UNIVERSITY OF Southampton

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Mark R. Taylor (2025) "Investigating spatio-temporal patterns in subsurface chlorophyll using biogeochemical-Argo floats and novel statistical methods", University of Southampton, School of Ocean and Earth Science, PhD Thesis, pagination.

UNIVERSITY OF SOUTHAMPTON

Faculty of Life Sciences
School of Ocean and Earth Science

Investigating spatio-temporal patterns in subsurface chlorophyll using biogeochemical-Argo floats and novel statistical methods

Volume 1 of 1

by

Mark Robert Taylor
MMath
ORCiD: 0009-0008-6983-6128

A thesis for the degree of
Doctor of Philosophy

November 2025

University of Southampton

Abstract

Faculty of Life Sciences
School of Ocean and Earth Science

Doctor of Philosophy

Investigating spatio-temporal patterns in subsurface chlorophyll using biogeochemical-Argo floats and novel statistical methods

by Mark Robert Taylor

Chlorophyll concentration is a widely used proxy for phytoplankton biomass, and its continued monitoring is essential for understanding phytoplankton's role in the global carbon cycle and as the foundation of marine ecosystems. The work presented in this thesis explores the large-scale spatio-temporal variability of vertical chlorophyll structure, particularly its relationships with environmental conditions, using data collected by Biogeochemical-Argo floats through a suite of statistical approaches. First, a spatio-temporal modelling framework was employed to identify the drivers of subsurface chlorophyll maxima (SCMs) on a global scale. This method pooled observations across space and time and revealed that the euphotic depth (z_{eu}) was the main driver of SCM depth and intensity. However, these insights were limited by the need to extract SCM properties prior to modelling. Consequently, functional regression models were used to examine how environmental conditions influenced the profile shapes of chlorophyll and particle backscatter (b_{bp}), which enabled the study of SCMs within the context of entire profiles. Results showed that SCM depth was primarily governed by the z_{eu}, while peak b_{bp} was linked to the nitracline depth. Additionally, photoacclimation, the physiological response of phytoplankton to low light conditions, emerged as a key driver of SCMs throughout the low latitudes. Finally, a novel measure of variance for oceanographic profiles was applied to chlorophyll and temperature data, and their correlation was assessed. Spatio-temporal autocorrelation of both variables was examined in Eulerian and semi-Lagrangian perspectives. This analysis revealed that the similarity in spatio-temporal length scales of chlorophyll and temperature varies by region and spatial extent of the dataset. In summary, this work highlighted the importance of light in determining the vertical distribution of phytoplankton and how contemporary statistical tools improve ecological insights into subsurface biogeochemical observations from autonomous platforms.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:..... Date:.....

Acknowledgements

Firstly, I would like to thank all my supervisors — Steph Henson, Sujit Sahu, B.B. Cael, and Matt Hammond — for their guidance and support over the past four years. Thank you, Steph, for your advice to keep my work simple; it has helped me become a better scientist in the process. I appreciate your understanding that my interests were often focused on the statistics than the biogeochemistry. I am also grateful for your encouragement to attend conferences, workshops and a research cruise early in the project. Thank you, Sujit, for your assistance in developing the spatio-temporal model and for the excellent guidance provided by your book. Thank you, Cael and Matt, for your support during the early stages of the PhD, especially as I adjusted to working at the intersection of oceanography and statistics. I would also like to thank Marin Cornec, who kindly provided additional guidance on subsurface chlorophyll maxima and BGC-Argo data.

A special thanks to those I met while attending conferences or during time away from Southampton — those experiences were among the highlights of the PhD. It has been a pleasure to feel part of an international community of researchers, whose diverse backgrounds and interests have provided me with plenty of new perspectives. I would also like to thank Gwyn Evans for giving me the opportunity to join the DY146 research expedition, assisting with collecting water samples and conducting dissolved oxygen analyses. It was an unforgettable experience, given how few working hours I have spent away from a computer screen.

Thank you to my family for their support during my PhD, and for always encouraging me to "make the most of it!" I am especially thankful to my friends, housemates, office mates, and colleagues who have made the past few years so enjoyable. I have been very lucky to live and work with such wonderful people, and their friendship, support, and laughter have enriched my life in Southampton in countless ways.

In memory of Gramps,

who always encouraged me
to do what made me happiest.

Definitions and Abbreviations

ACF	Autocorrelation function
AIC	Akaike information criterion
BGC-Argo	Biogeochemical-Argo
Chl	Chlorophyll
Chl_{SCM}	Chlorophyll concentration of subsurface chlorophyll maximum
CO_2	Carbon dioxide
C_{phyto}	Phytoplankton carbon
CTD	Conductivity, temperature and depth
FDA	Functional data analysis
FRM	Functional regression model
GAM	Generalised additive model
GRF	Gaussian random field
GMRF	Gaussian Markov random field
INLA	Integrated nested Laplace approximation
LGM	Latent Gaussian model
MCMC	Markov chain Monte Carlo
ML	Machine learning
MLD	Mixed layer depth
NPQ	Non photochemical quenching
PAR	Photosynthetically available radiation
POC	Particulate organic carbon
QC	Quality control
REML	Restricted maximum likelihood
RMSE	Root mean square error
SAM	Subsurface photoacclimation maxima
SBM	Subsurface biomass maxima
SCM	Subsurface chlorophyll maxima
SPDE	Stochastic partial differential equations
SSH	Sea surface height
SSHA	Sea surface height anomaly
SST	Sea surface temperature
STAS	Spatio-temporal autocorrelation structure

WAIC	Wanatabe Akaike information criterion
WMO	World meteorological organisation
z_{eu}	Euphotic depth
z_{ncline}	Nitracline depth
z_{SCM}	Depth of subsurface chlorophyll maximum

Chapter 1

Introduction

1.1 Phytoplankton and the marine carbon cycle

Phytoplankton are microscopic plant-like organisms that form the foundation of the marine ecosystem. Through photosynthesis, they convert inorganic carbon in the form of carbon dioxide (CO_2) and dissolved nutrients into organic carbon, using sunlight as a source of energy. This process occurs at the top of the water column, where sunlight is strongest. Approximately 45% of all primary production occurs in aquatic ecosystems (Falkowski, 1994), despite the fact that phytoplankton only contributes less than 1% of the global biomass of plants (Field et al., 1998). Phytoplankton contain pigments including chlorophyll, which absorbs energy from the red and blue parts of the electromagnetic spectrum, giving phytoplankton a green appearance when viewed in large quantities. Chlorophyll is essential for photosynthesis, with chlorophyll-a the major light-harvesting pigment (Falkowski and Raven, 2013). There is a wide variety of phytoplankton species, ranging in size by several orders of magnitude (typically 0.2 - 200 µm) (Sieburth et al., 1978). Their small size yields a large surface area-to-volume ratio, which allows a high rate of photosynthesis per unit mass. Different species have different nutrient requirements; for example, diatoms need silicon to build hard shell-like structures. In addition, various species occupy different niches within the marine environment.

The ocean plays a vital role in the global carbon cycle both as a carbon sink and through its exchanges with the atmosphere, due to its large volume and the relatively long residence times of carbon within it (Figure 1.1). Anthropogenic activity has increased atmospheric carbon concentration by 52% above preindustrial levels (Friedlingstein et al., 2024), and 25% of atmospheric CO_2 has been absorbed by the ocean (Gruber et al., 2023). Although the solubility pump is thought to drive the uptake of anthropogenic CO_2, the biological carbon pump is essential to maintain the dissolved inorganic carbon gradient from the surface to the deep ocean. Without

biological uptake and storage of carbon, atmospheric CO_2 concentrations would be approximately 150-200 ppm higher than pre-industrial levels (Parekh et al., 2006; Tjiputra et al., 2025). Phytoplankton remove CO_2 from the sunlit euphotic zone near the surface, and when they die or are consumed by zooplankton, the associated particulate organic carbon may enter the mesopelagic zone. Here, further work by zooplankton and microbial respiration converts some of this organic carbon back into CO_2, but a significant fraction remains stored for hundreds or thousands of years (DeVries, 2022; Siegel et al., 2023). The biological carbon pump is responsible for sequestering approximately 10 Gt of carbon each year (DeVries and Weber, 2017). Given that phytoplankton abundance is not homogeneous throughout the global ocean and varies on multiple temporal scales, the strength of the biological carbon pump is not uniform. There is uncertainty surrounding the role phytoplankton will play in future biogeochemical cycles in the presence of warming oceans (Agusti et al., 2019; Jorda et al., 2020; Browning and Moore, 2023), although it has been reported that phytoplankton biomass may be reducing by 1% yr^{-1} (Boyce et al., 2010). Model projections suggest that global primary production may reduce by 3% by 2100, relative to pre-industrial values (multi-model CMIP mean, SSP5-8.5 scenario) (Kwiatkowski et al., 2020). Therefore, understanding the abundance and distribution of phytoplankton and their role in the marine carbon cycle is essential to predict future changes in the climate system.

1.2 The vertical distribution of phytoplankton and chlorophyll

1.2.1 Phytoplankton growth

Several key nutrients are required for photosynthesis by phytoplankton, typically split into macronutrients (silicate, nitrogen, phosphorus) and micronutrients (e.g. iron and manganese). Solar irradiance is also essential as it provides energy for photosynthesis. As phytoplankton grow and multiply, nutrients are used up, gradually reducing the availability of resources. Consequently, the supply of one of the nutrients, or the light intensity, eventually becomes insufficient for the continued net growth of phytoplankton, once accounting for losses due to mortality. Once the limiting factor becomes available again, the rate of cell division can increase, and another factor may become limiting.

The concentration of nutrients and the intensity of light vary with depth and, consequently, the ability of phytoplankton to grow and reproduce is not uniform throughout the water column. Solar irradiance attenuates as it passes through seawater according to Beer's law, which means that its highest intensity occurs at the

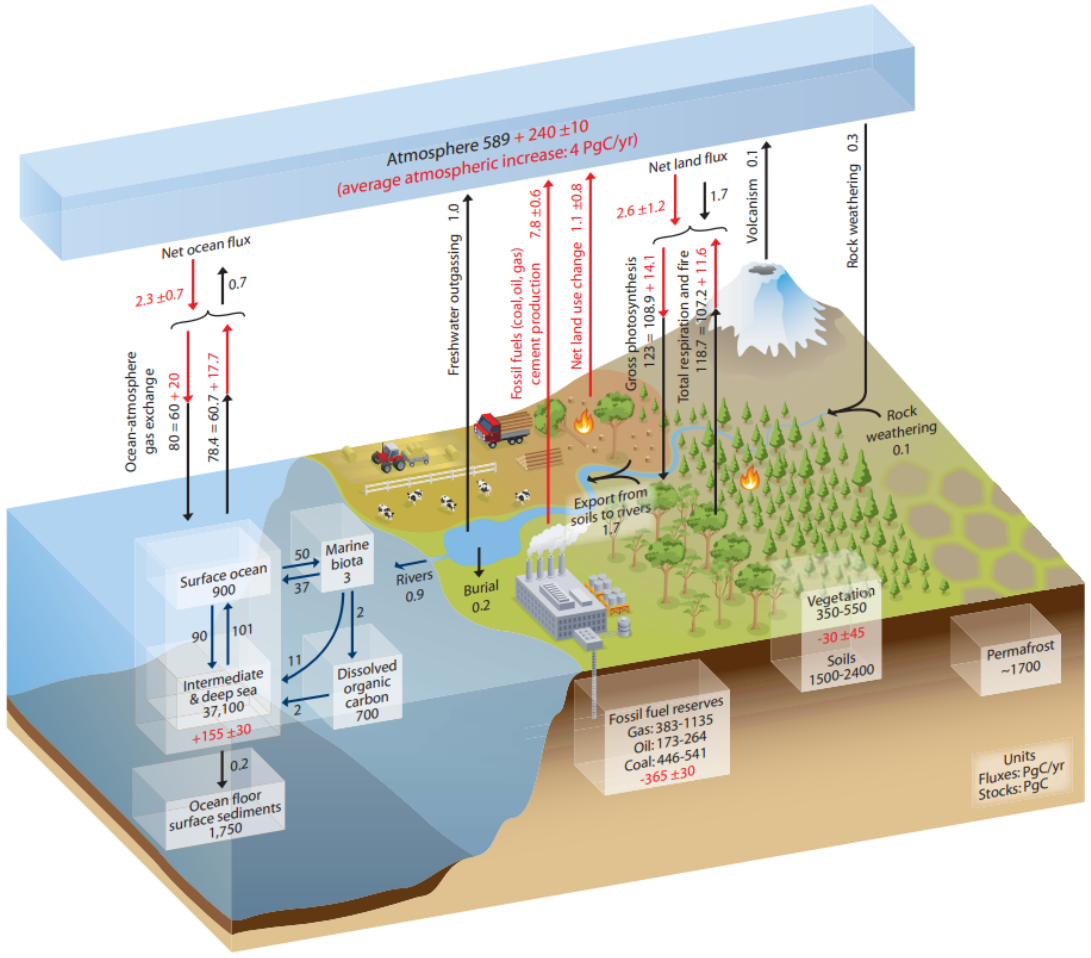

FIGURE 1.1: Diagram of the global carbon cycle, showing quantities of carbon stored in different components of the atmosphere, ocean, and terrestrial Earth, as well as the annual fluxes between them. Taken from Ciais et al. (2014).

surface and decays exponentially with depth. Consequently, in situations where nutrients are distributed uniformly across depth, the abundance of phytoplankton is highest near the surface and decreases with depth. Near-surface blooms are therefore a common phenomenon during spring at mid and high latitudes.

