. 2016). Despite correction and calibration progress, [307] and [309] found a drift in about 25% (mean drift -0.07% per year) and 70 % (mean -0.12 % per year) of analyzed floats, respectively. Positive and negative values of drift were found in [248] and [310]. Therefore, the development and dissemination of a post-deployment quality control (QC) method has been encouraged to avoid spurious results [311] and to distinguish between ocean signals or trends (e.g., deoxygenation) and potential optode drift.

The combination of in-situ observations and numerical models represents a promising approach to exploit the BGC-Argo potentialities. Indeed, BGC-Argo data are used for several BGC modelling tasks such as: (i) model tuning [311], (ii) validation (Terzić et al. 2019, Salon et al. 2019 and Wang et al. 2021), and (iii) assessment of the BGC ocean state and variability through data assimilation (DA) (Cossarini et al. 2019 and Teruzzi et al. 2021 d'Ortenzio et al. 2021). DA underpins decades of progress in ocean prediction [241] by increasing the type and number of observational datasets and associated uncertainty into a prediction model, and solving problems connected to uneven distribution and scarcity of the observations [256].

Given their capacity to approximate continuous functions [316], NN algorithms match specific DA tasks such as bias correction (Kumar et al. 2015 and Zhou et al. 2021), cross calibration [319], reformulation of observation operators [320] and new product creation or dataset reconstruction [319]. As an example, ocean color (OC) datasets were employed to test Multi-Layer Perceptrons (MLP, namely the most common NN) by retrieving past and long-term BGC time-series (Martinez et al. 2020, Martinez et al. 2020, Roussillon et al. 2023). Moreover, in [282], MLP serves to infer vertical BGC distribution from OC. High performance in predicting biogeochemical states (e.g., oxygen) from physical profiling floats measurements were achieved in [323] for the Black Sea.

In [215], an MLP-NN is used to approximate nutrient concentration and carbonate system from physical and oxygen profiles, and the updated version of [169] allows refining the previous work with the so-called Canyon-b NN method. A configuration to adapt global Canyon-b NN in the Mediterranean Sea region is developed by [216]. A further update of the application of the MLP method in the Mediterranean Sea is provided in

[1], by achieving a lower error in the predictions through a larger training dataset, a hyperparameter refinement and a two-step quality control of the input data.

In the context of operational oceanography, the biogeochemical modelling component of the Copernicus Marine Service for the Mediterranean Sea (MedBFM) provides analysis, short term forecast [313] and long term reanalysis [287], including the assimilation of satellites Ocean Colour (OC) and BGC-Argo observations [313].

The variational assimilation scheme, 3DVarBio, of MedBFM has evolved over time by including a greater number of observation types and variables. Starting from the first release [324], the assimilation has progressively included coastal OC observations [258], chlorophyll and nitrate profiles from BGC-Argo (Cossarini et al. 2019 and Teruzzi et al. 2021 respectively). Given the growing availability of O2 from BGC-Argo, in this paper we propose an additional upgrade of the MedBFM to include BGC-Argo oxygen assimilation, with a novel post-deployment quality control, and the integration of NN reconstructed profiles in the assimilation scheme.

The constant evolution of the observation networks and assimilation capacities requires an updated understanding of the impact of observation on the numerical model result [325]. This can be achieved by using the numerical assimilative models in Observing System Experiments (OSEs) where the impact of existing observations on the model performance is assessed [326].

In this paper, our OSE experiment, that combines data assimilation and neural network in a sequential modular approach (hereafter NN-MLP-MED), aims at quantifying how Argo and BGC-Argo networks can be exploited. Spatial and temporal impacts of the OSE have been evaluated using classic and new skill performance metrics in three two-year (2017-2018) numerical experiments performed using the MedBFM coupled with the 3DVarBio: a control run (HIND) without assimilation; a multivariate run (DAfl) with assimilation of BGC-Argo chlorophyll, nitrate, and oxygen; and a multivariate run that also assimilates in-situ observations and reconstructed ones (DAnn).

Given its particular characteristics and the high density of BGC-Argo profiles, the Mediterranean Sea represents an ideal site for OSE experiments to evaluate the potentiality of the BGC-profiles assimilation. The Mediterranean Sea is an anti-estuarine semi-enclosed sea characterized by specific physical and biogeochemical dynamics [327], with a complex horizontal circulation consisting of mesoscale and sub-basins scale gyre structures, transitional cyclonic and anticyclonic gyres and eddies, influenced by bathymetric features interconnected by currents and jets [328]. Despite its relatively limited extent in the mid-latitude temperate zone, the Mediterranean Sea has a considerable

BGC variability that can be roughly approximated in an oligotrophic West-East gradient with low nutrient availability at the surface, insufficient to sustain significant phytoplankton biomass (Siokou-Frangou et al. 2010 and Marañón et al. 2021) and a deeper nitracline in the east (¿120m) with respect to the west (¡100m). Additionally, chlorophyll has a particular seasonal cycle with pronounced winter/early spring surface blooms only in the western part and a few locations in the eastern part. During summer, a deep chlorophyll maximum follows the stratified and oligotrophic conditions at increasing depth moving eastward (¿100m at East and ¡100m in the West) [255]. Dissolved oxygen has a subsurface maximum at about 50m, with higher values in the west (partly due to the dependence of oxygen solubility on temperature). Noticeable differences are observed in the intermediate layers where the oxygen minimum ranges between 300 (west) and 1000 m (east) [331].

The paper is organized as follows. After a brief presentation of the OSE approach, each component and the experimental setup are described in detail (Section 2). In the following section (Section 3), we describe the results of the novel NN-MLP-MED and the assimilation simulations by using different skill metrics to assess model capability in reproducing the main biogeochemical seasonal dynamics. A discussion of some key issues involved in the NN and DA is provided in Section 4, then the paper closes with some final remarks (Section 5).

A.1 Methods

A novel combined Neural Network (NN-MLP-MED) and Data Assimilation (3DVar-Bio) approach is included in the Mediterranean MedBFM model system to integrate BGC-Argo and reconstructed profiles into biogeochemical simulations of the Mediterranean Sea.

Our OSE experiment is based on a sequential modular approach [256] consisting of a post-deployment quality control method of O2, hereafter QC O2 procedure, a trained multi-layer perceptron NN [1] and a data assimilation scheme (the 3DVarBio variational scheme of MedBFM, Figure A.1).

The input of the first two modules, QC O2 and NN-MLP-MED are the BGC-Argo and Argo datasets, while the final 3DVarBio module also takes the enhanced datasets as input: quality checked O2 (qcO2) and reconstructed nitrate (recNO3, Figure A.1).

In the following sections, the novel modules of the MedBFM system (i.e., the QC O2 procedure and the NN-MLP-MED scheme) and the dataset (BGC-Argo and reconstructed datasets) are described together with the revised 3D-VarBio.

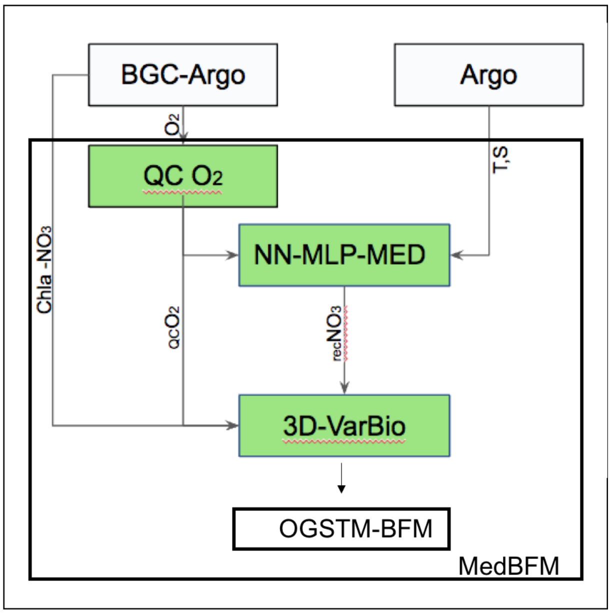


FIGURE A.1: Flowchart of the NN-MLP-MED and DA approach.
In green boxes: the modules. In plain boxes: the datasets. Arrows refers to Argo (temperature and salinity) and BGC Argo profiles of chlorophyll (Chla), oxygen (qcO_2) nitrate (NO_3) and reconstructed nitrate ($recNO_3$)

A.1.1 The regional model for the Mediterranean Sea, MedBFM

The MedBFM consists of the OGS transport model (OGSTM) based on the OPA 8.1 system [332] and updated according to the [333] and [334] versions, the BFM, Biogeochemical Flux Model described in [335] and [336], and the 3DVarBio variational assimilation scheme as in [324] and [258].

OGSTM solves advection, diffusion, sinking terms and the free surface and variable volume-layer effects on the transport of tracers [313], and it is forced by the output (current, T, S and sea surface height) of the NEMO3.2 model. OGSTM and NEMO3.2 share the same bathymetry and z* grid configuration, open boundary and river conditions [337].

The Biogeochemical Flux Model, BFM, is a biomass and functional group based marine ecosystem model. BFM solves governing equations for nine living-organic state variables (diatoms, autotrophic nanoflagellates, picophytoplankton, dinoflagellates, carnivorous and omnivorous mesozooplankton, bacteria, heterotrophic nanoflagellates, and microzooplankton, macro-nutrients (nitrate, phosphate, silicate and ammonium) and labile, semi-labile, and refractory organic matter and oxygen. In addition, the BFM includes a carbonate system model (Cossarini et al. 2015 and Canu et al. 2015).

A.1.2 3DVarBio data assimilation scheme

Based on 3DVarBio (Teruzzi et al. 2014, Teruzzi et al. 2018, Cossarini et al. 2019 and Teruzzi et al. 2021), the assimilation scheme adopted in the present work integrates oxygen, chlorophyll and nitrate to update all the assimilated variables as well as all the phytoplankton biomasses and phosphate.

The 3DVarBio is a variational data assimilation scheme [324] based on the minimization of a cost function *(J)* which relies on the misfit between the model background (x_b) and the observations (y) weighted with the respective error covariance matrices (B and R) as follows:

$$J(x_a) = (x_a - x_b)^T B^{-1} (x_a - x_b) + (y - H(x_b))^T R^{-1} (y - H(x_b)) \tag{A.1}$$

Here, the observation operator (H) maps the values of model background state in the observation space. Following [340], the background error covariance matrix, B, is factorised as B=VV^T with V=V$_V$V$_H$V$_B$. The V operators describe different aspects of the error covariances: the vertical error covariance (V_V), the horizontal error covariance (V_H), and the state variable error covariance (V_B). V_V is defined by a set of reconstructed profiles evaluated by means of an Empirical Orthogonal Function (EOF) decomposition applied to a validated multi-annual 1998-2015 run [258]. EOFs are computed for 12 months and 30 coastal and open sea sub-regions in order to account for the variability of biogeochemical anomaly fields. V$_H$ is built using a Gaussian filter whose correlation radius modulates the smoothing intensity. A non-uniform and direction-dependent correlation radius has been implemented as in [254]. V$_B$ operator consists of prescribed monthly and sub-region varying covariances among the biogeochemical variables (e.g., nitrate to phosphate). Specifically, for the assimilation of chlorophyll, the V$_B$ operator includes a balance scheme that maintains the ratio among the phytoplankton groups and preserves the physiological status of the phytoplankton cells (i.e., preserve optimal values for the internal chlorophyll, carbon and nutrients quota). The operators V_V and V_B of the 3DVarBio have been updated for the assimilation of oxygen. V_V involved the calculation of specific EOF profiles for oxygen including a localization function to avoid spurious assimilation in the deepest part of the water column. V_B included only a new direct relation for oxygen (i.e., oxygen assimilation update only the oxygen itself), given that it has been shown that it barely affects other variables [341].

Assimilated observations are composed by the QC BGC-Argo listed in Table A.1. Oxygen and nitrate profiles in the 0-600 m layer are used in the assimilation, while chlorophyll is assimilated in the 0-200 m layer.

The observation error covariance matrix R is diagonal with a monthly varying error in chlorophyll [254]. In nitrate and reconstructed nitrate profiles, the observation error is constant in time and increases along the vertical with a constant values of 0.25 mmol m^{-3} in the 0-450 meters layer [293] and linearly increasing of up to 0.35 mmol m^{-3} between 450-600m (nitrate maximum assimilation depth). Although the accuracy of the reconstruction of profiles is 0.87 mmol m^{-3} [1], we decided to not use different values of error for the two nitrate subsets in order to show the highest potential impact of the OSE.

Observation error for oxygen is set to 5 mmol m^{-3} in the upper 200 meters of depth and gradually goes to 20 mmol m^{-3} in correspondence of the maximum assimilation depth.

A.1.3 The neural network architecture and the reconstructed nitrate dataset

The NN-MLP-MED [1] is the evolution of previous MLP architectures developed to predict low-sampled variables (e.g., nutrients) starting from high-sampled ones (e.g., temperature) (Sauzède et al. 2017 , Bittig et al. 2018 and Fourrier et al. 2020)

The NN-MLP-MED presents some novel elements with respect to the mentioned methods (and in particular with respect to Canyon-Med in Fourrier et al. 2020), which lead to improved results. Firstly, the input dataset includes a larger sample size and broader coverage of the Mediterranean Sea region. Secondly, the quality of the input dataset benefits from a two-step quality check process, removing noisy and unreliable samples. The neural network architecture was also modified to enhance prediction performance by incorporating nonlinear functions, adjusting neuron count, and optimizing the training algorithm. The error of reconstructed nitrate, obtained by using the EMODnet as validation dataset, was 0.5 mmol m^{-3} [1]

In our OSE experiment, the trained NN-MLP-MED is used to reconstruct nitrate profiles from temperature and salinity (Argo), oxygen (BGC-Argo) and float date, latitude and longitude. The reconstruction also includes a vertical smoothing (running mean of 5-10 m window) and an adjustment to the 600 m climatology derived from EMODnet [313].

The input nitrate dataset for assimilation is made up of 938 BGC-Argo profiles and 2146 reconstructed nitrate profiles (Table A.1). The reconstructed nitrate profiles are located 61% in the western and 39% in the eastern Mediterranean Sea, thus providing a larger and more homogeneous spatial coverage as shown in Figures A.2.

A.1.4 BGC-Argo data and post-deployment oxygen quality control

BGC-Argo profiles available between the 2017-2018 period were downloaded from Coriolis GDAC (last visit on July 2022). Adjusted and delayed mode data were selected for oxygen and chlorophyll, while exclusively DM data were considered for nitrate.

Table A.1 reports the total number of BGC-Argo profiles, characterized by a significant number of oxygen and chlorophyll data against the relative paucity of nitrate. Figure A.2 shows the spatial distribution of the BGC profiles of chlorophyll and nitrate, while oxygen coverage can be approximated by merging nitrate and reconstructed nitrate profiles locations. Despite the lack of data in specific sub-basins (Alboran, South Western Ionian and Northern Adriatic Seas), all the three BGC variables have a fairly homogeneous spatial coverage between the western and eastern Mediterranean Sea.

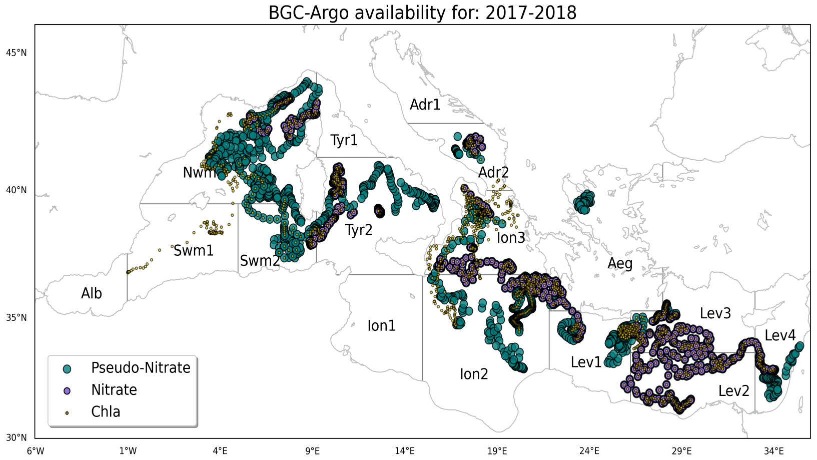


FIGURE A.2: BGC-profiles of chlorophyll-a (red), Nitrate in-situ (orange) and reconstructed Nitrate (grey) assimilated in Mediterranean Sea (2017-2018). Subdivision of the Mediterranean domain in sub-basins used for the validation. According to data availability and to ensure consistency and robustness of the metrics, different subsets of the sub-basins or some combinations among them can be used for the different metrics: lev=lev1+lev2+lev3+lev4; ion=ion1+ion2+ion3; tyr=tyr1+tyr2; adr=adr1+adr2; swm=swm1+swm2.

Since oxygen sensors may drift and lose accuracy over time, the accurate determination of dissolved oxygen is typically more challenging and requires some form of correction [308]. Expressed in % per year, the loss of accuracy is observed over the time, particularly 12 months after the deployment (https://www.euro-argo.eu). Deep ocean drift is considered as a proxy for oxygen sensor drift because of the lack of seasonal and annual signals for oxygen at depth [306].

Here, the optode sensor in-situ drift is evaluated through non parametric tests (RANSAC and Theil Sen) at two different depths (600 and 800 meters). Tests are applied when the life of a float is longer than 1 year.

The RANSAC and Theil Sen methods split the oxygen dataset into a set of inliers and outliers and drift is estimated only using inliers, avoiding possible biases due to the outliers (Dang et al. 2008 and Fischler and Bolles 1981).

The presence of a drift is established when all four drift estimates agree in sign and their average value is greater than 1 mmol m^{-3} y. This threshold was chosen on the basis of results in [310].

When detected, the suspicious drift is removed from the oxygen profiles by setting the computed drift values (i.e., the average of the four estimates) at 600 meters and linearly interpolating toward the surface, where drift is set equal to zero. Indeed, it can be assumed that O2 values at surface are already fixed by the GDACs [305].

A.1.5 Design of numerical experiments

Three numerical experiments were performed to analyze the impact of different assimilation setups. Simulated period is 1.1.2017-31.12.2018, and the MedBFM setup mostly corresponds to the standard adopted in the Mediterranean Analysis and Forecast biogeochemical system of the Marine Copernicus Service. Set up includes open boundary conditions in the Atlantic, climatological input of nutrients, carbon and alkalinity for 39 rivers and the Dardanelles Straits; and initial conditions from EMODnet dataset (Details are provided in Salon et al. 2019).

Our experimental setup differs from the standard one for the physical forcing, which are from Mediterranean Copernicus reanalysis [344], and the initial conditions of oxygen which are retrieved from BGC-Argo float climatology computed after QC O2 procedure (described in section A.1.4).

The three simulations, which share the same setup except for the assimilated datasets, are: (1) control run without assimilation (HIND); (2) assimilation of BGC-Argo chlorophyll, nitrate and oxygen and (DAfl) (3) assimilation of additional reconstructed nitrate profiles used to enhance the DAfl assimilative set up (DAnn).

Before integrating data in the 3D-VarBio, the same pre-assimilation assessments described in [255] were applied for chlorophyll. Nitrate profiles are rejected if concentration at the surface is higher than 3 mmol m^{-3}. Finally, oxygen exclusion is evaluated on the basis of the difference with the oxygen saturation values, using a threshold of

Test Case	Chl	O2	NO3	Updated variables
HIND	–	–	–	–
DAfl	1773	1924	938	phyto biomass, NO3 , O2 and PO4
DAnn	1773	1924	2146	phyto biomass, NO3 , O2 and PO4

TABLE A.1: Summary of the numerical experiments and assimilated BGC-profiles

10 mmol m^{-3} and comparing oxygen data at 600m with respect to a reference dataset using a threshold of 2 times the standard deviation of the reference dataset. For the reference dataset, we used the EMODnet2018_int data collection that integrates the in-situ aggregated EMODnet data [63] and the datasets listed in [334] and [345]. The EMODnet2018_int dataset is available for 16 sub-basins (see Figure A.2) in the Mediterranean Sea.

During the data assimilation, profiles can be excluded when model-observation misfit is higher than given thresholds. For chlorophyll the threshold is set 2 mg m^{-3} and it must be found in at least 5 vertical levels in the 0-50m layer. For nitrate, the misfit thresholds are set to 2 and 3 mmol m^{-3} in 0-50 m and 250-600m layers, respectively. Exceeding misfit has to be found in at least 5 vertical levels. Oxygen profiles are discarded by defining misfit thresholds of 30 and 50 mmol m^{-3} in at least 5 vertical levels in the 0-150m and 150-600m layers respectively.

A.2 Results

A.2.1 The oxygen post-deployment quality check method

The post deployment oxygen QC method allowed to automatically correcting in-situ sensor drifts.

Of the 40 floats available in the 2017-2018 period, the drift analysis was applied to 16 floats, while 24 floats could not be analyzed due to the limited length of the timeseries.

Over the 16 floats, a significant drift was found in 13 of them: 4 were affected by a positive drift and 9 by a negative one. The remaining 3 floats had lower drift values than the prescribed threshold (Section A.1.4).

The absolute average correction of the 13 floats is about 4.3$mmol\ m^{-3}$ performed at 600 meters of depth. This quantity is in line with the ranges expressed in terms of sensor drift percentage in [346] $(1 - 1.5\%)$.

Figure A.3 shows an example of the evolution of oxygen profiles of a quasi-stationary float (6902687) detected for a drift correction. In line with the conclusion reported in several works such as [346] and [307], the presence of a drift, like the one detected by our protocol, may reveal the possible tendency of the oxygen sensor to slowly degrade over time. After 2 years, the bias due to the drift was approximately 5 mmol m^{-3} (1st December 2017 profiles in Figure A.3.)

After removing of drift, the deep oxygen concentrations results to be closer to the EMODnet climatological data, allowing to include a higher number of profiles (example in Figure A.3, green star) in the assimilated O2 datasets (Figure A.1).

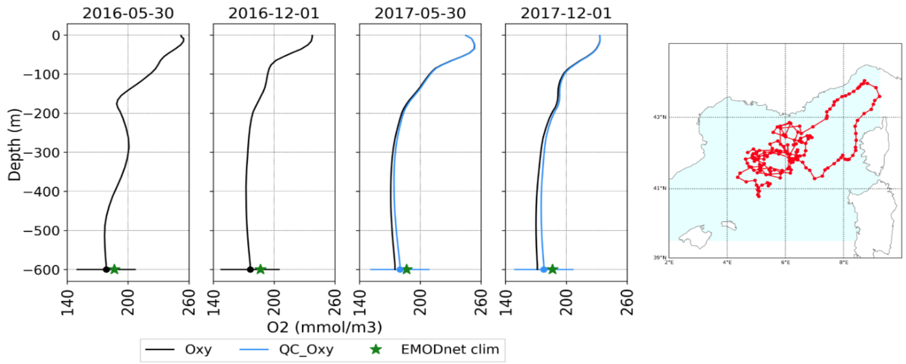


FIGURE A.3: Original (black) and corrected (blue) oxygen profiles of float 6902687 on four selected dates.
EMODnet climatology for nwm subbasin is reported (green star)

A.2.2 Validation using Satellite and BGC-Argo datasets

Skills performance of the simulations listed in Table A.1 are evaluated by comparing model results with satellite OC chlorophyll and BGC-Argo profiles. While for the satellite comparison the model daily averages are considered, the model first guess (i.e. the model state before the assimilation) is used for metrics based on BGC-Argo.

Root Mean Square Error (RMSE) metric is evaluated in winter (from February to April, FMA) and summer (from June to August, JJA) to investigate the model's capacity to reproduce the specific bloom and stratification conditions in 16 Mediterranean Sea subbasins or in an aggregated combination of them (Figure A.2).

Satellite L3 products from Copernicus Marine Service catalogue were interpolated from 1 km to the model resolution and a composite weekly average was computed to ensure gap-free maps, as in [324].

The Winter RMSE with respect to OC chlorophyll in HIND spans between ca. 0.09 to 0.21 mg m^{-3} with a maximum in the Alboran Sea (Figure A.4). The addition of multivariate DA (DAfl) has a positive impact with a reduction of surface errors of by 6.5% mainly observed in the eastern sub-basins. A further reduction of RMSE (up to 10%) with respect to HIND is then obtained with DAnn showing that enlarging the nitrate float network leads to improvements in reproducing surface phytoplankton dynamics. All the Mediterranean sub-basins show an RMSE reduction with the exception of alb, swm and nwm. A generalized slight worsening in the assimilated runs can generally be observed during the summer stratification period. The RMSE with respect to OC chlorophyll, which increases in all sub-basins, is fairly similar in the two assimilation runs: about 6% and 7.5% in DAfl and DAnn, respectively. It should be noted that the RMSE values in summer are an order of magnitude lower than in winter, reflecting the seasonal chlorophyll variability in the Mediterranean Sea (i.e., the very low values of chlorophyll at the surface).

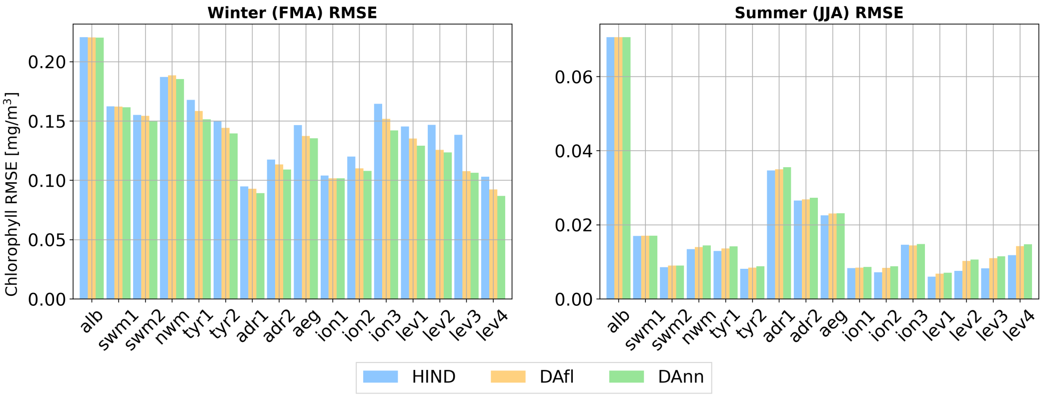


FIGURE A.4: Seasonal chlorophyll RMSE: Winter bloom and Summer stratification seasons in the Mediterranean Sea sub-basins for the HIND run (light blue), the DAfl run (in orange) and the DAnn run (green).

The RMSE metrics based on BGC-Argo are computed for six selected aggregated macro-basins (Alboran, South West Mediterranean, North West Mediterranean, Tyrrhenian, Ionian and Levantine Seas) and in selected layers (0-10m, 10-30m, 30-60m, 60-100m, 100-150m, 150-300m and 300-600m) and are shown in Figure A.5 for nitrate (top panel), chlorophyll (middle panel) and oxygen (bottom panel). It is worth recalling that only in-situ profiles are used in the validation (i.e., reconstructed NN profiles are used only for assimilation).

As expected, the assimilation of in-situ BGC-Argo considerably improves the quality of modelled nitrate with respect to the HIND run (Figure A.5). Winter RMSE reduction goes from 40% (DAfl) to approximately 46%, when also reconstructed profiles are assimilated, while the reduction of summer RMSE increases from 59% in DAfl to 63% in DAnn. Maximum RMSE reduction of RMSE of DAnn with respect to DAfl is observed in nwm and tyr (winter) and in ion (summer). This impact can be directly ascribed to the increased number of reconstructed nitrate in these sub-basins and seasons where additional profiles generate more persistent corrections.

The advantages of assimilating chlorophyll profiles have been already shown in [255]. Here, improvements related to chlorophyll assimilation can be observed in nwm, ion and lev in winter and at depth in tyr, ion and lev in summer (Figure A.5 middle panel). Even if phytoplankton dynamics depend on nutrients dynamics, the positive impact of DAnn on nitrate RMSE does not transfer to the the vertical chlorophyll statistics in the DAnn. This is mainly because the DAfl and DAnn simulations assimilate the same chlorophyll dataset, and the direct chlorophyll assimilation is more effective than the dynamical model adjustment after nitrate and reconstructed nitrate assimilation in the areas close to the observed chlorophyll profiles.

Assimilating oxygen profiles enable reducing the model-BGC floats RMSE by about 30% during winter and summer. The correction involves the whole water columns with a maximum correction between 150-600m during winter in the west and along the entire profiles in the east (ion and lev) in both winter and summer. The addiction of reconstructed profiles in the DAfl run does not significantly affect the quality of oxygen.

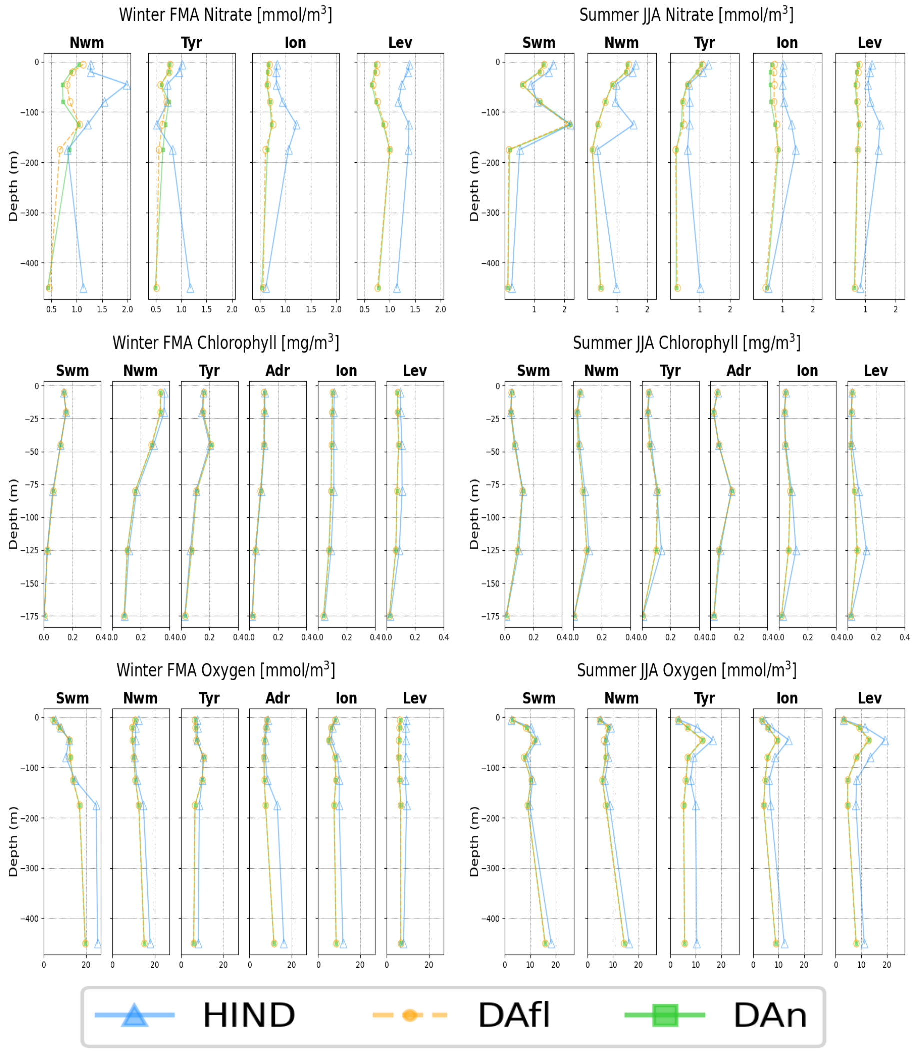


FIGURE A.5: Seasonal Nitrate, Chlorophyll and Oxygen RMSE (top, middle, bottom): Bloom (left) and Stratification (right) seasons in the Mediterranean Sea aggregated sub-basins for the HIND run (pale blue), DAfl run (orange) and DAnn (green).