In highly stratified water columns where surface nutrients are depleted and not resupplied from below through mixing, the growth of phytoplankton near the surface is reduced. Nutrient concentrations increase with depth, in contrast to decreasing light intensity. This creates a scenario in which optimal conditions for phytoplankton growth occur well below the surface, resulting in a peak in phytoplankton biomass (C_{phyto}) (Beckmann and Hense, 2007; Barbieux et al., 2019; Martin et al., 2010). However, this does not always align with the peak in chlorophyll, as the ratio the mass of between cellular chlorophyll and phytoplankton biomass (Chl:C_{phyto}) typically increases under low light conditions – a physiological adaptation called photoacclimation. This was first proposed by Steele (1964), who identified the potential for the chlorophyll peak to be located deeper than the peak in C_{phyto}.

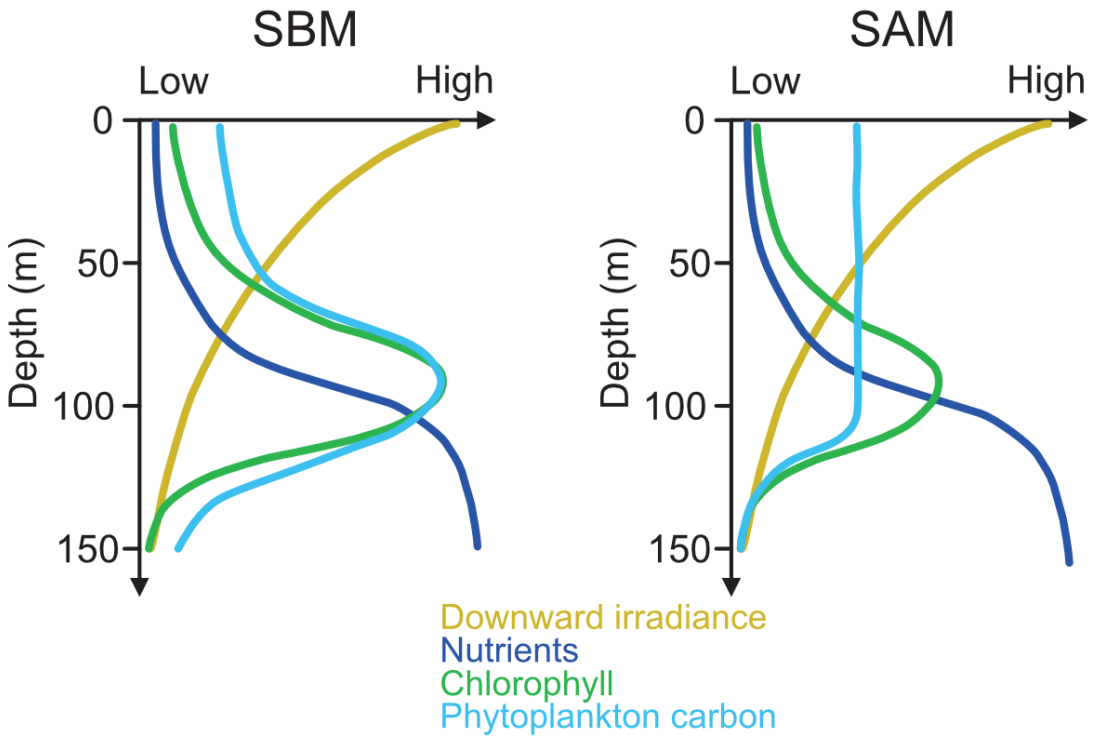

FIGURE 1.2: Schematic diagrams showing the typical light and nutrient profiles that give rise to subsurface chlorophyll maxima (SCMs), along with the characteristic bio-optical profiles of the two general types of SCM: subsurface biomass maxima (SBMs; left) and subsurface photoacclimation maxima (SAMs; right).

Consequently, it is now common to distinguish between subsurface biomass maxima (SBMs) and subsurface acclimation maxima (SAMs), under the umbrella of subsurface chlorophyll maxima (SCMs) (Figure 1.2). It is well established that the chlorophyll concentration at the SCM decreases with the depth of the peak (Herbland and Voituriez, 1979; Uitz et al., 2006; Cornec et al., 2021a). SCMs in high latitudes have been shown to form due to simultaneous increases in biomass and photoacclimation (Su et al., 2021).

Beyond surface blooms and SCMs, there is a wide variety of chlorophyll profile shapes depending on environmental conditions. For example, profiles with two distinct peaks have been observed (Muñoz-Anderson et al., 2015), or without a peak (that is, showing low chlorophyll throughout an entire profile) which occurs when a limiting factor is completely absent for a sustained period (Cornec et al., 2021a). The thickness of the SCM peak also varies, from < 5 m (Cowles et al., 1998; Dekshenieks et al., 2001; Durham and Stocker, 2012) to tens of metres (Uitz et al., 2006; Xu et al., 2022b).

1.2.2 Models of phytoplankton growth

For several decades, models of phytoplankton growth throughout the water column have been proposed in order to understand the dynamics between resources,

phytoplankton, and their consumers. Riley (1949) provided the first explanation of the interactions between nutrients, phytoplankton, and zooplankton in a vertically dynamic system. Their key improvement to previous work was the inclusion of turbulence that allowed nutrients and phytoplankton to be transported vertically over small distances. Steele and Yentsch (1960) added a variable sinking rate that decreased with depth, which allowed SCMs to form deeper at the base of the euphotic zone. Steele (1964) was the first to consider including a variable Chl:C_{phyto} ratio, allowing the distinction between the accumulation of phytoplankton biomass and the photoacclimation of phytoplankton. Further modelling approaches by Kiefer et al. (1976) and Fennel and Boss (2003) provided evidence that physiological adaptations contributed to the formation of SCMs. Whilst stratification is recognised as a fundamental prerequisite for the formation of SCMs (Cullen, 2015), and supported by theoretical modelling (Beckmann and Hense, 2007), it has also been shown that increased stratification can also lead to more unpredictable SCM behaviour (Huisman et al., 2006). The impacts of sinking rates (Hodges and Rudnick, 2004; Li and Hansell, 2016) and mesoscale ocean physics (Varela et al., 1992; Li and Hansell, 2016) have also been studied through models.

1.2.3 Chlorophyll profiles and environmental conditions

Many studies using observations, modelling, and theory have identified relationships between chlorophyll profiles and environmental conditions. I will briefly summarise the key findings from the literature.

Light Light is required for photosynthesis to occur, but the relationship between light and chlorophyll profiles still contains some uncertainty. Photosynthetically available radiation is the total intensity of light that photosynthetic plankton can use for photosynthesis, and typically encompasses wavelengths between approximately 400 nm and 700 nm. Surface irradiance has been used to infer the structure of subsurface chlorophyll (Gong et al., 2015; Joy-Warren et al., 2019). The presence of clouds can reduce surface irradiance and promote photoacclimation in phytoplankton (Begouen Demeaux et al., 2025). Studies of SCMs typically compare the profile shape with the euphotic depth (z_{eu}), typically defined as the depth at which the downwelling irradiance reaches a fraction of its surface intensity, often 1% (Ryther, 1956), although other values are also used (Wu et al., 2021). Several studies have found a positive correlation between z_{eu} and the depth of the SCM (z_{SCM}) (Xu et al., 2022b; Garg et al., 2024; Miyares et al., 2024). According to Cornec et al. (2021a), a minimum light level within the mixed layer was a necessary but insufficient condition for the formation of SCMs. This agrees with Mignot et al. (2014) and Xing et al. (2023) who identified a connection between vertical chlorophyll structure and isolumes.

Nitrate Nitrogen is a limiting factor for primary production in most of the global ocean, especially subtropical gyres (Browning and Moore, 2023), so it is not surprising that the nitracline depth (z_{SCM}), which describes the rapid increase in nitrate concentration from low concentrations typically found near the surface, is an important indicator of subsurface chlorophyll structure. Cullen (2015) suggests that the maximum phytoplankton biomass occurs at the nutricline. Herbland and Voituriez (1979) identified a strong positive relationship between the $z_{ncline}$and z_{SCM} in stratified tropical waters and a strong negative relationship between z_{ncline} and the magnitude of the subsurface peak (Chl_{SCM}). In some studies, z_{SCM} is coupled with z_{ncline} (Herbland and Voituriez, 1979; Cullen and Eppley, 1981), whereas in others they were correlated but did not occur at the same depth (Miyares et al., 2024; Garg et al., 2024). Gong et al. (2017) found that, analytically, the z_{SCM} should be shallower than z_{ncline} because phytoplankton in the SCM function as a barrier to the upward flux of nutrients. The steepness of the nitracline gradient indicates the magnitude of the upward nitrate flux, with steeper gradients shown to promote more intense and thinner SCMs (Gong et al., 2015, 2017).

Other nutrients Nutrients other than nitrogen can be (co-) limiting for phytoplankton growth (Moore et al., 2013), including iron (Hawco et al., 2021), silicate (Egge and Aksnes, 1992), phosphate (Lin et al., 2016) and manganese (Hawco et al., 2022). The case of iron as a limiting factor is well documented in the Southern Ocean (Moore et al., 2013). Terrestrial sources of iron lead to elevated chlorophyll concentrations at both the surface and subsurface (Baldry et al., 2020). Various sources, such as sea ice melt (Wang et al., 2014), islands (Graham et al., 2015; Robinson et al., 2016) and wildfires (Tang et al., 2021; Weis et al., 2022), can release sufficient iron to trigger surface blooms of phytoplankton. Studies on the effects of other nutrients on SCMs are relatively rare in the literature. An example is that of Firdaus et al. (2024), who demonstrated that water columns in Indonesia containing an SCM were nitrogen-limited given the high ratio of phosphate and silicate to nitrate.

Temperature and density The mixing of oceans determines the availability of nutrients in the euphotic zone, and it is therefore unsurprising that the physical structure of the water column has been shown to affect vertical chlorophyll distribution. For example, Carranza et al. (2018) observed that sigmoid-shaped profiles of chlorophyll developed during Southern Ocean storms which were usually followed by the formation of SCMs several days later, coinciding with a decrease in wind-driven mixing and a shoaling of the MLD. Several regional studies have found that the depth of SCMs was governed by the depth of a particular isotherm (Jayaram et al., 2021) or isopycnal (Xu et al., 2022b). Zampollo et al. (2023) found that the bottom of the pycnocline was a good predictor of z_{SCM} in a local, coastal setting.