A.2.3 Integration of NN DA: the impact

A.2.3.1 Impacts on biogeochemical vertical dynamics

As expected, the profile assimilation plays a major role in changing the vertical gradients of biogeochemical variables. Figures A.6, A.7, A.8, and A.9 show the impact of the assimilation in two sub-basins where the number of reconstructed profiles is high (Figure A.2) and in the Mediterranean Sea (third column). The two sub-basins represent two different trophic conditions in the Mediterranean Sea. The North Western Mediterranean (nwm) has higher level of nutrient concentration and more intense surface bloom in winter (Siokou-Frangou et al. 2010 and Di Biagio et al. 2022). In summer, nwm has higher chlorophyll concentration at the deep chlorophyll maximum (DCM), shallow nitracline, and shallow subsurface oxygen maximum (SOM) (first column in Figures A.6, A.7, A.8, and A.9). On the contrary, more oligotrophic conditions and deeper nitracline and DCM are found in the eastern subbasin (ion2, second column of Figures A.6, A.7, A.8, and A.9). The assimilation of nitrate corrects a general positive bias of the model in all the Mediterranean areas (blue pattern in Figure A.7). The addition of reconstructed profiles makes the corrections stronger. At the Mediterranean scale, the nitrate concentration below the nitracline decreases by 8% and 11% in DAfl and DAnn runs, respectively. Nitracline depth changes (i.e. deeper values) by DAfl assimilation by a few (nwm) and tens of meters (ion2). The deepening of the nitracline becomes more intense with the inclusion of reconstructed profiles (DAnn). The differences between the assimilation and reference runs accumulates over time and eventually reach a stationary phase in the second year in the sub-basins with a high number of BGC-Argo and reconstructed profiles, such as nwm and ion2. On the other hand, considering the whole Mediterranean Sea, which comprises some under-sampled areas (e.g., southern Ionian and southern western basin), the effect of corrections is still propagating after the two years (third column of Figure A.7)

Very similar patterns are also observed in the Hovmöller diagrams of phosphate (Figure A.8), which is an updated variable of the multivariate variational assimilation scheme through nitrate. In fact, the general negative corrections on phosphate fields are linked to the high positive values of the covariance matrix between nitrate and phosphate [255]

As a consequence of both the direct assimilation of chlorophyll profiles and the dynamical model adjustment after nitrate assimilation, the main effects of DAfl are to slightly reduce the intensity of chlorophyll concentration in the DCM (e.g., variation smaller than 5% with respect to HIND simulation) and in adjusting the timing of the surface winter blooms (second row in Figure A.6). Even if the chlorophyll validation (Figure A.5) has not shown significant differences between DAfl and DAnn, the basin wide averages of DAnn display more intense corrections with respect to DAfl in terms of DCM depth and chlorophyll intensity and overall chlorophyll concentration (figure A.6). Over the 0-200 layer of the entire Mediterranean Sea, the chlorophyll decreases with respects to HIND are 4% and 5% for DAfl and DAnn, respectively.

BGC-Argo oxygen profiles assimilation (DAfl, second row in Figure A.9) provides positive or negative corrections depending on the observation-model bias which varies in time and space (e.g., mostly negative in nwm and mostly positive in ion2). On a Mediterranean basin-wide scale, the average correction is of 0.2% for the 0-200m layer. The addition of the nitrate reconstructed profiles does not alter the correction pattern with an average correction of 0.3%. The only noticeable difference between the two assimilation runs can be spotted in areas with a high density of reconstructed profiles during summer (e.g., nwm, first column in Figure A.9). As observed in the nitrate and chlorophyll figures, the assimilation of reconstructed profiles causes a decrease of the summer productivity in the DCM layer. Consequently, less oxygen is produced generating the negative changes in the DCM layer in the bottom left panel of Figure A.9. Because of the smaller amount of subsequent sinking organic matter, less oxygen is consumed in remineralization processes in layers below the DCM in late summer and autumn, and positive oxygen changes are generated, particularly during 2018.

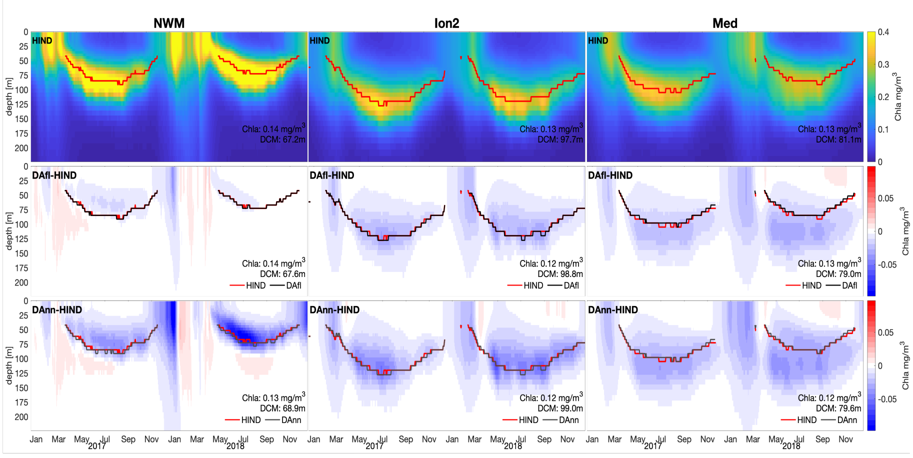


FIGURE A.6: Hovmöller diagram of chlorophyll of hindcast simulation (first row) and differences between assimilation runs and hindcast (second and third rows) for 2 sub-basins (nwm and ion2) and the Mediterranean Sea (med).

Evolution of the depth of nitracline for the three runs: red (hind) and black (DAfl and DAnn) lines. The averages of 0-200m concentration and of nitracline for the simulated period are reported.

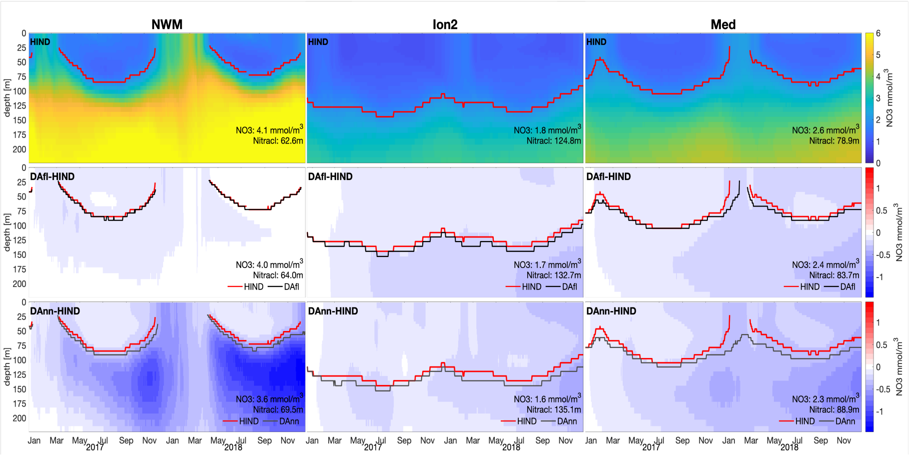


FIGURE A.7: Hovmöller diagram of nitrate of hindcast simulation (first row) and differences between assimilation runs and hindcast (second and third rows) for 2 subbasins (nwm and ion2) and the Mediterranean Sea (med).

Evolution of the depth of nitracline for the three runs: red (hind) and black (DAfl and DAnn) lines. The averages of 0-200m concentration and of nitracline for the simulated period are reported.

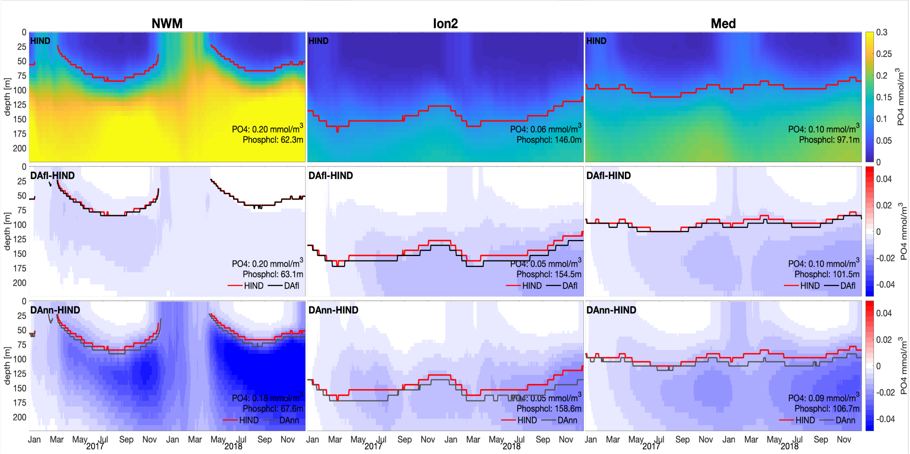


FIGURE A.8: Hovmöller diagram of phosphate of hindcast simulation (first row) and differences between assimilation runs and hindcast (second and third rows) for 2 subbasins (nwm and ion2) and the Mediterranean Sea (med).

Evolution of the depth of phosphocline for the three runs: red (hind) and black (DAfl and DAnn) lines. The averages of 0-200m concentration and of phosphocline for the simulated period are reported.

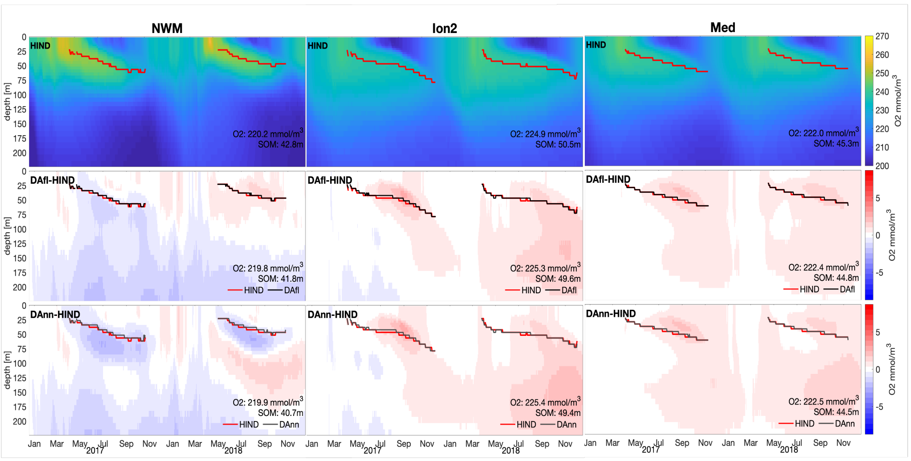


FIGURE A.9: Hovmöller diagram of oxygen of hindcast simulation (first row) and differences between assimilation runs and hindcast (second and third rows) for 2 sub-basins (nwm and ion2) and the Mediterranean Sea (med).

Evolution of the depth of subsurface oxygen maximum (SOM) for the three runs: red (hind) and black (DAfl and DAnn) lines). The averages of 0-200m concentration and of SOM for the simulated period are reported.

A.2.3.2 Impact on ecosystem indicator

Net primary production (NPP) integrates phytoplankton growth and respiration processes which are at the basis of the marine trophic food web. Assimilation of chlorophyll and nitrate together with the updates of phosphate, directly and indirectly affect primary production, since they impact phytoplankton biomass and nutrient availability. Thus, the comparison of primary production among the three simulations reveals how the assimilation impacts on a key indicator that integrates several marine ecosystem processes. Seasonal maps of net primary production integrated over the 0-200 m layer in the HIND, DAfl and DAnn simulations (Figure A.10) confirm that the assimilation impact varies spatially and temporally. In the DAfl simulation, the largest differences in primary production with respect to the HIND simulation are located in the eastern Mediterranean with a decrease of nearly 10% in the Levantine sub-basins and in the Ionian Sea close to the Greek coast. Reductions are slightly larger in winter. In the western Mediterranean the impacts on primary production are negligible in both seasons with the only exception of a 5% reduction in winter in the Tyrrhenian Sea. Areas with changes on NPP corresponding to the areas with assimilation of float profiles that include nitrate. In the DAnn simulation, the impacts on primary production are more intense than in DAfl and the impacted areas are larger. In particular, primary production is decreased also in areas such as the western Mediterranean in summer (8%) and in the northwestern Mediterranean and in the Tyrrhenian Sea in winter (5%).

Moreover, a further reduction of NPP occurs in the Ionian Sea in both seasons. In general, the impact on primary production is greater where nitrate observations or nitrate reconstructed observations are assimilated, suggesting a dynamical bottom-up control of primary production. In fact, the weaker fertilization of the surface layer in DAnn, which occurs for both macronutrients after assimilation (Figure A.7 and A.8), appears to be the main cause of reduced NPP, outweighing the effects due to changes in phytoplankton biomass after chlorophyll assimilation.

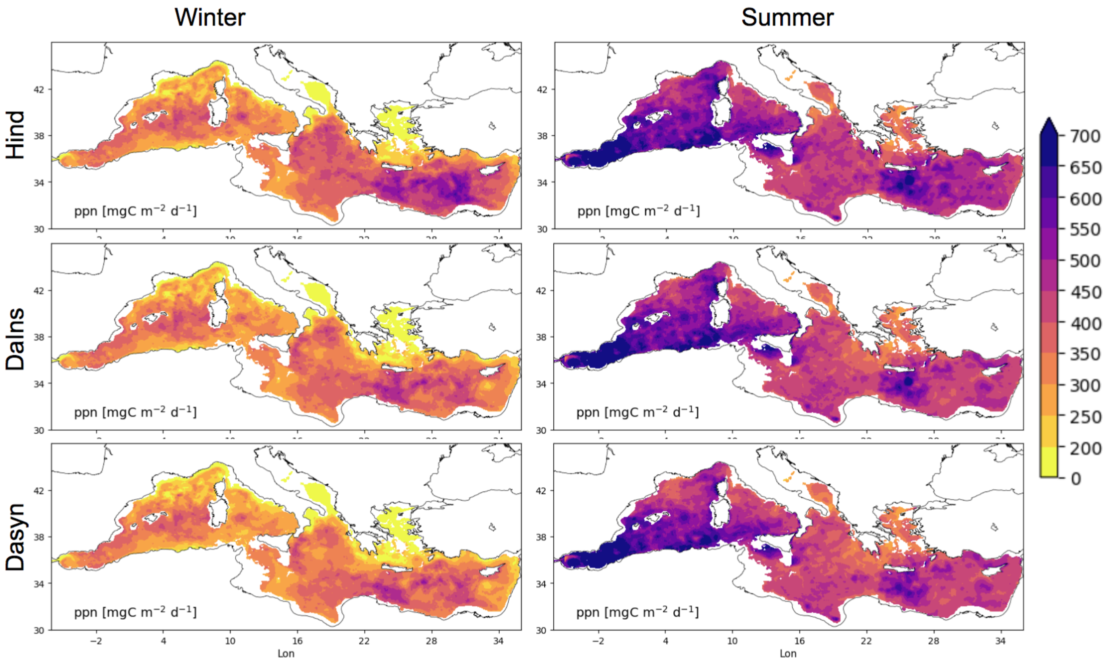


FIGURE A.10: Maps of net winter (FMA) and summer (JJA) primary production [NPP, mgC m-2 d-1] in the three runs: A. HIND B. DAfl C. DAnn.

Seasonal averages are computed considering the period 2017-2018.

A.2.3.3 Impact on Argo Observing system design

Analyzing the departure of an assimilated simulation from a reference solution provides insights into the impact of the observing system design and several data impact indicators can be used (Ford 2021, Teruzzi et al. 2021 and Raicich and Rampazzo 2003). In this work, we adopt the impact indicator $I_{ij}(t)$ described in [255], in order to quantify the integrated (0-300m) response of assimilating BGC Argo profiles with respect to the no assimilative run:

$$I_{ij}(t) = \frac{|Sim_{ij}(t) - HIND_{ij}(t)|_{300}}{(HIND_{300})_{mean}} \tag{A.2}$$

Here, HIND is here the reference, while Sim refers to one of the different DA set-ups. $|Sim_{ij}(t) - HIND_{ij}(t)|$ is the absolute difference between simulations (for each day and grid point) while the subscript 300 represents the integral over the 0–300m. The indicator $I_{ij}(t)$ quantifies how much an assimilated run deviates from the reference (HIND) simulation. Figures A.11 and A.12 show the nitrate and chlorophyll $I_{ij}(t)$ 95th percentile of the seasonal indicator in winter (left column) and in summer (right column) in the DAfl (first row) and DAnn (second row) simulations. The 95th percentile of the indicator shows that the BGC-Argo assimilation impacts (DAfl) both chlorophyll and nitrate.

In DAfl, the extent of nitrate $I_{ij}(t)$ 95th above 0.1 is 16.5% and 18.7% in winter and in summer respectively, with clear spatial distribution mapping the BGC-Argo density. The introduction of reconstructed profiles in DAnn make it possible to increase the nitrate impacted areas up to about 35% and 39% in winter and summer respectively. The DAnn impact increase is mainly localized in the western Mediterranean Seas and in the Ionian Sea, while the less evident impact in the Levantine, especially in summer, is mainly due to the low number of NN reconstructed nitrate in the area.

Chlorophyll impact maps (Figure A.12) show that besides the direct impact of chlorophyll profiles assimilation, phytoplankton is also affected by the reconstructed nitrate assimilation. Compared to a threshold of 0.4, the impacted areas increase from 18.2% to 29.8% in winter and from 10.8% to 14.5 in summer in the DAfl and DAnn runs. These results suggest that the inclusion of reconstructed nitrate assimilation can potentially extend the impact to almost all the Mediterranean Sea, with the only exclusion of the marginal seas (Adriatic and Aegean) and the southern part of Ionian and Western subbasins.

Oxygen impact maps (not shown) are very similar to nitrate DAnn maps and do not show differences between the two DA simulation, since the same QC oxygen dataset was assimilated in DAfl and DAnn.

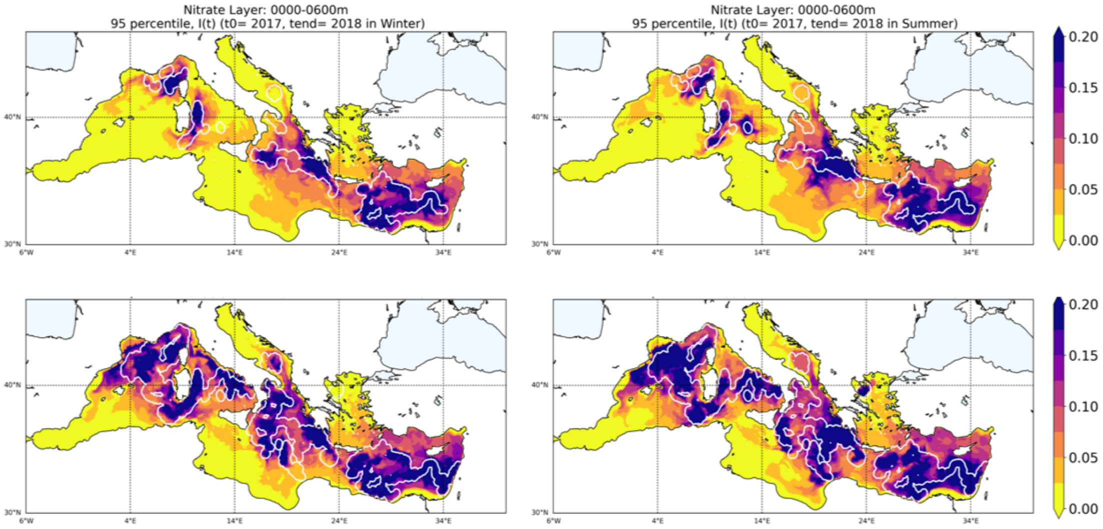

Nitrate Layer: 0000-0600m
95 percentile, I(t) (t0= 2017, tend= 2018 in Winter)
Nitrate Layer: 0000-0600m
95 percentile, I(t) (t0= 2017, tend= 2018 in Summer)
40°N
30°N
6°W
4°E
14°E
24°E
34°E
0.20
0.15
0.10
0.05
0.00

FIGURE A.11: Maps of Iij(t) 95th percentiles for Nitrate in winter (left column) and summer (right column) in the DAfl (first row) and DAnn (second row); white contours identify the areas within three correlation radii from the float profiles

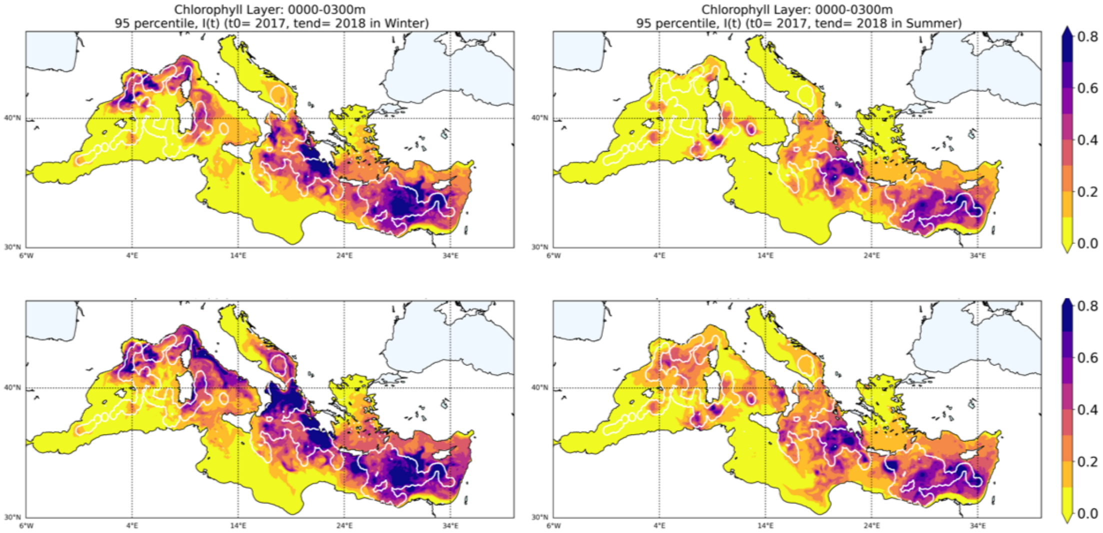

Chlorophyll Layer: 0000-0300m
95 percentile, I(t) (t0= 2017, tend= 2018 in Winter)
Chlorophyll Layer: 0000-0300m
95 percentile, I(t) (t0= 2017, tend= 2018 in Summer)
40°N
30°N
6°W
4°E
14°E
24°E
34°E
0.8
0.6
0.4
0.2
0.0

FIGURE A.12: Maps of Iij(t) 95th percentiles for Chlorophyll in winter (left column) and summer (right column) in the DAfl (first row) and DAnn (second row); white contours identify the areas within three correlation radii from the float profiles

A.3 Discussion

Our quality check procedure for oxygen drift detection and comparison with a reference dataset can successfully integrates the official QC BGC-Argo, making oxygen BGC-Argo a robust and valuable dataset for initial conditions, data assimilation, validation and new product reconstruction. Even if the distinction between real oxygen depletion signals and optode drift can remain problematic without in-situ high quality data, we believe that literature and prior knowledge can be used as a baseline for drift discrimination.

In particular, recent studies have revealed a decrease of oxygen concentrations in the Mediterranean Sea (Sisma-Ventura et al. 2021 and Di Biagio et al. 2022); such tendency has been defined as a multidecadal shifts (Coppola et al. 2018 and Mavropoulou et al. 2020) and also patchy deoxygenation [351]. One of the most evident signals is the early 1990's East Mediterranean Transient (EMT) associated with the variations of thermohaline circulation, which has caused a strong interannual variability of oxygen [348]. Based on this literature and considering the recent years, a threshold of 1 mmol m^{-3} y at 600 and 800 meters appears a prudent limit for sensor drift discrimination from real long term signals.

Up to now, the oceanographer visual check is required to distinguishing ocean signals from sensor drift [311] and the debate on how to replace visual check to automatic statistical procedures is still open. Thus, our work can contribute by proposing a new tool to automatically handle deep ocean signal or optode drift issues.

The search of drift is based on a robust combination of several factors: (i) the float history has to be longer than 1 year, (ii) the calculation is done with two robust non-parametric trend methods and (iii) at two different depths (i.e. 600m and 800m) sampling different water masses that can have different long-term dynamics. Lastly, the sensor drift signal must be confirmed by the four results and has to be compared with typical basin-wide trend values reported by scientific literature (i.e., the choice of the threshold).

The method can be further developed by applying the oxygen drift analysis at some fixed isopycnals together with the one at constant isobaths. Thus, possible oxygen concentration changes due to floats moving across different water masses can be filtered out.

The assimilation of vertical profiles provides complementary information with respect to satellite ocean color assimilation (Cossarini et al. 2019 and Verdy and Mazloff 2017), which however is still the most commonly used in operational systems [285]. In fact, the effectiveness of the profiles assimilation, which has the capability to constrain vertical biogeochemical dynamics in subsurface layers (Kaufman et al. 2018, Teruzzi et al. 2021 and Wang et al. 2022), lies in the amount of available BGC-Argo data, that are generally insufficient to constrain a basin wide simulation. In [255], results of the impact indicator principally showed the efficiency of ocean color assimilation in constraining chlorophyll dynamics mostly during winter. In this work, important impacts are also observed in summer for all variables, as a consequence of the increased number of assimilated profiles.

In fact, through the integration of NN and DA, the number of nitrate profiles ingested can significantly increase, with a density of more than 30 profiles every ten days over a basin of 2.5M km^2. Indeed, as shown in the results, Argo and BGC-Argo oxygen sensors can potentially support a density of biogeochemical profiles up to 1 in each 2.5° x 2.5° box every 10 days. This means that seasonal sub-basins scale dynamics (e.g., bloom or stratification) can effectively be constrained while, up to now, mesoscale dynamics can probably be exclusively locally studied [315]. Apart from an increase in the numbers of floats, improvements in the simulation of mesoscale dynamics can be achieved by redefining horizontal covariance errors in the data assimilation scheme. Indeed, benefits of non-uniform correlation radius in the horizontal scale have been previously investigated [254] and additional improvements could be provided by a 3D varying correlation radius [355].

Looking at the recent evolution in the availability of BGC-Argo sensors (Figure A.13), our combined NN and DA approach would allow keeping the benefits of the BGC-Argo OS in the Mediterranean operational system. Even if nitrate and chlorophyll profiles have dramatically decreased after 2020, the assimilation of reconstructed profiles can potentially overcome this lack. Nevertheless, as shown in our OSE experiments (Figure A.11 and A.12), there are still undersampled areas by the Argo and oxygen sensors, such as Alboran, Southern Ionian seas and the marginal seas (Northern Adriatic and Northern Aegean Sea) which would require specific deployments.

With respect to previous BGC OSSE experiments (Yu et al. 2018, Ford 2021), we show how to exploit the current Argo and BGC-Argo networks for reconstructing biogeochemical variables.

MLP feed-forward methods to reconstruct biogeochemical variables are good enough (Bittig et al. 2018, Sauzède et al. 2020 Fourrier et al. 2021 and Pietropolli et al. 2023) to reach our purposes, even if their application to generate smooth and consistent profiles still has some limitations [1]. The MLP-NN-MED method has a validation error of 0.50 mmol^2 m^{-3} for nitrate and 0.87 mmol^2 m^{-3} when it is used to predict BGC-Argo data [1] . These values are slightly higher than the BGC-Argo error estimated from the triple collocation method [293], which is used as the observation error. We recognized that different error values can be used for BGC-Argo and reconstructed profiles to take account of their uncertainties. Then, it is intuitive that with higher error, the reconstructed dataset impact would have been lower.

Indeed, the larger error in MLP-NN-MED prediction of BGC-Argo profiles derives the fact that the MLP methods, being pointwise based, are unaware of the vertical gradient (e.g., typical shape) of the profiles of biogeochemical variables that they seek to infer. This fact can lead to irregularities and lack of smoothness in the predicted profiles [1], which we partly solved by adding a smoothing operator. However, one way to increase the reliability of profile reconstruction would be to include information with a physical meaning from observed data [256]. 1D Convolutional Neural Networks represent a viable alternative approach considering their ability to treat the coherence of the 1D signals (e.g., typical shapes of profiles) as shown in [358].

Integration of NN and DA have been tested in several geoscience applications (Buizza et al. 2022, Brajard et al. 2021, Stanev et al. 2022) to infer unresolved spatial scales or reproduce missing data. In our application, the integration of NN, which retrieves a large number of profiles [1], and DA, which can apply the correction to all nutrients through error covariances [255], allows spatial and multivariate changes to be captured both at the local scale and across the basin to constrain Mediterranean productivity (Figure A.10). Although the corrections take time to extend to the entire basin (Figure A.7), our simulations have shown that constraining bottom-up ecosystem processes (e.g., productivity, organic matter sink) has proven effective and should be used in conjunction with the classical ocean color correction to phytoplankton biomass.

Any plan to learn directly from observations will have to face with some challenges, such as the use of observations whose time and space coverage is uneven or related to specific processes [241]. The modular approach followed in this work represents a successful example of exploiting the strengths of neural networks and data assimilation to enhance the observing system impact in the operational biogeochemical system of the Mediterranean Sea.

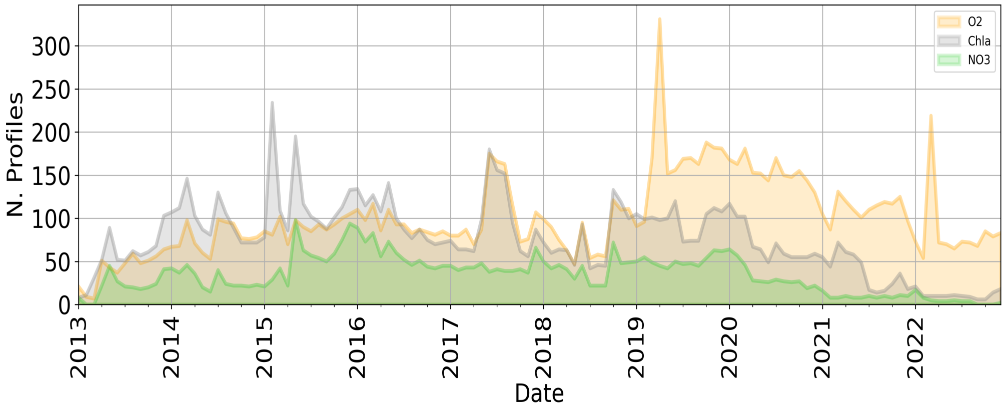


FIGURE A.13: Timeserie of BGC-Argo profiles availability (2013-2022) for: nitrate (green), chlorophyll (grey) and oxygen (yellow)

A.4 Conclusions

Combining deterministic Feed-Forward Neural Network and Data Assimilation to design an Observing System Experiment has enabled demonstrating the enhanced positive impact of profiles assimilation in the Copernicus Operational System for Short-Term Forecasting of the Biogeochemistry of the Mediterranean Sea (MedBFM).

The development of the oxygen QC procedure allowed to statistically deal with optode in-situ drift and to derive accurate reconstructed profiles of nitrate, keeping the number of assimilated observations at a much higher level despite the current negative trend in BGC-Argo availability. Achieved BGC profiles density provides valuable and additional information to complement that of ocean colour in describing phytoplankton seasonal blooms and stratification dynamics at sub-basins scale.

The assimilation of BGC-Argo nitrate corrects a general positive bias of the model in several Mediterranean areas, and the addition of reconstructed profiles makes the correction stronger.

Together with nitrate assimilation, the phosphate update through error covariances, sustains spatial and multivariate changes that are capable of correcting key biogeochemical processes (e.g., nitracline and deep chlorophyll maximum) and to constrain ecosystem processes (e.g., productivity) at basin-wide scale.

Bibliography

[1] Gloria Pietropolli, Luca Manzoni, and Gianpiero Cossarini. Multivariate relationship in big data collection of ocean observing system. *Applied Sciences*, 13(9): 5634, 2023.

[2] Carolina Amadio, Anna Teruzzi, Gloria Pietropolli, Luca Manzoni, Gianluca Coidessa, and Gianpiero Cossarini. Combining neural networks and data assimilation to enhance the spatial impact of argo floats in the copernicus mediterranean biogeochemical model. *EGUsphere*, 2023:1–28, 2023.

[3] Gloria Pietropolli, Luca Manzoni, and Gianpiero Cossarini. Ppcon 1.0: Biogeochemical argo profile prediction with 1d convolutional networks. *EGUsphere*, 2023:1–23, 2023.

[4] Gloria Pietropolli, Gianpiero Cossarini, and Luca Manzoni. Gans for integration of deterministic model and observations in marine ecosystem. In *Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings*, pages 452–463. Springer, 2022.

[5] Mauro Castelli, Luca Manzoni, Luca Mariot, Giuliamaria Menara, and Gloria Pietropolli. The effect of multi-generational selection in geometric semantic genetic programming. *Applied Sciences*, 12(10):4836, 2022.

[6] Alberto Leporati, Luca Manzoni, Giancarlo Mauri, Gloria Pietropolli, and Claudio Zandron. Inferring p systems from their computing steps: An evolutionary approach. *Swarm and Evolutionary Computation*, 76:101223, 2023.

[7] Giorgia Nadizar and Gloria Pietropolli. A grammatical evolution approach to the automatic inference of p systems. *Journal of Membrane Computing*, pages 1–15, 2023.