Strongly stratified water columns provide the ideal conditions for the formation of SCMs, provided that light is sufficient (Beckmann and Hense, 2007; Cullen, 2015; Cornec et al., 2021a; Bock et al., 2022). Consequently, Miyares et al. (2024) found that in low latitudes the z_{SCM} was almost always located at or below the MLD, and frequently much deeper.

Surface chlorophyll Morel and Berthon (1989) first analysed the subsurface chlorophyll distribution in relation to surface chlorophyll concentration from a global dataset. Their work was built upon by Uitz et al. (2006) who created typical chlorophyll profile shapes based on their surface concentration in the subtropics. They found that lower surface concentrations result in deeper SCM with lower peak concentrations and thicker peaks. This provided significant evidence that surface chlorophyll could be used as a predictor of subsurface profile shape and characteristics of SCMs (Mignot et al., 2011). Quartly et al. (2023) extended this research to interannual timescales, removing seasonal variability and identifying that higher surface concentrations produce shallower and less intense SCMs. Xu et al. (2022b) and Miyares et al. (2024) both found a power law connecting surface chlorophyll and z_{SCM}.

1.2.4 Community composition and the wider ecosystem

There is a great diversity of phytoplankton species across the global ocean, which allows them to occupy a range of ecological niches within the epipelagic where they form the foundation of the marine ecosystem. Their size varies over several orders of magnitude, with the smallest species having a high surface area-to-volume ratio, and a high Chl:C_{phyto} ratio (Falkowski and Kiefer, 1985; Cloern et al., 1995), allowing them to grow at greater depths. Consequently, differences in species composition can be linked with the depth of SCMs (Uitz et al., 2006; Latasa et al., 2016, 2017; Sato et al., 2022; Miyares et al., 2024) meaning that similar communities are observed on large spatial scales (Bouman et al., 2006; Brewin et al., 2010; Ward et al., 2014). Latasa et al. (2023) found evidence that resource partitioning within SCMs is influenced by species following specific isolumes, supporting the findings of Sato et al. (2022), who observed consistent vertical ordering of phytoplankton groups across ocean basins. Nutrient limitation varies regionally (Moore et al., 2013; Arteaga et al., 2014) which can also influence species composition (Elizondo et al., 2021). For example, diatoms dominate the Southern Ocean community because of the availability of silicate, which they use to build hard shells and is much more limiting in other regions. On a smaller scale, the variable availability of nutrients within an SCM can lead to several species contributing to the overall peak in chlorophyll (Latasa et al., 2016). The relative availability of light and nutrients can lead to changes in the dominant phytoplankton species (Brewin et al., 2022; Xing et al., 2023). There is evidence to suggest that

productivity is highest in communities with moderate diversity, in which a variety of ecological niches are occupied (Vallina et al., 2014).

Zooplankton play an important role in determining phytoplankton abundance (Steinberg and Landry, 2017). Studies have found that grazing can produce more pronounced SCMs by reducing phytoplankton biomass near the surface (Pilati and Wurtsbaugh, 2003; Moeller et al., 2019). This suggests that SCMs can be partly controlled by top-down processes, in addition to bottom-up mechanisms, and the interaction between these two factors has been studied (Ward et al., 2014; Rodríguez-Gálvez et al., 2023).

1.3 Monitoring subsurface chlorophyll on a global scale

Subsurface chlorophyll concentration was first measured using discrete bottle samples at a small number of depths (Cullen, 2015). From even quite limited observations, some theoretical models were developed that provided a possible explanation for subsurface chlorophyll peaks (Riley, 1949; Steele and Yentsch, 1960). Subsequently, methods were developed to continuously measure chlorophyll throughout the water column (Lorenzen, 1966; Strickland, 1968) using in vivo flow-through fluorometry. Ship-based observations provided the majority of subsurface chlorophyll concentration measurements for several decades, although these were restricted in quantity and spatio-temporal distribution for logistical reasons (Smith et al., 2019). Consequently, observations were very sparse in space and time, especially in remote regions (such as the Southern Ocean) or during unfavourable conditions (for example, during winter or under sea ice).

In contrast to ship-based observations, satellite measurements of ocean colour, from which surface chlorophyll can be estimated, sacrificed the ability to measure below the ocean surface to vastly increase the spatio-temporal coverage of observations. Since the deployment of the SeaWIFS (Sea-Viewing Wide Field-of-View Sensor) satellite in 1997, there has been a continuous record of satellite-derived ocean colour observations. This can provide chlorophyll estimates to a high resolution (1-10 km) over much of the ocean, as well as a variety of other relevant biogeochemical variables (Brewin et al., 2021). The polar regions are not sampled during the winter because they receive too little sunlight. Furthermore, some locations that are often covered by cloud or sea ice have reduced temporal coverage (Tan et al., 2020). Estimation of subsurface chlorophyll concentration from surface measurements is not trivial (see Section 1.2.3), as SCMs are not directly detectable by satellite.

Technological advances have led to the development of various underwater instruments capable of measuring subsurface chlorophyll without the need for research vessels, for example. Biogeochemical-Argo (BGC-Argo) floats are an

autonomous platform that measures physical and biogeochemical properties in the upper 2000 m of the water column, typically once every ten days. The floats form a global array with deployments in every major ocean basin. The first BGC-Argo floats were deployed in the late 2000s, although the first to be equipped with bio-optical sensors were deployed in 2010. BGC-Argo floats are an extension of the so-called core Argo floats (which measure profiles of temperature and salinity) and have been fitted with a selection of sensors capable of measuring dissolved oxygen, chlorophyll fluorescence, backscatter by particles (b_{bp}, a proxy for particulate organic carbon), pH, nitrate concentration, and downwelling irradiance (Claustre et al., 2020). As of May 2025, there are 539 operational BGC-Argo floats equipped with bio-optical sensors. They are often deployed in batches under projects funded by different countries and organisations. For example, the Southern Ocean Carbon and Climate Observing and Modelling (SOCCOM) project (Sarmiento et al., 2023) deployed a large number of floats in the Southern Ocean. Once deployed from a research ship, they require no further physical assistance from scientists and therefore can be deployed in regions that experience harsh conditions at other times of the year, even in the presence of sea ice. The floats are semi-Lagrangian, meaning that they approximately follow the horizontal flow of water. Consequently, floats can end up sampling areas far from their deployment location. There are targets to reach an array of 1000 fully equipped BGC-Argo floats worldwide (Owens et al., 2022; Thierry et al., 2025). Some of the more technical aspects of BGC-Argo floats and their sensors are described in Section 2.1.

Other observing platforms include gliders (Testor et al., 2019; Carvalho et al., 2020), autonomous underwater vehicles, and remotely operated vehicles (Whitt et al., 2020), which can be programmed to perform high-resolution monitoring over a particular route. In particular, these platforms are beneficial for identifying sub-mesoscale and (sub-) diurnal variability. However, these are deployed on a smaller scale than the BGC-Argo array, which provides the largest global observing system for subsurface chlorophyll. Ship-based observations of phytoplankton abundance and species composition are still vital to fully understand the effects of environmental conditions, which cannot be fully observed through other sources (Garg et al., 2024; Miyares et al., 2024). Figure 1.3 demonstrates how the combination of different observing systems is beneficial for a complete understanding of phytoplankton dynamics and distribution across the three-dimensional global ocean and over a range of temporal scales. In summary, the quantity of subsurface chlorophyll measurements has increased significantly in the past couple of decades, enabling the study of the SCMs on a global scale.

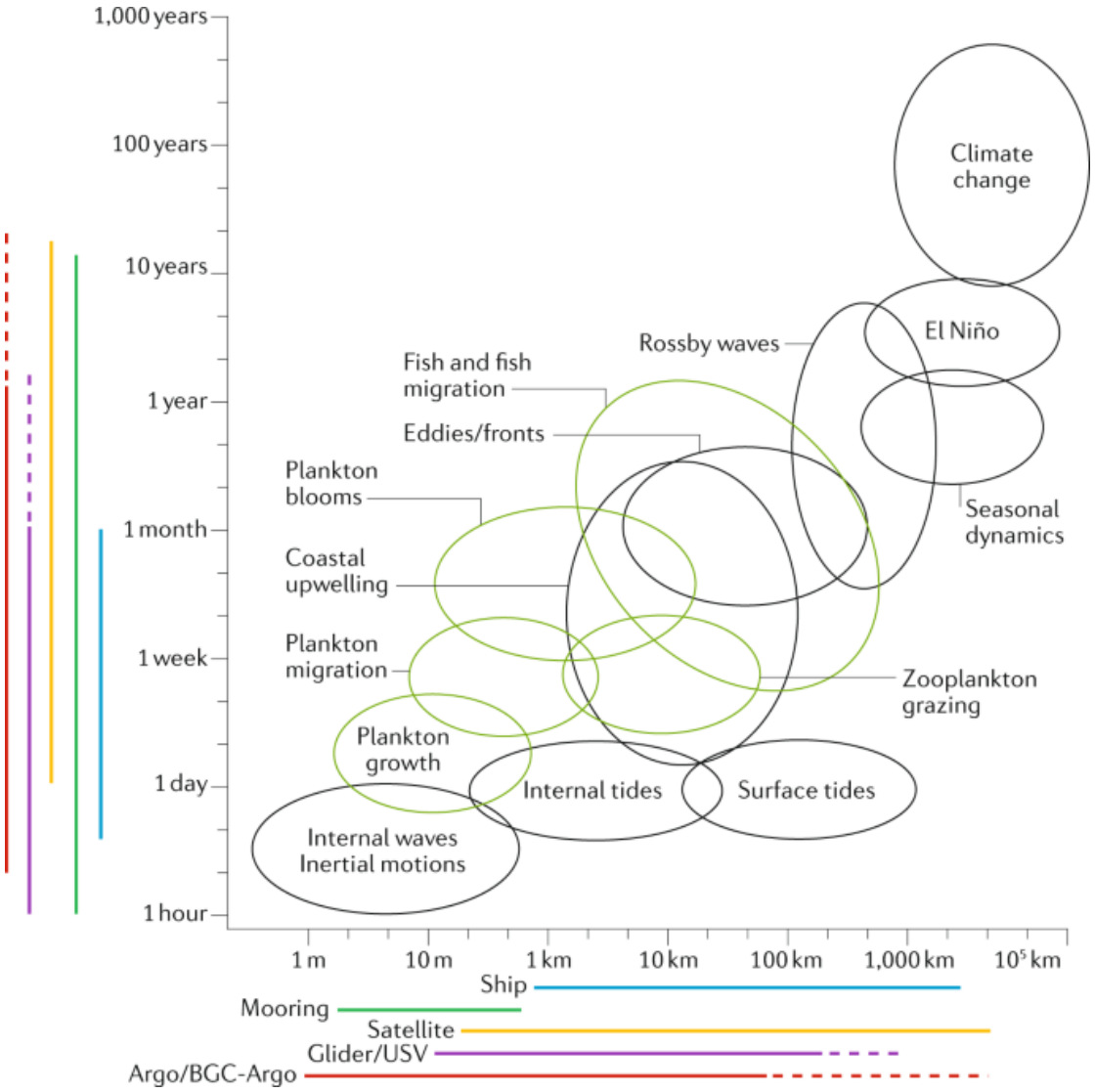

FIGURE 1.3: A diagram displaying the spatio-temporal scales of processes affecting biogeochemistry and the scales measured by different observing systems. Taken from Chai et al. (2020).

1.4 Spatio-temporal distribution of subsurface chlorophyll across the global ocean

1.4.1 Biogeographical regions

Similarly to terrestrial biomes, the global ocean can be partitioned into distinct biogeographical regions, each describing a characteristic marine ecosystem. The first such division was suggested by Longhurst et al. (1995) containing 56 regions with similar values for several variables including sea surface temperature, surface chlorophyll concentration and sea ice coverage. Alternative regions were suggested by Sarmiento et al. (2004) and Gurney et al. (2008), although none of these had dynamic boundaries. Reygondeau et al. (2013) developed a seasonally varying set of

regions, before Fay and McKinley (2014) defined biomes with complex and dynamic boundaries. Biogeographical regions have also been defined using phytoplankton community composition (Elizondo et al., 2021). All the aforementioned approaches used gridded data products to define a region for nearly all of the global ocean. In contrast, Bock et al. (2022) used data from the BGC-Argo array to identify six biome types based on their seasonal cycles of chlorophyll and b_{bp} profiles. Their approach extracted the main components of seasonal variation in profiles, although they were unable to define biome boundaries because of the sparse coverage of BGC-Argo floats. I will briefly describe the key differences in vertical chlorophyll distribution within each major biome, particularly the different mechanisms for SCM formation.

Subtropics Subtropical regions typically offer the ideal environmental conditions for the formation of SCMs, namely stratified water columns with year-round high irradiance (Beckmann and Hense, 2007). Several studies have shown that the oligotrophic subtropical gyres, which have some of the lowest surface chlorophyll concentrations (< 0.15 mg m^{-3}), generally have the deepest SCMs, which can be located at > 160 m depth (Uitz et al., 2006; Mignot et al., 2014; Cornec et al., 2021a; Yasunaka et al., 2021; Bock et al., 2022), although the depth of SCMs varies by several tens of metres seasonally due to changes in surface solar irradiance (Letelier et al., 2004). These regions cover a significant proportion of the ocean surface making the subtropical SCM a globally important ecosystem. These regions typically favour smaller phytoplankton (Latasa et al., 2017; Sato et al., 2022), which have a high Chl:C_{phyto} ratio, meaning that the peak in chlorophyll is deeper than the peak in phytoplankton biomass (Fennel and Boss, 2003; Cornec et al., 2021a; Masuda et al., 2021).

Tropics Locations near the equator tend to display an SCM between depths of 50 m and 80 m, typically around the nitracline, with little seasonal variation (Cornec et al., 2021a; Bock et al., 2022; Miyares et al., 2024). The shallower SCMs are driven by upwelling, bringing nutrients into the euphotic zone and supporting SCMs with higher biomass. The increase in biomass is a product of elevated light availability and high-nutrient flux conditions (Bock et al., 2022). Bock et al. (2022) also found that the nitracline was 29 m deeper than the MLD on average, indicating that SCMs in the equatorial regions are primarily supported by nutrient supply from below the mixed layer, rather than by mixing within it. Masuda et al. (2021) found that photoacclimation is still important in the equatorial regions despite shallower SCMs, finding similar Chl:C_{phyto} ratios, which peaked at a depth of around 120 m.

Mid latitudes The seasonal variability of water column characteristics increases with distance from the equator, resulting in SCMs forming only during summer in the

mid latitudes (Bock et al., 2022). Shoaling of the MLD increases stratification and reduces the supply of nutrients to sunlit surface waters, allowing the formation of SCMs, which tend to be shallower than in the subtropics (Cornec et al., 2021a; Bock et al., 2022). Before the formation of summer SCMs, a spring bloom typically occurs near the surface (Martinez et al., 2011; Chiswell et al., 2015), with high chlorophyll concentrations observed (> 1 mg m^{-3}). Sometimes, a smaller surface bloom occurs during autumn, as the MLD deepens, resupplying the surface with nutrients (Wihsgott et al., 2019). The Mediterranean Sea provides an interesting case study: the western end experiences seasonal SCMs, whereas the eastern more closely resembles subtropical conditions and has a year-round SCM (Lavigne et al., 2015; Barbieux et al., 2019). These seasonal variations in chlorophyll vertical distribution reflect a change in species composition (Bolaños et al., 2020).

Polar regions The polar oceans represent some of the most remote environments for studying phytoplankton and their role in the wider ecosystem. The deployment of BGC-Argo floats in these regions (including through the SOCCOM project) has helped the study of their phenology throughout the entire year, including during periods under sea ice. The extreme seasonal variation in light availability is reflected in the phenology of high-latitude phytoplankton populations. In spring, chlorophyll concentrations increase, even before the sea ice has melted (Hague and Vichi, 2021; Vives et al., 2024b). Once the ice has completely disappeared, strong surface blooms occur (Uchida et al., 2019; Kubryakova et al., 2025). Although they show very different physical conditions than the water columns traditionally considered ideal for SCM formation, recent research has established that they commonly form at high latitudes (Holm-Hansen and Hewes, 2004; Venables and Moore, 2010; Baldry et al., 2020, 2024; Kubryakova et al., 2025). The dominant phytoplankton species in the Southern Ocean are diatoms, whose buoyancy may aid the formation of SCMs (Baldry et al., 2020). Iron limitations in the Southern Ocean can also lead to SCMs (Baldry et al., 2020). In the Arctic, SCMs typically form at a depth of between 20-50 m due to low light levels (Martin et al., 2010; Arrigo et al., 2011) after the spring bloom (Ardyna et al., 2013). Unsurprisingly given the low light levels, photoacclimation plays a significant role in SCM formation in both polar regions (Baldry et al., 2020; Masuda et al., 2021).

1.4.2 Scales of variability

As outlined earlier, vertical chlorophyll structure varies by location and, in many cases, season. These differences tend to occur over ocean basin scales (Karl, 1999) due to global thermohaline circulation, stratification and the intensity of solar irradiance. However, within a biome, there is also significant variability on the mesoscale ($\sim$100 km), due to the presence and polarity of eddies (Huang and Xu, 2018; Cornec et al.,

2021b; Xu et al., 2022b; Strutton et al., 2023; Wang and Liu, 2024). These studies have established that cold-core (cyclonic) eddies bring nutrients closer to the surface, causing SCMs to rise in the water column. The opposite is true for warm-core (anticyclonic) eddies. Ocean fronts can also lead to small-scale variability both at the surface (Sokolov and Rintoul, 2007) and throughout the water column (Tripathy et al., 2015; McKee et al., 2023). Interannual surface chlorophyll variability has been shown to be dominated by small-scale processes (Keerthi et al., 2022; Prend et al., 2022), although it has not yet been shown that the same applies to chlorophyll profiles. On a vertical scale, SCMs can have very thin peaks (Cowles et al., 1998; Durham and Stocker, 2012), which can indicate the presence of precise conditions that favour a particular phytoplankton functional group. The vertical position of the SCM can change due to diel migration of phytoplankton in the water column, reflecting a behavioural adaptation to access light near the surface during the day and greater nutrient availability in deeper waters during the night (Cullen, 1985; Cullen and MacIntyre, 1998; Wirtz et al., 2022).

1.5 The usefulness of spatio-temporal modelling

A spatio-temporal dataset consists of observations that are each associated with a specific location and time. This type of data is common in environmental science, where variations over space and time are often of significant interest. There are several types of spatio-temporal datasets. Firstly, there are geostatistical datasets in which the variable of interest has been measured at specific point locations. Second, there are areal datasets, where each measurement is associated with a region. Third, there are point patterns in which the locations of measurements are the primary interest, rather than measurements at those locations. An example of this is the distribution of trees in a forest. In this thesis, I am only interested solely in geostatistical datasets, and this will be referred to as spatio-temporal modelling throughout this thesis, unless otherwise made clear.

There are a number of reasons why including spatio-temporal information within an analysis is desirable. For example, we can account for the spatio-temporal dependencies between observations. Tobler's first law of geography states that observations close to each other are more similar than those far apart (Miller, 2004), and this is a characteristic that should be integrated into a statistical model when measurements are associated with locations and times. Consequently, spatio-temporal models pool information over space and time through the inclusion of latent effects. Such latent effects allow for the partitioning of variance between that attributable to measured covariates and that arising from unmeasured covariates that may vary over space and time. This has the added benefit of better quantifying estimates and their corresponding uncertainties in covariate effects, since the presence of a

spatio-temporal latent effect can substantially affect covariate effects (Willems et al., 2022). By controlling for spatial variability, it becomes more feasible to accurately detect temporal patterns such as seasonality and long-term trends (Laurini, 2019; Hammond et al., 2020). A common objective of spatio-temporal modelling is to produce maps of a variable of interest, and possibly to forecast, whilst incorporating uncertainty quantification. Interpolation through the use of a latent spatio-temporal effect aids in this by accounting for autocorrelation between observations. One of the disadvantages of traditional spatio-temporal modelling methods is often the computation time, given the necessity of calculating a distance matrix between all observations. Several approaches have been suggested to alleviate this burden by modelling covariance between locations using an approximation, which can be solved using computationally efficient methods (Lindgren et al., 2011; Anderson et al., 2022; Pereira et al., 2022). These approximations significantly reduce computational time and memory requirements, making it feasible to apply spatio-temporal models to large datasets.

The literature on spatio-temporal modelling is extensive and contains a wide range of techniques, especially in the development of statistical software for such analyses. Excellent overviews of the topic are provided by Cressie and Wikle (2011) and Sahu (2022). A range of spatio-temporal models have been developed for the analysis of oceanography datasets (Du et al., 2018; Hammond et al., 2020; Hildeman et al., 2021). Stein (2020) identified statistical challenges associated with analysing the Argo float data. Firstly, from a spatial modelling perspective, it is quite a large dataset. This has ramifications in estimating the covariance between all observations based on the distance between them, which is a computationally expensive process. Second, the fact that the floats drift over time such that no specific location is ever sampled twice means that there are challenges in estimating temporal autocorrelation. Nevertheless, several approaches for spatio-temporal modelling Argo float data have been proposed. Sahu and Challenor (2008) suggested a hierarchical modelling approach to account for seasonality. Roemmich and Gilson (2009) produced the first global temperature and salinity maps from the core Argo float data by first estimating a mean field and then performing kriging. Kuusela and Stein (2018) extended this by performing local regression, in which smaller 'local' models are fitted to subsets of the data through a moving window to improve the characterisation of anomalies.

Geostatistical analyses are mainly focussed on datasets whose spatial coordinates lie on a two-dimensional surface (for example, a flat plane or the Earth's surface). This means that for many oceanographic studies, such as those on sea surface temperature, many spatio-temporal modelling approaches are applicable. However, the Argo dataset is inherently three-dimensional in space, since each measurement has an associated longitude, latitude, and depth. These techniques are unsuitable for analysing spatio-temporal data across multiple depths simultaneously, as traditional

modelling approaches cannot capture this structure and require depth-by-depth modelling, as was done by Sahu and Challenor (2008), Roemmich and Gilson (2009) and Kuusela and Stein (2018). Including depth as a third spatial dimension is not trivial, since the length scales of variability in the vertical are several orders of magnitude smaller than in the horizontal. Therefore, alternative methods are required to fully utilise the dependency within and between profiles.

Machine learning (ML) algorithms provide an alternative to true parametric spatio-temporal models to construct maps from observations due to their ability to extract complex relationships between several variables and make reasonable predictions for complex systems. In oceanography, it has become common to use ML for three-dimensional interpolation between sparse observations of a variable of interest using a more widely measured variable, often obtained by satellites, as a predictor (Sauzède et al., 2016; Chen et al., 2021; Renosh et al., 2023). A limitation with ML approaches is that weaker inferences can be made from a statistical perspective since dependence structures between locations and times are not explicitly described by the model framework. This is where spatio-temporal models are advantageous. While some results from spatio-temporal modelling might not immediately appear as preferable to ML, the depth of inference – particularly in partitioning variability between covariate effects and latent spatio-temporal variability – offers sufficient justification for these methods.