[8] Gloria Pietropolli, Giuliamaria Menara, Mauro Castelli, et al. A genetic programming based heuristic to simplify rugged landscapes exploration. *Emerging Science Journal*, 7(4):1037–1051, 2023.

[9] Gloria Pietropolli, Luca Manzoni, Alessia Paoletti, and Mauro Castelli. Combining geometric semantic gp with gradient-descent optimization. In *European Conference on Genetic Programming (Part of EvoStar)*, pages 19–33. Springer, 2022.

[10] José Ferreira, Mauro Castelli, Luca Manzoni, and Gloria Pietropolli. A self-adaptive approach to exploit topological properties of different gas' crossover operators. In *European Conference on Genetic Programming (Part of EvoStar)*, pages 3–18. Springer, 2023.

[11] Gloria Pietropolli, Federico Julian Camerota Verdù, Luca Manzoni, and Mauro Castelli. Parametrizing gp trees for better symbolic regression performance through gradient descent. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pages 619–622, 2023.

[12] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[13] Malte Meinshausen, Nicolai Meinshausen, William Hare, Sarah CB Raper, Katja Frieler, Reto Knutti, David J Frame, and Myles R Allen. Greenhouse-gas emission targets for limiting global warming to 2 c. *Nature*, 458(7242):1158–1162, 2009.

[14] Wilfried Thuiller. Climate change and the ecologist. *Nature*, 448(7153):550–552, 2007.

[15] James Garvey. The ethics of climate change. *The Ethics of Climate Change*, pages 1–192, 2008.

[16] Nicolas Gruber. Warming up, turning sour, losing breath: ocean biogeochemistry under global change. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1943):1980–1996, 2011.

[17] Lisa M Campbell, Noella J Gray, Luke Fairbanks, Jennifer J Silver, Rebecca L Gruby, Bradford A Dubik, and Xavier Basurto. Global oceans governance: New and emerging issues. *Annual review of environment and resources*, 41:517–543, 2016.

[18] J-P Gattuso, Alexandre Magnan, Raphaël Billé, William WL Cheung, Ella L Howes, Fortunat Joos, Denis Allemand, Laurent Bopp, Sarah R Cooley, Caroline M Eakin, et al. Contrasting futures for ocean and society from different anthropogenic co2 emissions scenarios. *Science*, 349(6243):aac4722, 2015.

[19] Hans-Otto Pörtner, Debra C Roberts, Valérie Masson-Delmotte, Panmao Zhai, Melinda Tignor, Elvira Poloczanska, Katja Mintenbeck, Andrés Alegría, Maike

Nicolai, Andrew Okem, et al. Ipcc special report on the ocean and cryosphere in a changing climate. *IPCC Intergovernmental Panel on Climate Change: Geneva, Switzerland*, 1(3):1–755, 2019.

[20] Susan Wijffels, Dean Roemmich, Didier Monselesan, John Church, and John Gilson. Ocean temperatures chronicle the ongoing warming of earth. *Nature Climate Change*, 6(2):116–118, 2016.

[21] Terry P Hughes, Michele L Barnes, David R Bellwood, Joshua E Cinner, Graeme S Cumming, Jeremy BC Jackson, Joanie Kleypas, Ingrid A Van De Leemput, Janice M Lough, Tiffany H Morrison, et al. Coral reefs in the anthropocene. *Nature*, 546(7656):82–90, 2017.

[22] Robert S Nerem, Brian D Beckley, John T Fasullo, Benjamin D Hamlington, Dallas Masters, and Gary T Mitchum. Climate-change–driven accelerated sea-level rise detected in the altimeter era. *Proceedings of the national academy of sciences*, 115(9): 2022–2025, 2018.

[23] Michael Oppenheimer, Bruce Glavovic, Jochen Hinkel, Roderik Van de Wal, Alexandre K Magnan, Amro Abd-Elgawad, Rongshuo Cai, Miguel Cifuentes-Jara, Robert M Deconto, Tuhin Ghosh, et al. Sea level rise and implications for low lying islands, coasts and communities. 2019.

[24] John M Guinotte and Victoria J Fabry. Ocean acidification and its potential effects on marine ecosystems. *Annals of the New York Academy of Sciences*, 1134(1):320–342, 2008.

[25] Scott C Doney, Victoria J Fabry, Richard A Feely, and Joan A Kleypas. Ocean acidification: the other co2 problem. *Annual review of marine science*, 1:169–192, 2009.

[26] Agathe Euzen, Françoise Gaill, Denis Lacroix, and Ohilippe Cury. The ocean revealed, 2017.

[27] Kerry Emanuel. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, 436(7051):686–688, 2005.

[28] The global ocean observing system. URL https://www.goosocean.org/. https://www.goosocean.org/ (accessed: 13.09.2022).

[29] Linda Sofie Lindström, Eva Karlsson, Ulla M Wilking, Ulla Johansson, Johan Hartman, Elisabet Kerstin Lidbrink, Thomas Hatschek, Lambert Skoog, and Jonas Bergh. Clinically used breast cancer markers such as estrogen receptor,

progesterone receptor, and human epidermal growth factor receptor 2 are unstable throughout tumor progression. *J Clin Oncol*, 30(21):2601–2608, 2012.

[30] Toste Tanhua, Sylvie Pouliquen, Jessica Hausman, Kevin O'brien, Pip Bricher, Taco De Bruin, Justin JH Buck, Eugene F Burger, Thierry Carval, Kenneth S Casey, et al. Ocean fair data services. *Frontiers in Marine Science*, 6:440, 2019.

[31] LD Talley, RA Feely, BM Sloyan, R Wanninkhof, MO Baringer, JL Bullister, CA Carlson, SC Doney, RA Fine, E Firing, et al. Changes in ocean heat, carbon content, and ventilation: a review of the first decade of go-ship global repeat hydrography. *Annual review of marine science*, 8:185–215, 2016.

[32] Dean Roemmich and Argo Steering Team. Argo: the challenge of continuing 10 years of progress. *Oceanography*, 22(3):46–55, 2009.

[33] James KB Bishop, Russ E Davis, and Jeffrey T Sherman. Robotic observations of dust storm enhancement of carbon biomass in the north pacific. *Science*, 298 (5594):817–821, 2002.

[34] BG Mitchell. Resolving spring bloom dynamics in the sea of japan. In *ALPS: Autonomous and Lagrangian Platforms and Sensors, Workshop Report, eds DL Rudnick and MJ Perry (La Jolla, CA: ALPS)*, pages 26–27, 2003.

[35] E Boss, D Swift, L Taylor, P Brickley, R Zaneveld, S Riser, MJ Perry, and PG Strutton. Observations of pigment and particle distributions in the western north atlantic from an autonomous float and ocean color satellite. *Limnology and Oceanography*, 53(5part2):2112–2122, 2008.

[36] Gayathri K Devi, BP Ganasri, and GS Dwarakish. Applications of remote sensing in satellite oceanography: A review. *Aquatic Procedia*, 4:579–584, 2015.

[37] R Rajeesh and GS Dwarakish. Satellite oceanography–a review. *Aquatic Procedia*, 4:165–172, 2015.

[38] Anna Teruzzi, Pierluigi Di Cerbo, Gianpiero Cossarini, Eric Pascolo, and Stefano Salon. Parallel implementation of a data assimilation scheme for operational oceanography: The case of the medbfm model system. *Computers & Geosciences*, 124:103–114, 2019.

[39] Katja Fennel, Jann Paul Mattern, Scott C Doney, Laurent Bopp, Andrew M Moore, Bin Wang, and Liuqian Yu. Ocean biogeochemical modelling. *Nature Reviews Methods Primers*, 2(1):76, 2022.

[40] Jorn Bruggeman and Karsten Bolding. A general framework for aquatic biogeochemical models. *Environmental modelling & software*, 61:249–265, 2014.

[41] Patrick Henry Winston. *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[43] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[44] Myeongsuk Pak and Sanghoon Kim. A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*, pages 1–3. IEEE, 2017.

[45] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[46] Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3):261–273, 2015.

[47] Jennifer M Durden, Jessica Y Luo, Harriet Alexander, Alison M Flanagan, and Lars Grossmann. Integrating "big data" into aquatic ecology: Challenges and opportunities. *Limnology and Oceanography Bulletin*, 26(4):101–108, 2017.

[48] Patricia A Soranno and David S Schimel. Macrosystems ecology: big data, big ecology, 2014.

[49] Marjolaine Matabos, Maia Hoeberechts, Carol Doya, Jacopo Aguzzi, Jessica Nephin, Thomas E Reimchen, Steve Leaver, Roswitha M Marx, Alexandra Branzan Albu, Ryan Fier, et al. Expert, crowd, students or algorithm: who holds the key to deep-sea imagery 'big data'processing? *Methods in Ecology and Evolution*, 8(8):996–1004, 2017.

[50] Katrin Schroeder, SA Josey, Marine Herrmann, Laure Grignon, GP Gasparini, and HL Bryden. Abrupt warming and salting of the western mediterranean deep water after 2005: Atmospheric forcings and lateral advection. *Journal of Geophysical Research: Oceans*, 115(C8), 2010.

[51] Harry L Bryden, Julio Candela, and Thomas H Kinder. Exchange through the strait of gibraltar. *Progress in Oceanography*, 33(3):201–248, 1994.

[52] Wolfgang Roether, Beniamino B Manca, Birgit Klein, Davide Bregant, Dimitrios Georgopoulos, Volker Beitzel, Vedrana Kovačević, and Anna Luchetta. Recent changes in eastern mediterranean deep waters. *Science*, 271(5247):333–335, 1996.

[53] Paola Malanotte-Rizzoli, Beniamino B Manca, Maurizio Ribera d'Alcala, Alexander Theocharis, Stephen Brenner, Giorgio Budillon, and Emin Ozsoy. The eastern mediterranean in the 80s and in the 90s: the big transition in the intermediate and deep circulations. *Dynamics of Atmospheres and Oceans*, 29(2-4):365–395, 1999.

[54] Wolfgang Ludwig, Egon Dumont, Michel Meybeck, and Serge Heussner. River discharges of water and nutrients to the mediterranean and black sea: major drivers for ecosystem changes during past and future decades? *Progress in oceanography*, 80(3-4):199–217, 2009.

[55] EJ Rohling, G Marino, and KM Grant. Mediterranean climate and oceanography, and the periodic development of anoxic events (sapropels). *Earth-Science Reviews*, 143:62–97, 2015.

[56] Paul Falkowski, John Woodhead, and Katherine Vivirito. *Biogeochemical cycles: A treatise on global change*. Academic Press, 1998.

[57] William H Schlesinger and Emily S Bernhardt. *Biogeochemistry: an analysis of global change*. Academic press, 2013.

[58] Alfred C Redfield. The biological control of chemical factors in the environment. *American Scientist*, 46(3):230A–221, 1958.

[59] Susan M Libes. *Introduction to marine biogeochemistry*. Academic press, 2009.

[60] Walter Munk. Oceanography before, and after, the advent of satellites. In *Elsevier Oceanography Series*, volume 63, pages 1–4. Elsevier, 2000.

[61] Hervé Claustre, Kenneth S Johnson, and Yuichiro Takeshita. Observing the global ocean with biogeochemical-argo. *Annual review of Marine science*, 12:23–48, 2020.

[62] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197, 1991.

[63] L Buga, L Boicenco, A Giorgetti, G Sarbu, A Spinu, et al. Emodnet chemistry–data aggregation and product generations in the black sea. *Journal of Environmental Protection and Ecology*, 19(1):300–308, 2018.

[64] Henry C Bittig, Tanya L Maurer, Joshua N Plant, Catherine Schmechtig, Annie PS Wong, Hervé Claustre, Thomas W Trull, TVS Udaya Bhaskar, Emmanuel Boss, Giorgio Dall'Olmo, et al. A bgc-argo guide: Planning, deployment, data handling and usage. *Frontiers in Marine Science*, 6:502, 2019.

[65] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.

[66] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

[67] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[68] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[69] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[70] Anders Krogh. What are artificial neural networks? *Nature biotechnology*, 26(2): 195–197, 2008.

[71] Qing Rao and Jelena Frtunikj. Deep learning for self-driving cars: Chances and challenges. In *Proceedings of the 1st international workshop on software engineering for AI in autonomous systems*, pages 35–38, 2018.

[72] Mihalj Bakator and Dragica Radosav. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3):47, 2018.

[73] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[74] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[75] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49. Springer, 2008.

[76] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[77] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.

[78] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[79] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676): 354–359, 2017.

[80] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[81] Jens Kober and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[82] X Zhao, W Zhang, and J Wang. Interactive collaborative filtering. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1411–1420. ACM, 2013.

[83] Jinming Zou, Yi Han, and Sung-Sau So. Overview of artificial neural networks. *Artificial neural networks: methods and applications*, pages 14–22, 2009.

[84] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[85] Murray Rosenblatt. Independence and dependence. In *Proc. 4th Berkeley sympos. math. statist. and prob*, volume 2, pages 431–443, 1961.

[86] Marvin Minsky and Seymour Papert. An introduction to computational geometry. *Cambridge tiass., HIT*, 479(480):104, 1969.

[87] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[88] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[89] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.

[90] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[91] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

[92] Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer, 2012.

[93] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[94] Sarit Khirirat, Hamid Reza Feyzmahdavian, and Mikael Johansson. Mini-batch gradient descent: Faster convergence under data sparsity. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2880–2887. IEEE, 2017.

[95] David E Rumelhart, Richard Durbin, Richard Golden, and Yves Chauvin. Backpropagation: The basic theory. In *Backpropagation*, pages 1–34. Psychology Press, 2013.

[96] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[97] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[98] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[99] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Overview of supervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 9–41, 2009.

[100] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[101] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[102] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[103] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[104] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[105] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[106] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[107] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[108] T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

[109] Xuan Huang, Lei Wu, and Yinsong Ye. A review on dimensionality reduction techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, 33 (10):1950017, 2019.

[110] Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021.

[111] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[112] Hang Zhou, Ruiyao Cai, Tingwei Quan, Shijie Liu, Shiwei Li, Qing Huang, Ali Ertürk, and Shaoqun Zeng. 3d high resolution generative deep-learning network for fluorescence microscopy imaging. *Optics letters*, 45(7):1695–1698, 2020.

[113] Atsushi Teramoto, Tetsuya Tsukamoto, Ayumi Yamada, Yuka Kiriyama, Kazuyoshi Imaizumi, Kuniaki Saito, and Hiroshi Fujita. Deep learning approach to classification of lung cytological images: Two-step training using actual and synthesized images by progressive growing of generative adversarial networks. *PloS one*, 15(3):e0229951, 2020.

[114] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 299–315. Springer, 2020.

[115] Devendra Prakash Jaiswal, Srishti Kumar, and Youakim Badr. Towards an artificial intelligence aided design approach: application to anime faces with generative adversarial networks. *Procedia Computer Science*, 168:57–64, 2020.

[116] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[117] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[118] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[119] M Durgadevi et al. Generative adversarial network (gan): a general review on different variants of gan and applications. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pages 1–8. IEEE, 2021.

[120] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems*, 28, 2015.

[121] Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*, 2016.

[122] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[123] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, and Fei-Yue Wang. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, 2017.

[124] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[125] Anat Levin, Assaf Zomet, Shmuel Peleg, and Yair Weiss. Seamless image stitching in the gradient domain. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*, pages 377–389. Springer, 2004.

[126] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *Siam journal on imaging sciences*, 7(4):1993–2019, 2014.

[127] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[128] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, 51:2007–2028, 2020.

[129] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001.

[130] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999.

[131] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: Image and video synthesis using graph cuts. *Acm transactions on graphics (tog)*, 22(3):277–286, 2003.

[132] Rolf Köhler, Christian Schuler, Bernhard Schölkopf, and Stefan Harmeling. Mask-specific inpainting with deep neural networks. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 523–534. Springer, 2014.

[133] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[134] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3911–3919, 2017.

[135] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[136] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[137] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[138] Zhaoyi Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. High-resolution image inpainting using multi-scale neural patch synthesis. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[139] David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383, 1977.

[140] Geoffrey P Goodwin and PN Johnson-Laird. Reasoning about relations. *Psychological review*, 112(2):468, 2005.

[141] Matthew M Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, 12(5):201–208, 2008.

[142] Noam Chomsky. Logical structure in language. *Journal of the American Society for Information Science*, 8(4):284, 1957.

[143] David Heckerman. A tutorial on learning with bayesian networks. *Innovations in Bayesian networks: Theory and applications*, pages 33–82, 2008.