1.6 The usefulness of functional data analysis

Given the dependence between measurements in a single profile, it is no surprise that previous studies on chlorophyll profile shape have fitted profiles to families of mathematical curves (Lewis et al., 1983; Ardyna et al., 2013; Gong et al., 2015; Muñoz-Anderson et al., 2015; Carranza et al., 2018; Xu et al., 2022b). In doing so, entire profiles can be described using only a small number of parameters representing characteristics such as SCM depth or thickness. The main reason for this is to facilitate the interpretation of environmental changes that affect these parameters and, consequently, the profile shape. Choices for curves include Gaussians (Gong et al., 2015; Xu et al., 2022b), sigmoids (Carranza et al., 2018) or even combinations of multiple curves (Muñoz-Anderson et al., 2015; Carranza et al., 2018; Brewin et al., 2022; Sato et al., 2022). Theoretical results have suggested that these profile shapes should be observed; however, in practice, there are cases where profiles do not fit one of the assumed shapes. Additionally, the process of fitting curves to profiles can be time-consuming for large datasets, or for complicated curves, since they typically contain more parameters.

Functional data analysis (FDA) provides an alternative and appropriate set of techniques for ocean profile data. This branch of statistics comprises methods for data analysis in which a variable of interest is a continuous function of another variable, which in practice is represented using basis functions. This contrasts with scalar-valued data, which consists of finite sets of values, and are the building blocks of most familiar statistical methods. The difference between these two will be made clear where necessary throughout this thesis. FDA gained attention in the 1990s and a general guide to the topic is given by Ramsay and Silverman (2005), who describe methods for regression models for functional data. It is worth mentioning that functional regression can involve predicting functional data using scalar covariates, or vice versa, or predicting functional data with functional covariates (Ramsay and Dalzell, 1991). Mathematical details of functional regression are described in Section 2.2.2. The FDA literature has expanded significantly in the past two decades, with developments in regression (Greven and Scheipl, 2017), clustering (Zhang and Parnell, 2023) and, geostatistics (Mateu and Giraldo, 2021), who present a detailed selection of geostatistical models for functional data, including spatio-temporal kriging (optimal interpolation), analysis of variance and clustering. Recently, Urbano-Leon et al. (2023) developed a method for defining variance and correlation for functional data as a scalar value, enabling correlation analyses analogous to those for scalar datasets. Since this approach has not yet been applied to datasets of oceanographic profiles, I explore its use with Argo profiles in Chapter 5.

FDA provides a natural framework for oceanographic profiles, which are a function of depth. There are several immediate advantages to using FDA for Argo data. Firstly, the dependency between discrete measurements on a single profile is included within the framework by assuming a continuous function passes through all measurements. Moreover, Argo profiles often sample irregularly across a range of depths, and by describing profiles as functional data objects, this irregularity can be neglected in later analysis. Profiles as functional data objects allow for the extraction of the gradient and integrals of variables over depths, for example, the heat content of the ocean (Yarger et al., 2022). Despite these benefits, there have been relatively few applications of FDA in oceanography. To my knowledge the first was Assunção et al. (2020), who used it to study thermohaline structure from ship-based observations. Yarger et al. (2022) developed a spatio-temporal model designed specifically for Argo data. This approach involved fitting local mean and covariance functions to predict temperature and salinity profiles at unsampled locations. Korte-Stapff et al. (2022) proposed a method for interpolating between the sparse BGC-Argo using more widely sampled core Argo profiles throughout the Southern Ocean. Le Ster et al. (2023) used linear functional models to assess chlorophyll profile variability measured from biologging tags attached to marine mammals. Kande et al. (2024) also explored the use of FDA for spatial statistics in a marine ecology context.

1.7 Aims

My aims for this thesis are two-fold. Firstly, I seek to assess the spatio-temporal variability of chlorophyll profiles from BGC-Argo profiles, with a particular focus on those displaying SCMs. I aim to quantify the extent to which the presence and structure of SCMs are determined by environmental drivers such as light, nutrients, and the MLD, and physiological adaptations by phytoplankton. This will hopefully contribute to a greater understanding of the mechanisms of SCM formation and maintenance within ocean ecosystems. The second aim is to apply a selection of statistical approaches to data from BGC-Argo floats from spatio-temporal modelling and FDA, assessing their usefulness for future applications. Each of Chapters 3-5 addresses a slightly different biogeochemical question and demonstrates a different statistical approach. Collectively, the thesis demonstrates both the ecological insights gained from BGC-Argo float data and the methodological advances that assist their interpretation.

1.8 Thesis structure

In this thesis, I demonstrate techniques from spatio-temporal modelling and FDA to investigate spatio-temporal variability in subsurface chlorophyll using data from BGC-Argo floats. I will first introduce the Argo dataset in Chapter 2, as well as a more technical background to the statistical approaches. In Chapter 3, I investigate drivers of SCM characteristics on a global scale using a spatio-temporal model. In Chapter 4, I employ functional regression models to assess how environmental conditions affect SCMs at low latitudes. In Chapter 5, I apply a novel method for defining variance and correlation for functional data to oceanographic profiling data, using chlorophyll and temperature BGC-Argo profiles across the global ocean as an example. Finally, in Chapter 6, I summarise the collective findings of the entire thesis and suggests directions for future research.

Chapters 3-5 are intended to be submitted to scientific journals, and there may be some repetition across chapters, particularly when reviewing previous work and describing the aims for each chapter. All references have been collated at the end of the thesis.

Chapter 2

Materials and Methods

2.1 Argo float data

Argo floats are autonomous platforms equipped with a range of sensors to measure the properties of the top 2000m of the marine water column. In addition to measuring the temperature, salinity, and pressure of the water column, biogeochemical-Argo (BGC-Argo) floats carry a suite of specialised sensors for monitoring biogeochemical processes (Claustre et al., 2020; Chai et al., 2020). They control their depth by changing the volume of an oil-filled bladder and transmit measurements back to land via satellite using an Iridium antenna. Floats are programmed to not surface in the presence of sea ice (Wong and Riser, 2011); instead, they continue profiling beneath the ice and transmit data once they re-emerge. The positions of under-ice profiles are later estimated (Chamberlain et al., 2018). A diagram of a BGC-Argo float is shown in Figure 2.1.

Most floats complete one profile of the upper 2000 m of the water column, typically every 10 days for 5-7 years, before descending to a parking depth of 1000 m. They can be programmed to complete profiles at particular times, meaning that some floats have much higher sampling frequency than others and consequently have a shorter lifespan. The vertical sampling structure varies between floats and sensors, but typically deeper measurements are sparser compared to the upper ocean. A schematic showing a cycle of an Argo float is shown in Figure 2.2.

All measurements collected are given an initial quality control flag by the Argo data centre (Table 2.1). Measurements that have quality control flags of 3 or 4 after adjustment are not included in any analyses. Similarly, I only used “Delayed” mode data, which have been more thoroughly quality controlled, in contrast to “Real time” data. All Argo float data were downloaded in netCDF format prior to analysis using the R package **argoFloats**. These methods are described in the relevant chapters.

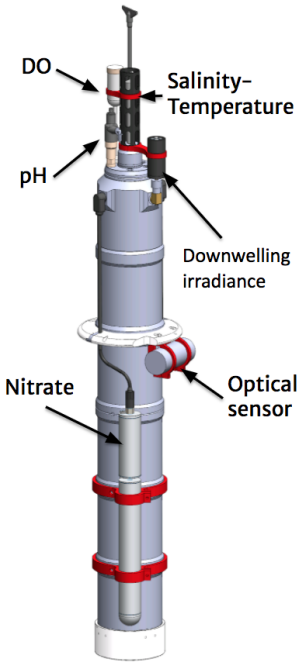

FIGURE 2.1: Diagram of a BGC-Argo float equipped with measuring temperature, salinity and the following six biogeochemical parameters: chlorophyll-a fluorescence, optical backscatter, pH, dissolved oxygen (DO), nitrate and downwelling irradiance. Figure taken from https://www.go-bgc.org/floats (last accessed 17/06/2025).

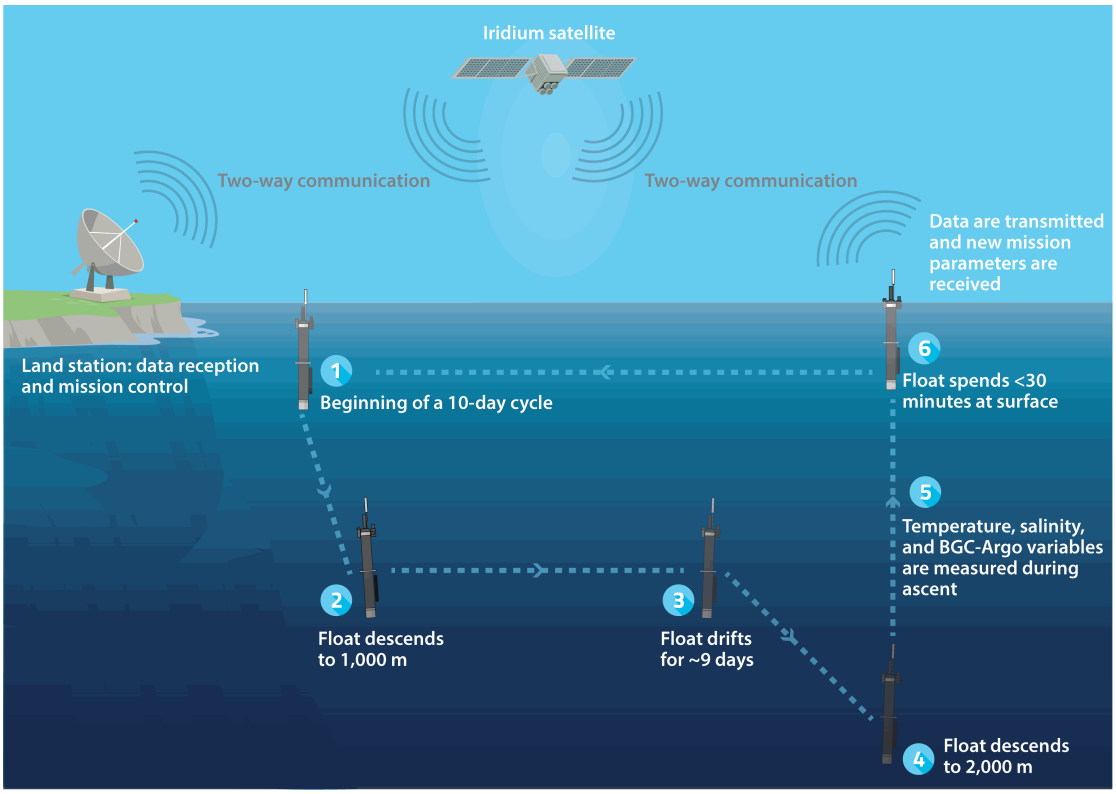

FIGURE 2.2: Schematic of a typical cycle of an Argo float. Taken from Claustre et al. (2020).

Flag	Meaning
0	No QC
1	Good
2	Probably good
3	Probably bad
4	Bad
5	Value changed
8	Interpolated value
9	Missing value

TABLE 2.1: The quality control flags for measurements from Argo floats. No measurements with a flag of 3 or 4 were included in any analyses.

2.1.1 Chlorophyll concentration

Some BGC-Argo floats are fitted with bio-optical sensors manufactured by SeaBird which can measure chlorophyll-a fluorescence. The sensor detects a change in the intensity of light at two different wavelengths. Specifically, when light is intercepted by chlorophyll a, a fraction of the photons absorbed at the blue end of the spectrum (470 nm) are re-emitted at the red end of the spectrum (695 nm), which carry less energy. The light emitted is the fluorescence of chlorophyll-a (F_{Chl}, [mole quanta m^{-3} s^{-1}]). The following equation is used to estimate chlorophyll-a fluorescence from the raw chlorophyll concentration $[\text{Chl}]_{\text{raw}}$

$$F_{\text{Chl}} = E[\text{Chl}]_{\text{raw}} a^* \Phi_f \tag{2.1}$$

where E is the excitation irradiance [mole quanta $\text{m}^{-2}\text{s}^{-1}$], a^* is the chlorophyll-a specific absorption coefficient [m^2 (mg Chl a)$^{-1}$] and Φ_f is the fluorescence yield [mole quanta emitted (mole quanta absorbed)$^{-1}$]. The recorded chlorophyll concentration (denoted as $[\text{Chl}]_{\text{rec}}$ with units [mg m^{-3}]) is then estimated by applying a float-specific conversion supplied by the manufacturer. The conversion takes the following form

$$[\text{Chl}]_{\text{rec}} = (F_{\text{Chl}} - \text{dark offset}) \times \text{scale factor}. \tag{2.2}$$

This method for estimating chlorophyll concentration by Argo data team allows for for several corrections (Schmechtig et al., 2023), given below:

1. In situ dark correction. This error refers to an offset to the measurement in the presence of zero chlorophyll. This correction is provided by the manufacturer but can change once the sensor is added to a float as well as during its active lifespan.

2. Non-photochemical quenching (NPQ). NPQ is a physiological response of phytoplankton to high light levels, resulting in lower fluorescence measurements than expected given the chlorophyll concentration. This is

corrected by assuming that any decrease in fluorescence below the maximum observed in the mixed layer is due to NPQ (Xing et al., 2012). For practical reasons, this is the method chosen for the BGC-Argo dataset, although more recent methods have been suggested (Roesler et al., 2017; Xing et al., 2017, 2018).

3. Scale factor. This describes the rescaling from fluorescence units [mole quanta m^{-3} s^{-1}] to concentration units [mg m^{-3}] and is supplied by the manufacturer. However, several factors can affect this including (but not limited to) location, depth, time of day and species composition.

From here on when I refer to chlorophyll concentration I mean the recorded chlorophyll concentration and I drop the square brackets around chlorophyll concentration for consistency. It should be noted that a recent study showed that the relationship between F_{Chl} and chlorophyll concentration varies between different phytoplankton communities (Petit et al., 2022). Moreover, detected fluorescence can originate from sources other than chlorophyll in the deep sea, such as in low oxygen conditions (Xing et al., 2017; Wojtasiewicz et al., 2020). However, this is unlikely to be a significant factor in my analyses since I focus on the top 250 m of the ocean. The chlorophyll fluorescence sensor on BGC-Argo floats does not show significant drift over time (Claustre et al., 2020).

These adjusted chlorophyll concentration values (labelled as CHLOROPHYLL_A_ADJUSTED in the Argo dataset) serve as the starting point for my analyses. Additional quality control procedures are implemented as needed, depending on the specific analysis, and are therefore described in the relevant chapters.

2.1.2 Particle backscatter

2.1.2.1 Measurement

The backscatter coefficient at wavelength λ, $b_b(\lambda)$, is calculated through several steps. First, the volume scattering function (VSF), a function of backscattering angle θ and wavelength λ, is estimated for a specific angle (often 117°) (Boss and Pegau, 2001). Then $b_{bp}(\lambda)$ is calculated by integrating the VSF between $\theta = 90°$ and $\theta = 180°$, including a scaling by a factor of 1.1 (Boss and Pegau, 2001). Finally, the total backscatter b_b is partitioned into the backscatter from seawater and the backscatter by particles, which can be interpreted as the combined concentration of detritus and phytoplankton in the open ocean. This is essential since seawater can cause up to 80% of backscatter in very clear water (Morel and Gentili, 1991). The unit of b_{bp} is m^{-1} and the measurement represents the probability per unit length that light is backscattered

by particles. On BGC-Argo floats, b_{bp} is measured at λ =700 nm (sometimes at other wavelengths as well) using the SeaBird WETLabs sensor.

2.1.2.2 Quality control

The assumption is that b_{bp} scales linearly with particle concentration, but in reality it also varies with particle size and composition. The real-time b_{bp} data are subject to the quality control tests developed by Dall'Olmo et al. (2023). The following tests are applied to the raw b_{bp} profiles.

1. Missing data test. This checks whether there is at least one measurement across a range of depth bins in the top 1000 m of the water column.
2. High deep value test. A particularly high b_{bp} measurement can indicate that the sensor has malfunctioned or that it has been subjected to biofouling. This test only affects values deeper than 700 m, which is deeper than any used in this work.
3. Negative b_{bp} test. Negative values can indicate that the sensor is out of the water and has malfunctioned.
4. Noisy profile test. A profile with more than 10% of measurements far from the median value (residuals are greater than 0.0005 m^{-1}), is given a QC flag of 3 ("Probably bad").
5. Parking hook test. Particles can accumulate on the sensor while it is at its parking depth of 1000 m. This test identifies when these particles detach from the sensor as it ascends in the water column. If this is detected, then these measurements receive a quality control flag of 4 ("Bad").

Full details of these tests can be found in the backscatter manual supplied by the BGC-Argo data team (Dall'olmo et al., 2023). The adjusted b_{bp} measurements (labelled BBP700_ADJUSTED in the Argo dataset) are combined with the adjusted chlorophyll measurements for my analysis in Chapter 4.

2.1.3 Temperature, salinity and depth

Measurements of conductivity, temperature, and depth (CTD) are collected by all Argo floats using SeaBird CTD sensors (such as the SBE-41 and other versions). These sensors are specifically designed for autonomous use in the ocean, providing stable and accurate measurements over the multi-year lifespans of the floats. The CTD measurements go through a rigorous quality control process that includes real-time

automated tests such as spike detection and checks for sensor drift followed by delayed-mode adjustments based on sensor drift, comparisons with ship-based reference profiles, and historical climatologies. These procedures ensure that the data used for oceanographic analyses are of the highest possible accuracy. Given the technical nature and complexity of these quality control procedures, I refer the reader to the Argo CTD data manual, which gives complete details of the quality control procedures (Wong et al., 2025). In this thesis, the CTD-derived variables are used within each of Chapters 3-5 and I will describe any further treatment of them in the relevant chapters.

2.2 Overview of statistical methods

2.2.1 Spatio-temporal models

2.2.1.1 Summary of Bayesian statistics

Bayesian statistics is a branch of statistics in which the parameters that govern random processes are not assumed to have fixed 'true' values. Instead, their values are described probabilistically. By combining prior beliefs with evidence (in the form of data), we update our understanding to form posterior beliefs. Uncertainty about parameter values is expressed through probability distributions, allowing us to represent not a single estimate but a full range of plausible values along with their associated uncertainties.

Suppose that we have some prior knowledge of a process and describe our uncertainty of a parameter with a given distribution. Suppose we then collect some data and calculate the likelihood function, a function of the probability of seeing the data given parameter values. The procedure to update our prior beliefs is given by the following relationship

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}. \tag{2.3}$$

The proportional symbol in Equation 2.3 is included to simplify the expression by removing a denominator of the probability of seeing the data, which is a complicated calculation and is unnecessary for parameter estimation. The shape of the posterior distribution is important, as it explains our updated uncertainty. If required, we can identify a summary statistic describing the posterior distribution (often the mean or median) as a single value updated parameter estimate. The location of the maximum is unaffected by the absent constant of proportionality, which means this denominator can be ignored in practice.

This perspective naturally allows for hierarchical model structures, whereby the parameters of some probability distribution are themselves described by probability

distributions. In this case, a prior distribution is provided for each parameter that forms the base of the hierarchy. For more extensive guides to Bayesian statistics, readers are referred to Bolstad and Curran (2016) and Sahu (2022).

2.2.1.2 Parameter estimation

In some Bayesian models, the choice of prior distribution allows for an analytical solution to the posterior distribution. However, in practice this is not often the case (often for complicated hierarchical models), and instead the posterior distribution is difficult to identify and needs to be estimated using a numerical approximation called Monte Carlo sampling, which relies on taking many random samples to estimate values. A common example of this is Markov chain Monte Carlo (MCMC) sampling, in which samples are generated sequentially, slowly exploring the parameter space. Depending on the problem, this can be a computationally expensive and time-consuming task to ensure that the (possibly high-dimensional) parameter space is well sampled, i.e., that parameter estimates have converged.

2.2.1.3 Integrated Nested Laplace Approximation

In recent years, approximate Bayesian inference has become increasingly popular. An example of this is the integrated nested Laplace approximation (INLA), as described by Rue et al. (2009), which is designed for the class of statistical models called latent Gaussian models (LGMs). LGMs are hierarchical models comprising observations, a latent Gaussian field, and hyperparameters. The observations $\boldsymbol{y}$ are assumed to be generated by a process belonging to the exponential family of distributions with hyperparameters $\boldsymbol{\theta}_1$, and the mean μ is connected to the linear predictor through a link function. A vector $\boldsymbol{x}$ contains the linear predictor and the covariate coefficients. The latent effect x is distributed according to a Gaussian Markov random field (GMRF), which is dependent on some hyperparameters $\boldsymbol{\theta}_2$. The set of all hyperparameters governing both the mean and the GMRF is denoted by $\boldsymbol{\theta}$. The data layer is given by

$$\boldsymbol{y}|\boldsymbol{x}, \theta \sim \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \tag{2.4}$$

where π denotes a generic probability distribution. The latent variables x represent a GMRF given by

$$\boldsymbol{x}|\boldsymbol{\theta} \sim N(\mu(\boldsymbol{\theta}), \boldsymbol{Q}(\boldsymbol{\theta})^{-1}) \tag{2.5}$$

where $\boldsymbol{Q}(\theta)$ is a precision matrix, defined as the inverse of a covariance matrix. Finally, there are prior distributions for the hyperparameters.

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \tag{2.6}$$

The objective is to estimate the values of the parameters $\boldsymbol{\theta}$, which control the mean μ and the GMRF. In contrast to MCMC sampling approaches, INLA is deterministic. INLA involves several steps, which I will briefly explain.

1. First, the marginal posterior of the hyperparameters is approximated $\pi(\theta|\boldsymbol{y}) \approx \tilde{\pi}(\theta|\boldsymbol{y})$ using the Laplace approximation, which involves estimating a Gaussian centred on the mode of the posterior distribution.
2. The latent Gaussian field is approximated given the hyperparameters $\pi(\boldsymbol{x}|\theta, \boldsymbol{y}) \approx \tilde{\pi}(\boldsymbol{x}|\theta, \boldsymbol{y})$.
3. Integrate over θ using a numerical method to get $\pi(x|y) \approx \int \tilde{\pi}(\boldsymbol{x}|\theta, \boldsymbol{y})\tilde{\pi}(\theta|\boldsymbol{y})d\theta$.

An advantage of INLA is the avoidance of sampling $\boldsymbol{x}$ directly, which is often high-dimensional. The sparsity of the precision matrix $\boldsymbol{Q}$ of the GMRF helps with computational efficiency. Spatio-temporal models are a type of LGM which means that their parameters can be estimated using INLA.

2.2.1.4 Stochastic Partial Differential Equation spatio-temporal models

There is a selection of spatial covariance functions that are used in spatial models. A common choice is the Matérn covariance function, which is defined by the following equation

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{d}{\rho}\right) \tag{2.7}$$

where d is the Euclidean distance between two locations, σ is the marginal covariance, $\nu > 0$ is the smoothness parameter, $\rho > 0$ is the range parameter, Γ is the Gamma function, and K_ν is the Bessel function of the second kind. The range controls how quickly the correlation decays with distance, and the smoothness controls the differentiability of the function and typically takes values of $\nu \in \{0.5, 1.5, 2.5, \dots\}$ for computational convenience. The Matérn is a popular choice of covariance function because both the range and the smoothness can be varied. Examples of Matérn covariance functions are shown in Figure 2.3.