[144] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[145] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[146] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[147] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.

[148] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7): 1235–1270, 2019.

[149] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

[150] Balamurugan Sadaiappan, Preethiya Balakrishnan, Vishal CR, Neethu T Vijayan, Mahendran Subramanian, and Mangesh U Gauns. Applications of machine learning in chemical and biological oceanography. *ACS omega*, 8(18):15831–15853, 2023.

[151] Changming Dong, Guangjun Xu, Guoqing Han, Brandon J Bethel, Wenhong Xie, and Shuyi Zhou. Recent developments in artificial intelligence in oceanography. *Ocean-Land-Atmosphere Research*, 2022, 2022.

[152] Stéphanie Manel, Jean-Marie Dias, and Steve J Ormerod. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a himalayan river bird. *Ecological modelling*, 120(2-3): 337–347, 1999.

[153] Jane Elith*, Catherine H. Graham*, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

[154] Anantha M Prasad, Louis R Iverson, and Andy Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9:181–199, 2006.

[155] Anne E Thessen. Adoption of machine learning techniques in ecology and earth science. Technical report, PeerJ PrePrints, 2016.

[156] Friedrich Recknagel. Applications of machine learning to ecological modelling. *Ecological modelling*, 146(1-3):303–310, 2001.

[157] Glenn De'ath and Katharina E Fabricius. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192, 2000.

[158] Maike Sonnewald, Redouane Lguensat, Daniel C Jones, Peter Dueben, Julien Brajard, and Venkatramani Balaji. Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, 2021.

[159] Richard H Norris and Martin C Thoms. What is river health? *Freshwater biology*, 41(2):197–209, 1999.

[160] Angel Borja, Suzanne B Bricker, Daniel M Dauer, Nicolette T Demetriades, João G Ferreira, Anthony T Forbes, Pat Hutchings, Xiaoping Jia, Richard Kenchington, João Carlos Marques, et al. Overview of integrative tools and methods in assessing ecological integrity in estuarine and coastal systems worldwide. *Marine pollution bulletin*, 56(9):1519–1537, 2008.

[161] Stephanie E Hampton, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3): 156–162, 2013.

[162] KA Dafforn, EL Johnston, Alastair Ferguson, CL Humphrey, Wendy Monk, SJ Nichols, SL Simpson, MG Tulbure, and DJ Baird. Big data opportunities and challenges for assessing multiple stressors across scales in aquatic ecosystems. *Marine and Freshwater Research*, 67(4):393–413, 2015.

[163] Robert N Miller. *Numerical modeling of ocean circulation*. Cambridge University Press, 2007.

[164] Baylor Fox-Kemper, Alistair Adcroft, Claus W Böning, Eric P Chassignet, Enrique Curchitser, Gokhan Danabasoglu, Carsten Eden, Matthew H England, Rüdiger Gerdes, Richard J Greatbatch, et al. Challenges and prospects in ocean circulation models. *Frontiers in Marine Science*, 6:65, 2019.

[165] Curtis D Mobley, Bernard Gentili, Howard R Gordon, Zhonghai Jin, George W Kattawar, Andre Morel, Phillip Reinersman, Knut Stamnes, and Robert H Stavn. Comparison of numerical models for computing underwater light fields. *Applied Optics*, 32(36):7484–7504, 1993.

[166] Argo webpage. `https://argo.ucsd.edu/`. Accessed: 2023-08-10.

[167] Robert M Key, Are Olsen, Steven van Heuven, Siv K Lauvset, Anton Velo, Xiaohua Lin, Carsten Schirnick, Alex Kozyr, Toste Tanhua, Mario Hoppema, et al. Global ocean data analysis project, version 2 (glodapv2). *Ornl/Cdiac-162, Ndp-093*, 2015.

[168] Are Olsen, Robert M Key, Steven Van Heuven, Siv K Lauvset, Anton Velo, Xiaohua Lin, Carsten Schirnick, Alex Kozyr, Toste Tanhua, Mario Hoppema, et al. The global ocean data analysis project version 2 (glodapv2)–an internally consistent data product for the world ocean. *Earth System Science Data*, 8(2):297–323, 2016.

[169] Henry C Bittig, Tobias Steinhoff, Hervé Claustre, Björn Fiedler, Nancy L Williams, Raphaëlle Sauzède, Arne Körtzinger, and Jean-Pierre Gattuso. An alternative to

static climatologies: Robust estimation of open ocean co2 variables and nutrient concentrations from t, s, and o2 data using bayesian neural networks. *Frontiers in Marine Science*, 5:328, 2018.

[170] OceanOPS webpage. https://www.ocean-ops.org/board. Accessed: 2023-08-10.

[171] Global climate observing system. https://gcos.wmo.int/en/home. Accessed: 2023-08-11.

[172] Arne Kortzinger, Jens Schimanski, Uwe Send, and Douglas Wallace. The ocean takes a deep breath. *Science*, 306(5700):1337–1337, 2004.

[173] Stephen C Riser and Kenneth S Johnson. Net production of oxygen in the subtropical ocean. *Nature*, 451(7176):323–325, 2008.

[174] Dean Roemmich, Matthew H Alford, Hervé Claustre, Kenneth Johnson, Brian King, James Moum, Peter Oke, W Brechner Owens, Sylvie Pouliquen, Sarah Purkey, et al. On the future of argo: A global, full-depth, multi-disciplinary array. *Frontiers in Marine Science*, 6:439, 2019.

[175] Kenneth S Johnson, Stephen C Riser, and David M Karl. Nitrate supply from deep to near-surface waters of the north pacific subtropical gyre. *Nature*, 465 (7301):1062–1065, 2010.

[176] E Boss and M Behrenfeld. In situ evaluation of the initiation of the north atlantic phytoplankton bloom. *Geophysical Research Letters*, 37(18), 2010.

[177] Amanda L Whitmire, Ricardo M Letelier, Victor Villagrán, and Osvaldo Ulloa. Autonomous observations of in vivo fluorescence and particle backscattering in an oceanic oxygen minimum zone. *Optics express*, 17(24):21992–22004, 2009.

[178] Seth M Bushinsky, Alison R Gray, Kenneth S Johnson, and Jorge L Sarmiento. Oxygen in the southern ocean from argo floats: Determination of processes driving air-sea fluxes. *Journal of Geophysical Research: Oceans*, 122(11):8661–8682, 2017.

[179] NL Williams, LW Juranek, RA Feely, KS Johnson, Jorge Louis Sarmiento, LD Talley, AG Dickson, AR Gray, R Wanninkhof, JL Russell, et al. Calculating surface ocean pco2 from biogeochemical argo floats equipped with ph: An uncertainty analysis. *Global Biogeochemical Cycles*, 31(3):591–604, 2017.

[180] A. P. Wong et al. Instrumental drift in argo float sensors and its implications for oceanographic research. *Journal of Ocean Technology*, 20(4):123–138, 2015.

[181] B. Claus et al. Challenges in real-time data transmission from argo floats. *Oceanographic Data Systems Journal*, 31(2):241–252, 2016.

[182] M. Takahashi et al. Environmental impacts and considerations for ocean-based observational instruments: A case study of argo floats. *Marine Pollution Bulletin*, 78(1):18–25, 2015.

[183] Felisberto Pereira, Ricardo Correia, and Nuno Borges Carvalho. Comparison of active and passive sensors for iot applications. In *2018 IEEE Wireless Power Transfer Conference (WPTC)*, pages 1–3. IEEE, 2018.

[184] Robert Frouin, Sam F Iacobellis, and Pierre-Yves Deschamps. Satellite ocean color observations of the tropical pacific ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, 41(2):271–288, 1994.

[185] Ewa J Kwiatkowska. Evolution of the radiometric calibration of seawifs. *Remote Sensing*, 11(10):1205, 2019.

[186] CS Bretherton and CL Smith. Decadal variations in climate associated with the north atlantic oscillation. *Climatic Change*, 14(3-4):213–234, 1990.

[187] Curtis D Mobley, Lydia K Sundman, and Emmanuel Boss. Underwater light field modulated by optically shallow bottoms: Modeling and experiments. *Limnology and Oceanography*, 49(1part2):433–445, 2004.

[188] Trevor Platt and Shubha Sathyendranath. Remote sensing of marine biogeochemistry and ecology: the re-emergence of bio-optics. *Progress in Oceanography*, 77 (2-3):91–96, 2008.

[189] Richard W Reynolds, Nick A Rayner, Thomas M Smith, Diane C Stokes, and Wanqiu Wang. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22):5473–5496, 2007.

[190] Shubha Sathyendranath, Albert D Gouveia, Satish R Shetye, Ponnumony Ravindran, and Trevor Platt. Remote sensing of phytoplankton pigments: a comparison of empirical and theoretical approaches. *International Journal of Remote Sensing*, 25(7-8):1469–1477, 2004.

[191] Mark R Drinkwater. Satellite observations of arctic change. *Arctic*, pages 147–161, 2001.

[192] James KB Bishop. Barriers to physical controls of co2 distributions in the atlantic ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, 51(10-11):1355–1380, 2004.

[193] Seelye Martin and Stanford B Hooker. Satellite remote sensing of surface ocean pco2. *Remote Sensing of Environment*, 82(2-3):269–280, 2002.

[194] Howard R Gordon and Minghua Wang. Atmospheric correction of ocean color imagery in the earth observing system era. *Journal of Geophysical Research: Atmospheres*, 102(D14):17081–17106, 1994.

[195] Frank J Wentz. Measurement of oceanic wind vectors using satellite microwave radiometers. *IEEE Transactions on Geoscience and Remote Sensing*, 30(5):960–972, 1992.

[196] Sean W Bailey and P Jeremy Werdell. A comparison of satellite bio-optical algorithms for global ocean chlorophyll-a. *Remote Sensing of Environment*, 100(3): 302–333, 2006.

[197] Philip A Harrison, John M Huthnance, and Alan Greig. Satellite oceanography from the ers synthetic aperture radar and radar altimeter: A brief review. *Advances in Space Research*, 44(8):1006–1015, 2009.

[198] Michael JR Fasham, Hugh W Ducklow, and Stuart M McKelvie. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. *Journal of Marine Research*, 48(3):591–639, 1990.

[199] Stephen M Griffies and Kirk L Bryan. Numerical modeling for oceanographers: Insights from decades of experience. *Ocean Modelling*, 110:74–90, 2017.

[200] Jessica S Hindell and Eric P Chassignet. Numerical models in oceanography: Challenges and limitations. *Journal of Marine Research*, 78(2):123–140, 2020.

[201] Christine L Borgman, Peter T Darch, Ashley E Sands, Irene V Pasquetto, Milena S Golshan, Jillian C Wallis, and Sharon Traweek. Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries*, 16:207–227, 2015.

[202] John Boyle. Biology must develop its own big-data systems. *Nature*, 499(7456): 7–7, 2013.

[203] Detlef Stammer, Richard D Ray, Ole B Andersen, Brian K Arbic, Wolfgang Bosch, Laurent Carrère, Yannan Cheng, Douglas S Chinn, Brian D Dushaw, Gary D Egbert, et al. Accuracy assessment of global barotropic ocean tide models. *Reviews of Geophysics*, 52(3):243–282, 2014.

[204] Carlo Heip, Herman Hummel, Pim van Avesaath, Ward Appeltans, Christos Arvanitidis, Rebecca Aspden, Melanie Austen, Ferdinando Boero, Tjeerd J Bouma,

Geoff Boxshall, et al. Marine biodiversity and ecosystem functioning: report of an international workshop. *Marine Ecology Progress Series*, 411:1–8, 2010.

[205] Peter Landschützer, Nicolas Gruber, Dorothee CE Bakker, and Ute Schuster. Recent variability of the global ocean carbon sink. *Global Biogeochemical Cycles*, 28 (9):927–949, 2014.

[206] Donata Giglio, Vyacheslav Lyubchich, and Matthew R Mazloff. Estimating oxygen in the southern ocean using argo temperature and salinity. *Journal of Geophysical Research: Oceans*, 123(6):4280–4297, 2018.

[207] Seth M Bushinsky, Peter Landschützer, Christian Rödenbeck, Alison R Gray, David Baker, Matthew R Mazloff, Laure Resplandy, Kenneth S Johnson, and Jorge L Sarmiento. Reassessing southern ocean air-sea co2 flux estimates with the addition of biogeochemical float observations. *Global Biogeochemical Cycles*, 33(11):1370–1388, 2019.

[208] Andrew J Watson, Ute Schuster, Jamie D Shutler, Thomas Holding, Ian GC Ashton, Peter Landschützer, David K Woolf, and Lonneke Goddijn-Murphy. Revised estimates of ocean-atmosphere co2 flux are consistent with ocean carbon inventory. *Nature communications*, 11(1):4422, 2020.

[209] Guillaume Maze, Herlé Mercier, Ronan Fablet, Pierre Tandeo, Manuel Lopez Radcenco, Philippe Lenca, Charlène Feucher, and Clement Le Goff. Coherent heat patterns revealed by unsupervised classification of argo temperature profiles in the north atlantic ocean. *Progress in Oceanography*, 151:275–292, 2017.

[210] Daniel C Jones, Harry J Holt, Andrew JS Meijers, and Emily Shuckburgh. Unsupervised clustering of southern ocean argo float temperature profiles. *Journal of Geophysical Research: Oceans*, 124(1):390–402, 2019.

[211] Isabel A Houghton and James D Wilson. El niño detection via unsupervised clustering of argo temperature profiles. *Journal of Geophysical Research: Oceans*, 125(9):e2019JC015947, 2020.

[212] Isabella Rosso, Matthew R Mazloff, Lynne D Talley, Sarah G Purkey, Natalie M Freeman, and Guillaume Maze. Water mass and biogeochemical variability in the kerguelen sector of the southern ocean: A machine learning approach for a mixing hot spot. *Journal of Geophysical Research: Oceans*, 125(3):e2019JC015877, 2020.

[213] Edward N Lorenz. *Empirical orthogonal functions and statistical weather prediction*, volume 1. Massachusetts Institute of Technology, Department of Meteorology Cambridge, 1956.

[214] Maike Sonnewald, Stephanie Dutkiewicz, Christopher Hill, and Gael Forget. Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science advances*, 6(22):eaay4740, 2020.

[215] Raphaëlle Sauzède, Henry C Bittig, Hervé Claustre, Orens Pasqueron de Fommervault, Jean-Pierre Gattuso, Louis Legendre, and Kenneth S Johnson. Estimates of water-column nutrient concentrations and carbonate system parameters in the global ocean: a novel approach based on neural networks. *Frontiers in Marine Science*, 4:128, 2017.

[216] Marine Fourrier, Laurent Coppola, Hervé Claustre, Fabrizio D'Ortenzio, Raphaëlle Sauzède, and Jean-Pierre Gattuso. A regional neural network approach to estimate water-column nutrient concentrations and carbonate system variables in the mediterranean sea: Canyon-med. *Frontiers in Marine Science*, 7: 620, 2020.

[217] S Thiria, C Mejia, F Badran, and M Crepon. A neural network approach for modeling nonlinear transfer functions: Application for wind retrieval from spaceborne scatterometer data. *Journal of Geophysical Research: Oceans*, 98(C12):22827–22841, 1993.

[218] Zied Ben Mustapha, Séverine Alvain, Cédric Jamet, Hubert Loisel, and David Dessailly. Automatic classification of water-leaving radiance anomalies from global seawifs imagery: application to the detection of phytoplankton groups in open ocean waters. *Remote sensing of environment*, 146:97–112, 2014.

[219] Christopher Chapman and Anastase Alexandre Charantonis. Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps. *IEEE Geoscience and Remote Sensing Letters*, 14(5):617–620, 2017.

[220] Anna Denvil-Sommer, Marion Gehlen, Mathieu Vrac, and Carlos Mejia. Lsce-ffnn-v1: a two-step neural network model for the reconstruction of surface ocean¡ i¿ p¡/i¿ co¡ sub¿ 2¡/sub¿ over the global ocean. *Geoscientific Model Development*, 12(5):2091–2105, 2019.

[221] Elodie Martinez, Thomas Gorgues, Matthieu Lengaigne, Clement Fontana, Raphaëlle Sauzède, Christophe Menkes, Julia Uitz, Emanuele Di Lorenzo, and Ronan Fablet. Reconstructing global chlorophyll-a variations using a non-linear statistical approach. *Frontiers in Marine Science*, 7:464, 2020.

[222] Marco Castellani. Identification of eddies from sea surface temperature maps with neural networks. *International journal of remote sensing*, 27(8):1601–1618, 2006.