A covariance matrix in traditional spatial statistics requires the calculation of all covariances between all pairs of observations. This communicates the full correspondence between any single point and all the others directly but requires a lot of computational power, especially as a dataset grows in size. A Gaussian random field can be approximated as a GMRF by discretising on a finite set of locations in a study region. This can be done by designing a mesh of the study region (often a Delauney triangulation, which is built from small triangles). It is recommended to construct a mesh that exceeds the boundaries of a study region to avoid boundary

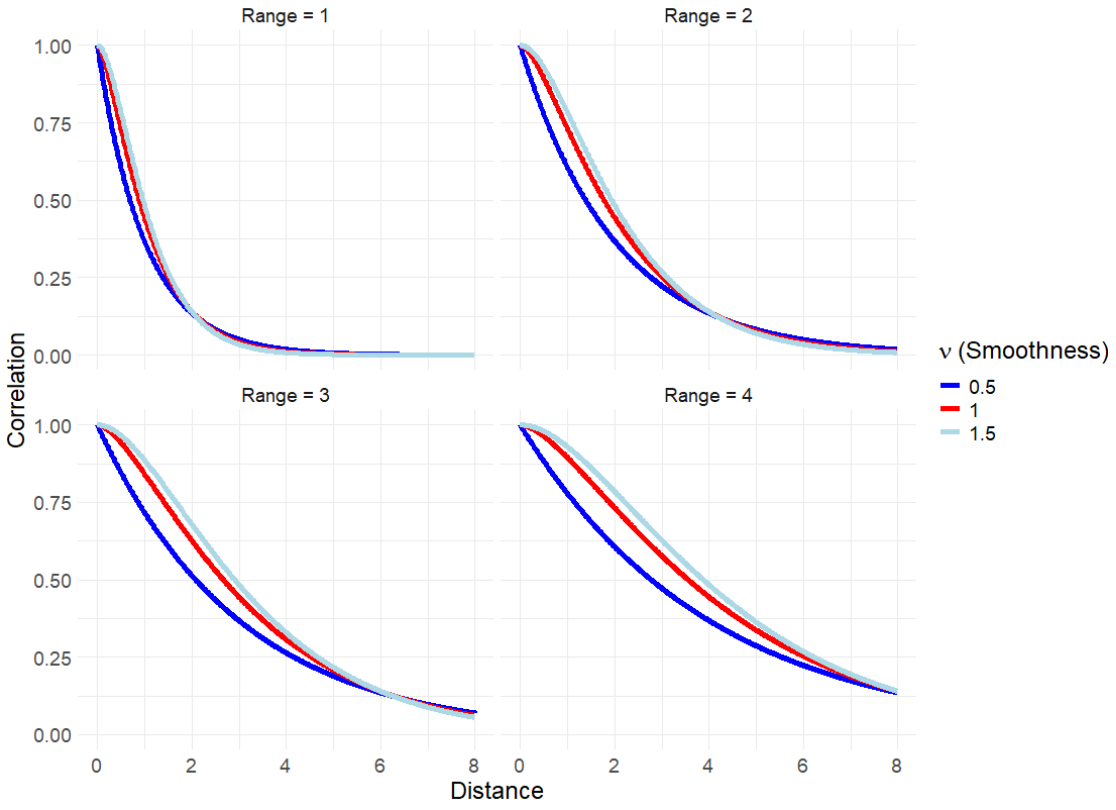

FIGURE 2.3: Examples of Matérn correlation functions with range $\rho = 1, 2, 3$, and 4 respectively and the smoothness $\nu = 0.5, 1$, and 1.5.

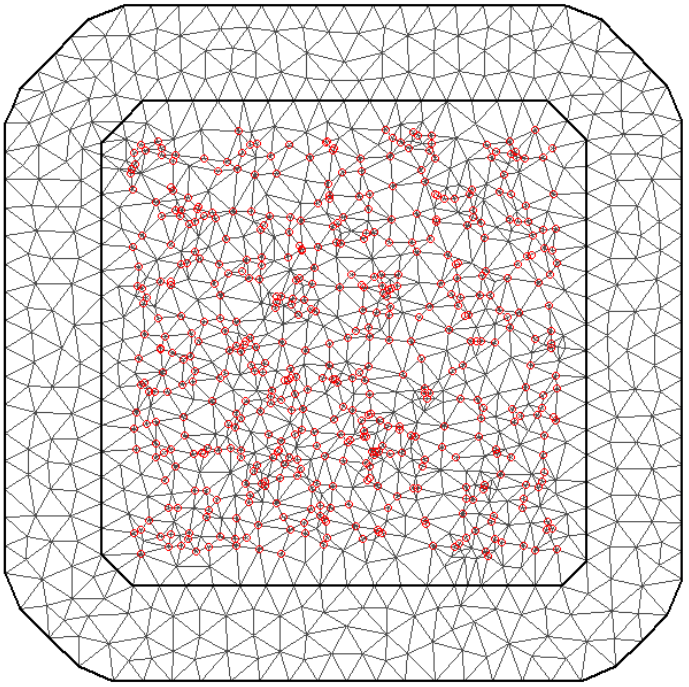

FIGURE 2.4: An example of a mesh described by a Delauney triangulation. Here the red dots show the locations of observations. The inner black line represents the boundary of the study region and any triangles outside that are incorporated to keep valid boundary conditions.

conditions affecting estimates of the GMRF at the edge of the study region. An example 2D mesh is shown in Figure 2.4. The GRF at any location is then approximated by a distance-weighted linear combination of piecewise linear basis functions ψ (one basis function for each vertex) of the GMRF at the vertices of the mesh (Figure 2.5). This approximation for spatial and spatio-temporal modelling was developed by Rue et al. (2009) and Lindgren et al. (2011). The spatial autocorrelation of the GMRF is represented in the precision matrix $\boldsymbol{Q}$ whose elements are zero unless they describe the covariance between vertices that share an edge in the mesh. This means that $\boldsymbol{Q}$ is largely comprised of zeros, which allows fast computations by leveraging methods from graph theory.

Consider the following stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}) \tag{2.8}$$

where $\mathcal{W}$ is spatial white noise, $u(\boldsymbol{s})$ is a Gaussian random field (GRF), κ is the spatial autocorrelation range of the GRF, and Δ is the Laplacian operator. Conveniently, the solution to this SPDE is the Matérn covariance function, which means this SPDE can be used to represent random fields in spatial models. This allows us to leverage numerical methods for solving partial differential equations to find approximations to the GRF using a mesh as described above. Examples of four samples from the same GMRF are shown in Figure 2.6. Note how the distance over which locations are strongly correlated remains similar across each sample.

The SPDE can be extended from a spatial model to a spatio-temporal model by allowing the GRF to vary in time. Temporal autocorrelation in GRFs is governed by a parameter $\alpha \in [-1, 1]$ and connects neighbouring GRFs at times t and $t+1$, denoted $u_t(\mathbf{s})$ and $u_{t+1}(\mathbf{s})$ respectively, by the relation $u_{t+1}(s) = \alpha u_t(s) + (1 - \alpha^2)x(\mathbf{s})$ where $x_t(\mathbf{s})$ is a random sample from the GRF. Examples of spatio-temporal GMRFs are shown in Figure 2.7 and Figure 2.8 with $\alpha = 0.5$ and $\alpha = 0.9$, respectively.

2.2.1.5 Barrier model

The covariance of a random field can be non-stationary, meaning its properties (such as its range) can vary over space, or time, or both (Bakka et al., 2018). One form of non-stationarity is the representation of physical barriers in the spatial autocorrelation via changing values in the precision matrix. The barrier model was first suggested and demonstrated by Bakka et al. (2019) for meshes in the plane. Using Bakka et al. (2019) as inspiration, I define a spatial autocorrelation structure for global oceanographic data, which represents land masses as barriers, across which correlation is heavily restricted.

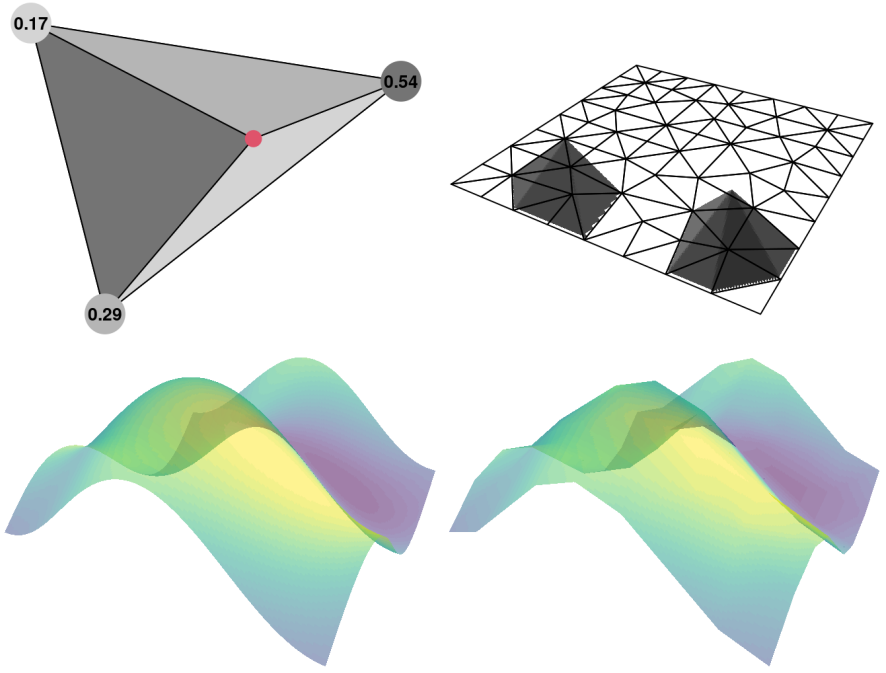

FIGURE 2.5: Illustration of how a GRF can be approximated by a GMRF. A discretisation of the study region in a Delauney triangulation allows for a finite number of locations to be approximated using linear combinations of linear basis functions. This approach means that the corresponding covariance matrix is very sparse (i.e., it contains many zeros) since most vertices are not neighbours of each other. Taken from Krainski et al. (2018).

Denoting the inner product (a generalisation of the dot product) as $\langle , \rangle$, and denoting the vector gradient as ∇, I define the following $n \times n$ matrices, where n is the number of vertices in the mesh.

$$\mathbf{C}_{ij} = \langle \psi_i, \psi_j \rangle \tag{2.9}$$

$$\mathbf{G}_{ij} = \langle \nabla\psi_i, \nabla\psi_j \rangle \tag{2.10}$$

$$\mathbf{K}_{ij} = \kappa^2 \mathbf{C}_{ij} + \mathbf{G}_{ij} \tag{2.11}$$

These matrices are the sums of contributions from all triangles in the mesh. The precision matrix $\mathbf{Q}$ for a given spatial autocorrelation range κ is given by

$$\mathbf{Q} = \mathbf{K}\mathbf{C}^{-1}\mathbf{K}. \tag{2.12}$$

For computational reasons, $\mathbf{C}$ can be replaced with a diagonal matrix $\tilde{\mathbf{C}}$, with $\tilde{\mathbf{C}}_{ij} = \langle \psi_i, 1 \rangle$. Refer to Lindgren et al. (2011) for justification of this.