[223] David Ian Duncan, Patrick Eriksson, Simon Pfreundschuh, Christian Klepp, and Daniel C Jones. On the distinctiveness of observed oceanic raindrop distributions. *Atmospheric Chemistry and Physics*, 19(10):6969–6984, 2019.

[224] PY Le Traon, F Nadal, and N Ducet. An improved mapping method of multisatellite altimeter data. *Journal of atmospheric and oceanic technology*, 15(2):522–534, 1998.

[225] Aida Alvera-Azcárate, Alexander Barth, Gaëlle Parard, and Jean-Marie Beckers. Analysis of smos sea surface salinity data using dineof. *Remote sensing of environment*, 180:137–145, 2016.

[226] Alexander Barth, Aida Alvera-Azcárate, Matjaz Licer, and Jean-Marie Beckers. Dincae 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations. *Geoscientific Model Development*, 13 (3):1609–1622, 2020.

[227] Thomas Bolton and Laure Zanna. Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, 11(1):376–399, 2019.

[228] Laure Zanna and Thomas Bolton. Data-driven equation discovery of ocean mesoscale closures. *Geophysical Research Letters*, 47(17):e2020GL088376, 2020.

[229] Julien Brajard, Alberto Carrassi, Marc Bocquet, and Laurent Bertino. Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A*, 379(2194):20200086, 2021.

[230] Niraj Agarwal, Dmitri Kondrashov, Peter Dueben, Eugene Ryzhov, and Pavel Berloff. A comparison of data-driven approaches to build low-dimensional ocean models. *Journal of Advances in Modeling Earth Systems*, 13(9):e2021MS002537, 2021.

[231] Peer Nowack, Peter Braesicke, Joanna Haigh, Nathan Luke Abraham, John Pyle, and Apostolos Voulgarakis. Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. *Environmental Research Letters*, 13(10):104016, 2018.

[232] Peter D Dueben and Peter Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10):3999–4009, 2018.

[233] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11): e2020MS002203, 2020.

[234] Rachel Furner, Peter Haynes, Dave Munday, Brooks Paige, Daniel C Jones, and Emily Shuckburgh. Sensitivity analysis of a data-driven model of ocean temperature. *Geoscientific Model Development Discussions*, pages 1–33, 2021.

[235] Allan R Robinson, Michael A Spall, Leonard J Walstad, and Wayne G Leslie. Data assimilation and dynamical interpolation in gulfcast experiments. *Dynamics of atmospheres and oceans*, 13(3-4):301–316, 1989.

[236] Carl Wunsch. Towards the world ocean circulation experiment and a bit of aftermath. In *Physical oceanography: developments since 1950*, pages 181–201. Springer, 2006.

[237] MichAEl J BEll, Michel Lefèbvre, Pierre-Yves Le Traon, Neville Smith, and Kirsten Wilmer-Becker. Godae: the global ocean data assimilation experiment. *Oceanography*, 22(3):14–21, 2009.

[238] Michael Ghil and Paola Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. In *Advances in geophysics*, volume 33, pages 141–266. Elsevier, 1991.

[239] Henry DI Abarbanel, Paul J Rozdeba, and Sasha Shirman. Machine learning: Deepest learning as statistical data assimilation problems. *Neural computation*, 30 (8):2025–2055, 2018.

[240] Marc Bocquet, Julien Brajard, Alberto Carrassi, and Laurent Bertino. Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Processes in Geophysics*, 26(3):143–162, 2019.

[241] Alan J Geer. Learning earth system models from observations: machine learning or data assimilation? *Philosophical Transactions of the Royal Society A*, 379(2194): 20200089, 2021.

[242] Redouane Lguensat, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet. The analog data assimilation. *Monthly Weather Review*, 145(10):4093–4107, 2017.

[243] Massimo Bonavita and Patrick Laloyaux. Machine learning for model error inference and correction. *Journal of Advances in Modeling Earth Systems*, 12(12): e2020MS002232, 2020.

[244] SG Penny, E Bach, K Bhargava, C-C Chang, C Da, L Sun, and T Yoshida. Strongly coupled data assimilation in multiscale media: Experiments using a quasi-geostrophic coupled model. *Journal of Advances in Modeling Earth Systems*, 11(6):1803–1829, 2019.

[245] Maddalena Amendola, Rossella Arcucci, Laetitia Mottet, Cesar Quilodran Casas, Shiwei Fan, Christopher Pain, Paul Linden, and Yi-Ke Guo. Data assimilation in the latent space of a neural network. *arXiv preprint arXiv:2012.12056*, 2020.

[246] Ronan Fablet, Bertrand Chapron, Lucas Drumetz, Etienne Mémin, Olivier Pannekoucke, and François Rousseau. Learning variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*, 13(10):e2021MS002572, 2021.

[247] Julian Mack, Rossella Arcucci, Miguel Molina-Solana, and Yi-Ke Guo. Attention-based convolutional autoencoders for 3d-variational data assimilation. *Computer Methods in Applied Mechanics and Engineering*, 372:113291, 2020.

[248] Kenneth Johnson and Hervé Claustre. Bringing biogeochemistry into the argo age. *Eos, Transactions American Geophysical Union*, 2016.

[249] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[250] David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469, 1995.

[251] Siv K Lauvset, Robert M Key, Are Olsen, Steven Van Heuven, Anton Velo, Xiaohua Lin, Carsten Schirnick, Alex Kozyr, Toste Tanhua, Mario Hoppema, et al. A new global interior ocean mapped climatology: The $1\times$ 1 glodap version 2. *Earth System Science Data*, 8(2):325–340, 2016.

[252] Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

[253] Patricia Miloslavich, Sophie Seeyave, Frank Muller-Karger, Nicholas Bax, Elham Ali, Claudia Delgado, Hayley Evers-King, Benjamin Loveday, Vivian Lutz, Jan Newton, et al. Challenges for global ocean observation: the need for increased human capacity. *Journal of Operational Oceanography*, 12(sup2):S137–S156, 2019.

[254] G Cossarini, L Mariotti, L Feudale, A Mignot, S Salon, V Taillandier, A Teruzzi, and F d'Ortenzio. Towards operational 3d-var assimilation of chlorophyll biogeochemical-argo float data into a biogeochemical model of the mediterranean sea. *Ocean Modelling*, 133:112–128, 2019.

[255] Anna Teruzzi, Giorgio Bolzon, Laura Feudale, and Gianpiero Cossarini. Deep chlorophyll maximum and nutricline in the mediterranean sea: emerging properties from a multi-platform assimilated biogeochemical model experiment. *Biogeosciences*, 18(23):6147–6166, 2021.

[256] Caterina Buizza, César Quilodrán Casas, Philip Nadler, Julian Mack, Stefano Marrone, Zainab Titus, Clémence Le Cornec, Evelyn Heylen, Tolga Dur, Luis Baca Ruiz, et al. Data learning: Integrating data assimilation and machine learning. *Journal of Computational Science*, 58:101525, 2022.

[257] Anna Teruzzi, Srdjan Dobricic, Cosimo Solidoro, and Gianpiero Cossarini. A 3-d variational assimilation scheme in coupled transport-biogeochemical models: Forecast of mediterranean biogeochemical properties. *Journal of Geophysical Research: Oceans*, 119(1):200–217, 2014.

[258] Anna Teruzzi, Giorgio Bolzon, Stefano Salon, Paolo Lazzari, Cosimo Solidoro, and Gianpiero Cossarini. Assimilation of coastal and open sea biogeochemical data to improve phytoplankton simulation in the mediterranean sea. *Ocean Modelling*, 132:46–60, 2018.

[259] The european marine observation and data network. URL `https://emodnet.ec.europa.eu/en`. `https://emodnet.ec.europa.eu/en` (accessed: 13.09.2021).

[260] Flora Sakketou and Nicholas Ampazis. On the invariance of the selu activation function on algorithm and hyperparameter selection in neural network recommenders. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 673–685. Springer, 2019.

[261] Leonardo Noriega. Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, 2005.

[262] David W Hosmer and Stanley Lemeshow. Confidence interval estimation of interaction. *Epidemiology*, pages 452–456, 1992.

[263] Tim Pearce, Mohamed Zaki, Alexandra Brintrup, N Anastassacos, and A Neely. Uncertainty in neural networks: Bayesian ensembling. *stat*, 1050:12, 2018.

[264] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[265] Stefano Salon, Gianpiero Cossarini, Giorgio Bolzon, Laura Feudale, Paolo Lazzari, Anna Teruzzi, Cosimo Solidoro, and Alessandro Crise. Marine ecosystem forecasts: Skill performance of the cmems mediterranean sea model system. *Ocean Sci. Discuss*, 1:35, 2019.

[266] Fabrizio d'Ortenzio, Hélöise Lavigne, Florent Besson, Hervé Claustre, Laurent Coppola, Nicole Garcia, Agathe Laës-Huon, Serge Le Reste, Damien Malardé, Christophe Migon, et al. Observing mixed layer depth, nitrate and chlorophyll concentrations in the northwestern mediterranean: A combined satellite and no3 profiling floats experiment. *Geophysical Research Letters*, 41(18):6443–6451, 2014.

[267] Alexandre Mignot, Hervé Claustre, Julia Uitz, Antoine Poteau, Fabrizio d'Ortenzio, and Xiaogang Xing. Understanding the seasonal dynamics of phytoplankton biomass and the deep chlorophyll maximum in oligotrophic environments: A bio-argo float investigation. *Global Biogeochemical Cycles*, 28(8):856–876, 2014.

[268] Steven R Jayne, Dean Roemmich, Nathalie Zilberman, Stephen C Riser, Kenneth S Johnson, Gregory C Johnson, and Stephen R Piotrowicz. The argo program: present and future. *Oceanography*, 30(2):18–28, 2017.

[269] ZQ Li, ZH Liu, and SL Lu. Global argo data fast receiving and post-quality-control system. In *IOP Conference Series: Earth and Environmental Science*, volume 502, page 012012. IOP Publishing, 2020.

[270] Giorgio Dall'Olmo, Udaya Bhaskar TVS, Henry Bittig, Emmanuel Boss, Jodi Brewster, Hervé Claustre, Matt Donnelly, Tanya Maurer, David Nicholson, Violetta Paba, et al. Real-time quality control of optical backscattering data from biogeochemical-argo floats. *Open Research Europe*, 2(118):118, 2022.

[271] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Jing Jiang, and Michael Blumenstein. Rethinking 1d-cnn for time series classification: A stronger baseline. *arXiv preprint arXiv:2002.10061*, pages 1–7, 2020.

[272] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

[273] Andrinandrasana David Rasamoelina, Fouzia Adjailia, and Peter Sinčák. A review of activation function for artificial neural network. In *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 281–286. IEEE, 2020.

[274] Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26, 2013.

[275] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2): 301–320, 2005.

[276] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[277] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[278] Craig J Donlon, Matthew Martin, John Stark, Jonah Roberts-Jones, Emma Fiedler, and Werenfrid Wimmer. The operational sea surface temperature and sea ice analysis (ostia) system. *Remote Sensing of Environment*, 116:140–158, 2012.

[279] Gianluca Volpe, Bruno Buongiorno Nardelli, Simone Colella, Andrea Pisano, and Rosalia Santoleri. Operational interpolated ocean colour product in the mediterranean sea. *New Frontiers in Operational Oceanography*, pages 227–244, 2018.

[280] Michela Sammartino, Bruno Buongiorno Nardelli, Salvatore Marullo, and Rosalia Santoleri. An artificial neural network to infer the mediterranean 3d chlorophyll-a and temperature fields from remote sensing observations. *Remote Sensing*, 12(24):4123, 2020.

[281] Aïda Alvera-Azcárate, Alexander Barth, Michel Rixen, and Jean-Marie Beckers. Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the adriatic sea surface temperature. *Ocean Modelling*, 9(4):325–346, 2005.

[282] R Sauzède, Hervé Claustre, J Uitz, C Jamet, G Dall'Olmo, F d'Ortenzio, B Gentili, A Poteau, and C Schmechtig. A neural network-based method for merging ocean color and argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscattering coefficient. *Journal of Geophysical Research: Oceans*, 121(4):2552–2571, 2016.

[283] David Ford. Assimilating synthetic biogeochemical-argo and ocean colour observations into a global ocean model to inform observing system design. *Biogeosciences*, 18(2):509–534, 2021.

[284] Liuqian Yu, Katja Fennel, Laurent Bertino, Mohamad El Gharamti, and Keith R Thompson. Insights on multivariate updates of physical and biogeochemical ocean variables using an ensemble kalman filter and an idealized model of upwelling. *Ocean Modelling*, 126:13–28, 2018.

[285] Katja Fennel, Marion Gehlen, Pierre Brasseur, Christopher W Brown, Stefano Ciavatta, Gianpiero Cossarini, Alessandro Crise, Christopher A Edwards, David

Ford, Marjorie AM Friedrichs, et al. Advancing marine biogeochemical and ecosystem reanalyses and forecasts as tools for monitoring and managing ecosystem health. *Frontiers in Marine Science*, 6:89, 2019.

[286] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[287] Gianpiero Cossarini, Laura Feudale, Anna Teruzzi, Giorgio Bolzon, Gianluca Coidessa, Cosimo Solidoro, Valeria Di Biagio, Carolina Amadio, Paolo Lazzari, Alberto Brosich, et al. High-resolution reanalysis of the mediterranean sea biogeochemistry (1999–2019). *Frontiers in Marine Science*, 8:1537, 2021.

[288] Simone Colella, Federico Falcini, Eleonora Rinaldi, Michela Sammartino, and Rosalia Santoleri. Mediterranean ocean colour chlorophyll trends. *PloS one*, 11 (6):e0155756, 2016.

[289] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. arxiv 2015. *arXiv preprint arXiv:1511.07122*, 615, 2016.

[290] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[291] Kenneth S Johnson, Luke J Coletti, Hans W Jannasch, Carole M Sakamoto, Dana D Swift, and Stephen C Riser. Long-term nitrate measurements in the ocean using the in situ ultraviolet spectrophotometer: sensor integration into the apex profiling float. *Journal of Atmospheric and Oceanic Technology*, 30(8):1854–1866, 2013.

[292] P Le Traon, V Abadie, A Ali, A Behrens, J Staneva, M Hieronymi, and H Krasemann. The copernicus marine service from 2015 to 2021: six years of achievements. 2021.

[293] A Mignot, F d'Ortenzio, V Taillandier, G Cossarini, and S Salon. Quantifying observational errors in biogeochemical-argo oxygen, nitrate, and chlorophyll a concentrations. *Geophysical Research Letters*, 46(8):4330–4337, 2019.

[294] Argo. Argo float data and metadata from global data assembly centre (argo gdac), 2023. URL `https://doi.org/10.17882/42182`.

[295] Pierre-Yves Le Traon. From satellite altimetry to argo and operational oceanography: three revolutions in oceanography. *Ocean Science*, 9(5):901–915, 2013.

[296] Marie Barbieux, Julia Uitz, Bernard Gentili, Orens Pasqueron de Fommervault, Alexandre Mignot, Antoine Poteau, Catherine Schmechtig, Vincent Taillandier, Edouard Leymarie, Christophe Penkerc'h, et al. Bio-optical characterization of subsurface chlorophyll maxima in the mediterranean sea from a biogeochemical-argo float database. *Biogeosciences*, 16(6):1321–1342, 2019.

[297] Fabrizio D'ortenzio, Vincent Taillandier, Hervé Claustre, Louis Marie Prieur, Edouard Leymarie, Alexandre Mignot, Antoine Poteau, Christophe Penkerc'h, and Catherine Marie Schmechtig. Biogeochemical argo: The test case of the naos mediterranean array. *Frontiers in Marine Science*, 7:120, 2020.

[298] Florian Ricour, Arthur Capet, Fabrizio d'Ortenzio, Bruno Delille, and Marilaure Grégoire. Dynamics of the deep chlorophyll maximum in the black sea as depicted by bgc-argo floats. *Biogeosciences*, 18(2):755–774, 2021.

[299] Marie Barbieux, Julia Uitz, Alexandre Mignot, Collin Roesler, Hervé Claustre, Bernard Gentili, Vincent Taillandier, Fabrizio d'Ortenzio, Hubert Loisel, Antoine Poteau, et al. Biological production in two contrasted regions of the mediterranean sea during the oligotrophic period: an estimate based on the diel cycle of optical properties measured by biogeochemical-argo profiling floats. *Biogeosciences*, 19(4):1165–1194, 2022.

[300] Arthur Capet, Emil V Stanev, Jean-Marie Beckers, James W Murray, and Marilaure Grégoire. Decline of the black sea oxygen inventory. *Biogeosciences*, 13(4): 1287–1297, 2016.

[301] Vincent Taillandier, Louis Prieur, Fabrizio d'Ortenzio, Maurizio Ribera d'Alcalà, and Elvira Pulido-Villena. Profiling float observation of thermohaline staircases in the western mediterranean sea and impact on nutrient fluxes. *Biogeosciences*, 17(13):3343–3366, 2020.

[302] Tao Wang, Fei Chai, Xiaogang Xing, Jue Ning, Wensheng Jiang, and Stephen C Riser. Influence of multi-scale dynamics on the vertical nitrate distribution around the kuroshio extension: An investigation based on bgc-argo and satellite data. *Progress In Oceanography*, 193:102543, 2021.

[303] Giorgio Dall'Olmo and Kjell Arne Mork. Carbon export by small particles in the norwegian sea. *Geophysical Research Letters*, 41(8):2921–2927, 2014.

[304] Bin Wang and Katja Fennel. An assessment of vertical carbon flux parameterizations using backscatter data from bgc argo. *Geophysical Research Letters*, 50(3): e2022GL101220, 2023.

[305] Virginie Thierry and Henry Bittig. Argo quality control manual for dissolved oxygen concentration. Report (qualification paper (procedure, accreditation support)), FRANCE, 2021.

[306] Yuichiro Takeshita, Todd R Martz, Kenneth S Johnson, Josh N Plant, Denis Gilbert, Stephen C Riser, Craig Neill, and Bronte Tilbrook. A climatology-based quality control procedure for profiling float oxygen data. *Journal of Geophysical Research: Oceans*, 118(10):5640–5650, 2013.

[307] Tanya L Maurer, Joshua N Plant, and Kenneth S Johnson. Delayed-mode quality control of oxygen, nitrate, and ph data on soccom biogeochemical profiling floats. *Frontiers in Marine Science*, 8:683207, 2021.

[308] Kenneth S Johnson, Joshua N Plant, Stephen C Riser, and Denis Gilbert. Air oxygen calibration of oxygen optodes on a profiling float array. *Journal of Atmospheric and Oceanic Technology*, 32(11):2160–2172, 2015.

[309] Seth M Bushinsky, Steven R Emerson, Stephen C Riser, and Dana D Swift. Accurate oxygen measurements on modified a rgo floats using in situ air calibrations. *Limnology and Oceanography: Methods*, 14(8):491–505, 2016.

[310] Henry C Bittig, Arne Körtzinger, Craig Neill, Eikbert Van Ooijen, Joshua N Plant, Johannes Hahn, Kenneth S Johnson, Bo Yang, and Steven R Emerson. Oxygen optode sensors: principle, characterization, calibration, and application in the ocean. *Frontiers in Marine Science*, 4:429, 2018.

[311] Bin Wang, Katja Fennel, Liuqian Yu, and Christopher Gordon. Assessing the value of biogeochemical argo profiles versus ocean color observations for biogeochemical model optimization in the gulf of mexico. *Biogeosciences*, 17(15): 4059–4074, 2020.

[312] Elena Terzić, Paolo Lazzari, Emanuele Organelli, Cosimo Solidoro, Stefano Salon, Fabrizio d'Ortenzio, and Pascal Conan. Merging bio-optical data from biogeochemical-argo floats and models in marine biogeochemistry. *Biogeosciences*, 16(12):2527–2542, 2019.

[313] Stefano Salon, Gianpiero Cossarini, Giorgio Bolzon, Laura Feudale, Paolo Lazzari, Anna Teruzzi, Cosimo Solidoro, and Alessandro Crise. Novel metrics based on biogeochemical argo data to improve the model uncertainty evaluation of the cmems mediterranean marine ecosystem forecasts. *Ocean Science*, 15(4):997–1022, 2019.

[314] Bin Wang, Katja Fennel, and Liuqian Yu. Can assimilation of satellite observations improve subsurface biological properties in a numerical model? a case study for the gulf of mexico. *Ocean Science*, 17(4):1141–1156, 2021.

[315] Fabrizio d'Ortenzio, V Taillandier, Hervé Claustre, Laurent Coppola, P Conan, Franck Dumas, X Durrieu du Madron, M Fourrier, A Gogou, A Karageorgis, et al. Bgc-argo floats observe nitrate injection and spring phytoplankton increase in the surface layer of levantine sea (eastern mediterranean). *Geophysical Research Letters*, 48(8):e2020GL091649, 2021.

[316] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[317] SV Kumar, CD Peters-Lidard, JA Santanello, RH Reichle, CS Draper, RD Koster, G Nearing, and MF Jasinski. Evaluating the utility of satellite soil moisture retrievals over irrigated areas and the ability of land data assimilation methods to correct for unmodeled processes. *Hydrology and Earth System Sciences*, 19(11): 4463–4478, 2015.

[318] Hao Zhou, Xu Yue, Yadong Lei, Chenguang Tian, Yimian Ma, and Yang Cao. Large contributions of diffuse radiation to global gross primary productivity during 1981–2015. *Global Biogeochemical Cycles*, 35(8):e2021GB006957, 2021.

[319] David J Lary, Gebreab K Zewdie, Xun Liu, Daji Wu, Estelle Levetin, Rebecca J Allee, Nabin Malakar, Annette Walker, Hamse Mussa, Antonio Mannino, et al. Machine learning applications for earth observation. *Earth observation open science and innovation*, 165, 2018.

[320] Andrea Storto, Giovanni De Magistris, Silvia Falchetti, and Paolo Oddo. A neural network–based observation operator for coupled ocean–acoustic variational data assimilation. *Monthly Weather Review*, 149(6):1967–1985, 2021.

[321] Elodie Martinez, Anouar Brini, Thomas Gorgues, Lucas Drumetz, Joana Roussillon, Pierre Tandeo, Guillaume Maze, and Ronan Fablet. Neural network approaches to reconstruct phytoplankton time-series in the global ocean. *Remote Sensing*, 12(24):4156, 2020.

[322] Joana Roussillon, Ronan Fablet, Thomas Gorgues, Lucas Drumetz, Jean Littaye, and Elodie Martinez. A multi-mode convolutional neural network to reconstruct satellite-derived chlorophyll-a time series in the global ocean from physical drivers. *Frontiers in Marine Science*, 10(1077623), 2023.

[323] EMIL V Stanev, KATHRIN Wahle, and JOANNA Staneva. The synergy of data from profiling floats, machine learning and numerical modeling: Case of

the black sea euphotic zone. *Journal of Geophysical Research: Oceans*, 127(8): e2021JC018012, 2022.

[324] Anna Teruzzi, Srdjan Dobricic, Cosimo Solidoro, and Gianpiero Cossarini. A 3-d variational assimilation scheme in coupled transport-biogeochemical models: Forecast of mediterranean biogeochemical properties. *Journal of Geophysical Research: Oceans*, 119(1):200–217, 2014.

[325] Florent Gasparin, Stephanie Guinehut, Chongyuan Mao, Isabelle Mirouze, Elisabeth Rémy, Robert R King, Mathieu Hamon, Rebecca Reid, Andrea Storto, Pierre-Yves Le Traon, et al. Requirements for an integrated in situ atlantic ocean observing system from coordinated observing system simulation experiments. *Frontiers in Marine Science*, 6:83, 2019.

[326] Pierre Yves Le Traon, Antonio Reppucci, Enrique Alvarez Fanjul, Lotfi Aouf, Arno Behrens, Maria Belmonte, Abderrahim Bentamy, Laurent Bertino, Vittorio Ernesto Brando, Matilde Brandt Kreiner, et al. From observation to information and users: The copernicus marine service perspective. *Frontiers in Marine Science*, 6:234, 2019.

[327] Nadia Pinardi, Marco Zavatarelli, Mario Adani, Giovanni Coppini, Claudia Fratianni, Paolo Oddo, Simona Simoncelli, Marina Tonani, Vladislav Lyubartsev, Srdjan Dobricic, et al. Mediterranean sea large-scale low-frequency ocean variability and water mass formation rates from 1987 to 2007: A retrospective analysis. *Progress in Oceanography*, 132:318–332, 2015.

[328] P Oddo, M Adani, N Pinardi, C Fratianni, M Tonani, and D Pettenuzzo. A nested atlantic-mediterranean sea general circulation model for operational forecasting. *Ocean science*, 5(4):461–473, 2009.

[329] Ioanna Siokou-Frangou, Urania Christaki, Maria Grazia Mazzocchi, Marina Montresor, Maurizio Ribera d'Alcalá, Dolors Vaqué, and Adriana Zingone. Plankton in the open mediterranean sea: a review. *Biogeosciences*, 7(5):1543–1586, 2010.

[330] Emilio Marañón, France Van Wambeke, Julia Uitz, Emmanuel S Boss, Céline Dimier, Julie Dinasquet, Anja Engel, Nils Haëntjens, María Pérez-Lorenzo, Vincent Taillandier, et al. Deep maxima of phytoplankton biomass, primary production and bacterial production in the mediterranean sea. *Biogeosciences*, 18(5): 1749–1767, 2021.

[331] Valeria Di Biagio, Stefano Salon, Laura Feudale, and Gianpiero Cossarini. Subsurface oxygen maximum in oligotrophic marine ecosystems: mapping the interaction between physical and biogeochemical processes. *Biogeosciences Discussions*, pages 1–33, 2022.

[332] Marie-Alice Foujols, Marina Lévy, Olivier Aumont, and Gurvan Madec. Opa 8.1 tracer model reference manual. *Institut Pierre Simon Laplace*, page 39, 2000.

[333] P Lazzari, C Solidoro, VALERİA Ibello, S Salon, A Teruzzi, K Béranger, S Colella, and A Crise. Seasonal and inter-annual variability of plankton chlorophyll and primary production in the mediterranean sea: a modelling approach. *Biogeosciences*, 9(1):217–233, 2012.

[334] P Lazzari, C Solidoro, S Salon, and G Bolzon. Spatial variability of phosphate and nitrate in the mediterranean sea: A modeling approach. *Deep Sea Research Part I: Oceanographic Research Papers*, 108:39–52, 2016.

[335] Marcello Vichi, Nadia Pinardi, and Simona Masina. A generalized model of pelagic biogeochemistry for the global ocean ecosystem. part i: Theory. *Journal of Marine Systems*, 64(1-4):89–109, 2007.

[336] Marcello Vichi, Nadia Pinardi, and Simona Masina. A generalized model of pelagic biogeochemistry for the global ocean ecosystem. part i: Theory. *Journal of Marine Systems*, 64(1-4):89–109, 2007.

[337] Giovanni Coppini, Emanuela Clementi, Gianpiero Cossarini, Stefano Salon, Gerasimos Korres, Michalis Ravdas, Rita Lecci, Jenny Pistoia, Anna Chiara Goglio, Massimiliano Drudi, et al. The mediterranean forecasting system. part i: evolution and performance. *EGUsphere*, pages 1–50, 2023.

[338] G Cossarini, P Lazzari, and C Solidoro. Spatiotemporal variability of alkalinity in the mediterranean sea. *Biogeosciences*, 12(6):1647–1658, 2015.

[339] Donata Melaku Canu, Andrea Ghermandi, Paulo ALD Nunes, Paolo Lazzari, Gianpiero Cossarini, and Cosimo Solidoro. Estimating the value of carbon sequestration ecosystem services in the mediterranean sea: An ecological economics approach. *Global Environmental Change*, 32:87–95, 2015.

[340] Srdjan Dobricic, Nadia Pinardi, Mario Adani, Marina Tonani, Claudia Fratianni, Alessandro Bonazzi, and Vicente Fernandez. Daily oceanographic analyses by the mediterranean basin scale assimilation system. *Ocean Science Discussions*, 3 (6):1977–1998, 2006.

[341] Jozef Skakala, David Ford, Jorn Bruggeman, Tom Hull, Jan Kaiser, Robert R King, Benjamin Loveday, Matthew R Palmer, Tim Smyth, Charlotte AJ Williams, et al. Towards a multi-platform assimilative system for north sea biogeochemistry. *Journal of Geophysical Research: Oceans*, 126(4):e2020JC016649, 2021.

[342] Xin Dang, Hanxiang Peng, Xueqin Wang, and Heping Zhang. Theil-sen estimators in a multiple linear regression model. *Olemiss Edu*, 2008.

[343] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[344] Romain Escudier, Emanuela Clementi, Andrea Cipollone, Jenny Pistoia, Massimiliano Drudi, Alessandro Grandi, Vladislav Lyubartsev, Rita Lecci, Ali Aydogdu, Damiano Delrosso, et al. A high resolution reanalysis for the mediterranean sea. *Frontiers in Earth Science*, page 1060, 2021.

[345] G Cossarini, S Querin, and C Solidoro. The continental shelf carbon pump in the northern adriatic sea (mediterranean sea): Influence of wintertime variability. *Ecological Modelling*, 314:118–134, 2015.

[346] Henry C Bittig, Arne Körtzinger, Craig Neill, Eikbert Van Ooijen, Joshua N Plant, Johannes Hahn, Kenneth S Johnson, Bo Yang, and Steven R Emerson. Oxygen optode sensors: principle, characterization, calibration, and application in the ocean. *Frontiers in Marine Science*, 4:429, 2018.

[347] F. Raicich and A. Rampazzo. Observing system simulation experiments for the assessment of temperature sampling strategies in the mediterranean sea. *Annales Geophysicae*, 21(1):151–165, 2003. doi: 10.5194/angeo-21-151-2003. URL https://angeo.copernicus.org/articles/21/151/2003/.

[348] Guy Sisma-Ventura, Nurit Kress, Jacob Silverman, Yaron Gertner, Tal Ozer, Eli Biton, Ayah Lazar, Isaac Gertman, Eyal Rahav, and Barak Herut. Post-eastern mediterranean transient oxygen decline in the deep waters of the southeast mediterranean sea supports weakening of ventilation rates. *Frontiers in Marine Science*, 7:598686, 2021.

[349] Laurent Coppola, Louis Legendre, Dominique Lefevre, Louis Prieur, Vincent Taillandier, and Emilie Diamond Riquier. Seasonal and inter-annual variations of dissolved oxygen in the northwestern mediterranean sea (dyfamed site). *Progress in Oceanography*, 162:187–201, 2018.

[350] Apostolia-Maria Mavropoulou, Vassilios Vervatis, and Sarantis Sofianos. Dissolved oxygen variability in the mediterranean sea. *Journal of Marine Systems*, 208:103348, 2020.

[351] Alan Maria Mancini, Giacomo Bocci, Caterina Morigi, Rocco Gennari, Francesca Lozar, and Alessandra Negri. Past analogues of deoxygenation events in the mediterranean sea: A tool to constrain future impacts. *Journal of Marine Science and Engineering*, 11(3):562, 2023.

[352] Ariane Verdy and Matthew R Mazloff. A data assimilating model for estimating s outhern o cean biogeochemistry. *Journal of Geophysical Research: Oceans*, 122(9): 6968–6988, 2017.

[353] Daniel E Kaufman, Marjorie AM Friedrichs, John CP Hemmings, and Walker O Smith Jr. Assimilating bio-optical glider data during a phytoplankton bloom in the southern ross sea. *Biogeosciences*, 15(1):73–90, 2018.

[354] Shuaitao Wang, Nicolas Flipo, Thomas Romary, and Masihullah Hasanyar. Particle filter for high frequency oxygen data assimilation in river systems. *Environmental Modelling & Software*, 151:105382, 2022.

[355] Andrea Storto, Simona Masina, and Srdjan Dobricic. Estimation and impact of nonuniform horizontal correlation length scales for global ocean physical analyses. *Journal of Atmospheric and Oceanic Technology*, 31(10):2330–2349, 2014.

[356] Raphaëlle Sauzède, J Johnson, Hervé Claustre, G Camps-Valls, and A Ruescas. Estimation of oceanic particulate organic carbon with machine learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2:949–956, 2020.

[357] Marine Fourrier, Laurent Coppola, Hervé Claustre, Fabrizio D'Ortenzio, Raphaëlle Sauzède, and Jean-Pierre Gattuso. Corrigendum: A regional neural network approach to estimate water-column nutrient concentrations and carbonate system variables in the mediterranean sea: Canyon-med. *Frontiers in Marine Science*, 8:650509, 2021.

[358] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021.