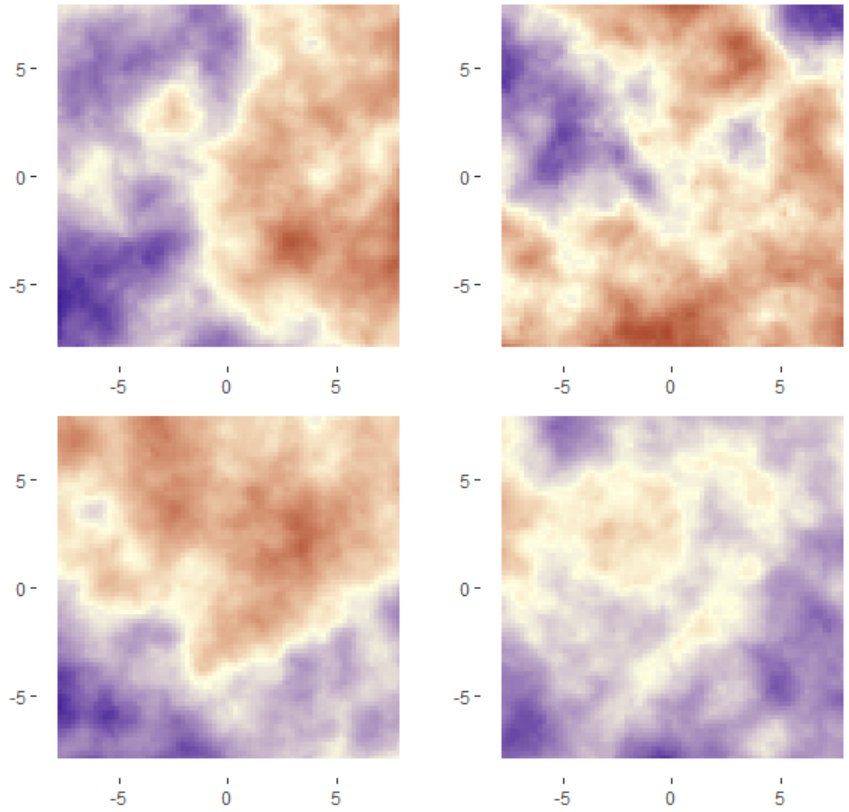

FIGURE 2.6: Four samples from a random field with range $\kappa = 0.2$. This random field is isotropic (the range is constant in all directions). Here the warmer and cooler colours represent positive and negative values in the random field respectively. Note how the spatial correlation length scale is similar across samples.

Suppose that a mesh triangle T has vertices $T = (\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2)$, each located in $\mathbb{R}^3$, with the corresponding edges

$$\mathbf{e}_0 = \mathbf{v}_2 - \mathbf{v}_1, \tag{2.13}$$

$$\mathbf{e}_1 = \mathbf{v}_0 - \mathbf{v}_2, \tag{2.14}$$

$$\mathbf{e}_2 = \mathbf{v}_1 - \mathbf{v}_0. \tag{2.15}$$

Suppose that the triangle T has area $|T|$, then the contributions of triangle T to $\tilde{\mathbf{C}}$ and $\mathbf{G}$ are

$$[\tilde{C}_{i,i}(T)]_{i=0,1,2} = \frac{|T|}{3}(1 \quad 1 \quad 1), \tag{2.16}$$

$$[G_{i,j}(T)]_{i,j=0,1,2} = \frac{1}{4|T|}(\mathbf{e}_0 \quad \mathbf{e}_1 \quad \mathbf{e}_2)^T(\mathbf{e}_0 \quad \mathbf{e}_1 \quad \mathbf{e}_2). \tag{2.17}$$

where $i = 0, 1, 2$ are the mesh vertex indices (i.e., nine elements for each triangle).

To implement the physical barrier condition, I reduced the contributions to $\tilde{\mathbf{C}}$ and $\mathbf{G}$ by mesh triangles whose centres are located over land by 80%. This reduces the spatial

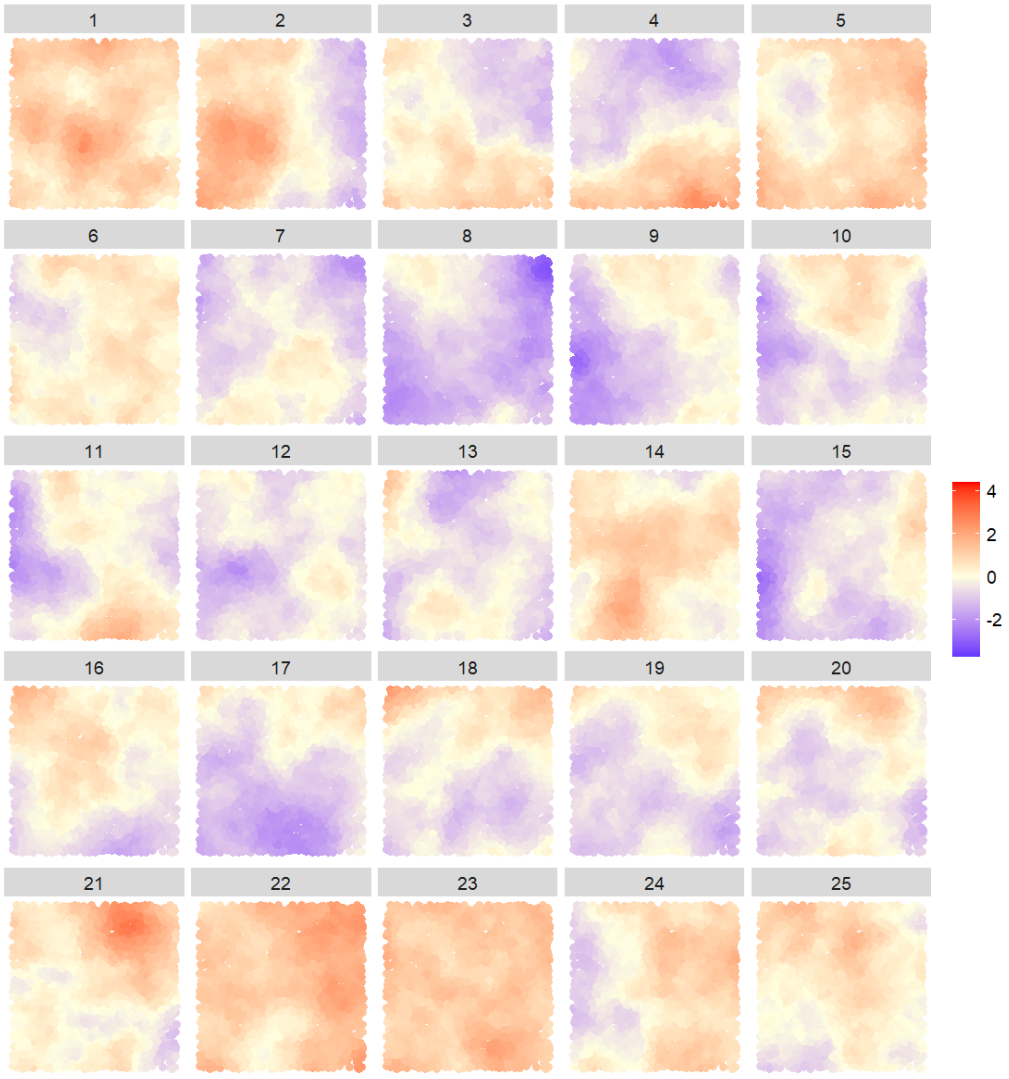

FIGURE 2.7: Example of a spatio-temporal GMRF with the autoregressive coefficient $\alpha = 0.5$. Each panel shows one of 25 regularly spaced times.

autocorrelation range significantly so that correlation takes 'the path of least resistance' around land masses spreading out radially from a point. Although this method differs from Bakka et al. (2019), the spatial autocorrelation structure behaves in a similar way. An example of this non-stationary correlation structure is presented in Figure 2.9. This type of covariance structure is what I use in the spatio-temporal models in Chapter 3.

2.2.2 Functional data analysis

2.2.2.1 Background

Functional data analysis (FDA) is a branch of statistics that contains methods for data objects that take the form of curves. Formally, this means that a single observation y

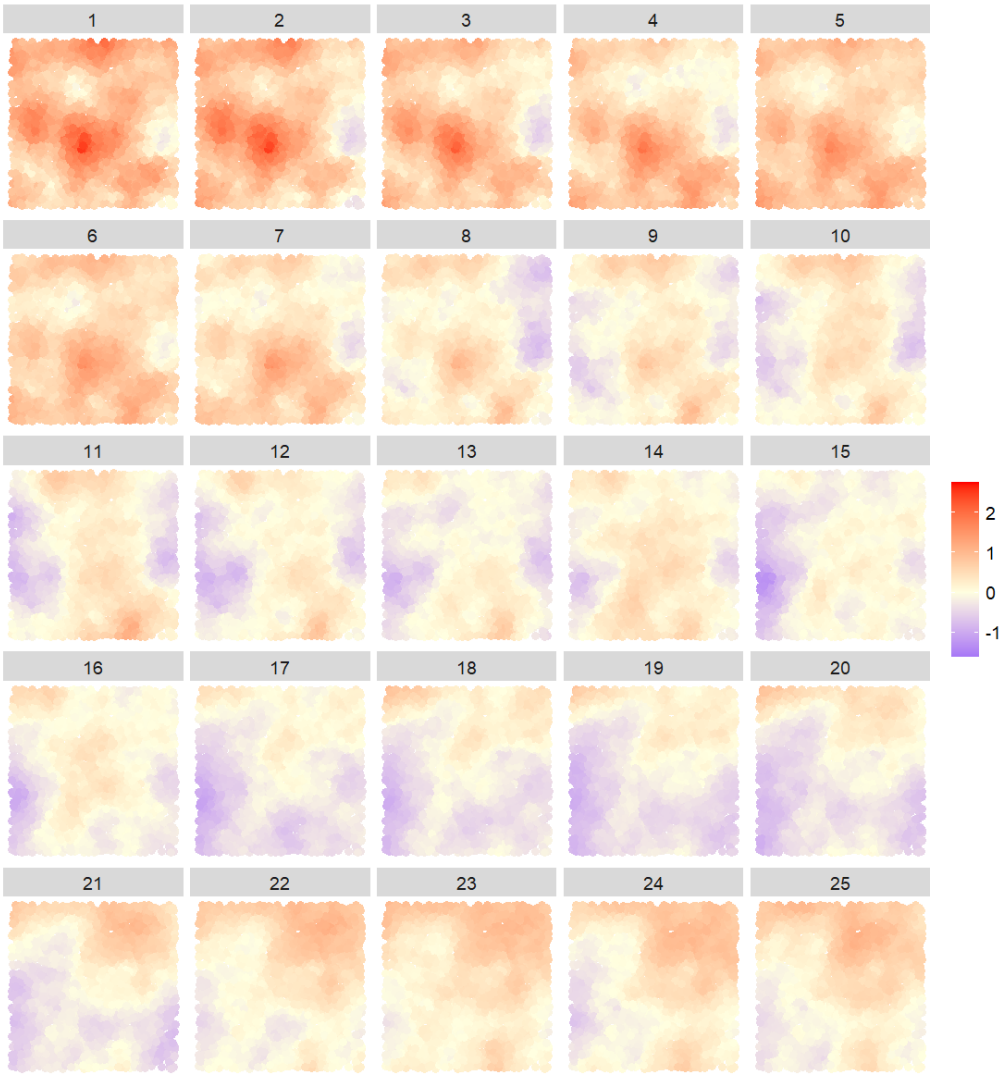

FIGURE 2.8: Example of a spatio-temporal GMRF with the autoregressive coefficient $\alpha = 0.9$. Each panel shows one of 25 regularly spaced times.

can be written in the form $y = f(t)$, where t belongs to some continuous domain. This contrasts with classical statistical methods, which are typically designed for scalar or vector data—that is, single values or finite-dimensional sets. In contrast, functional datasets are inherently infinite-dimensional, motivating the development of specialised statistical theory and methods. Oceanographic profiles provide an excellent opportunity to take advantage of these techniques (Figure 2.10). In the following, I briefly outline the mathematical foundations of FDA, providing context to some of the methods used in this thesis.

First, I define a random function X as a function of t over some closed and bounded interval $[a, b]$, denoted $X(t)$. Here, t can be referred to as the argument, indexing variable, or domain variable, with common examples being time and wavelength. Formally, random functions of the form $X(t)$ lie in the space L^2, which contains all continuous functions defined on a closed and bounded interval $[a, b]$, for which all

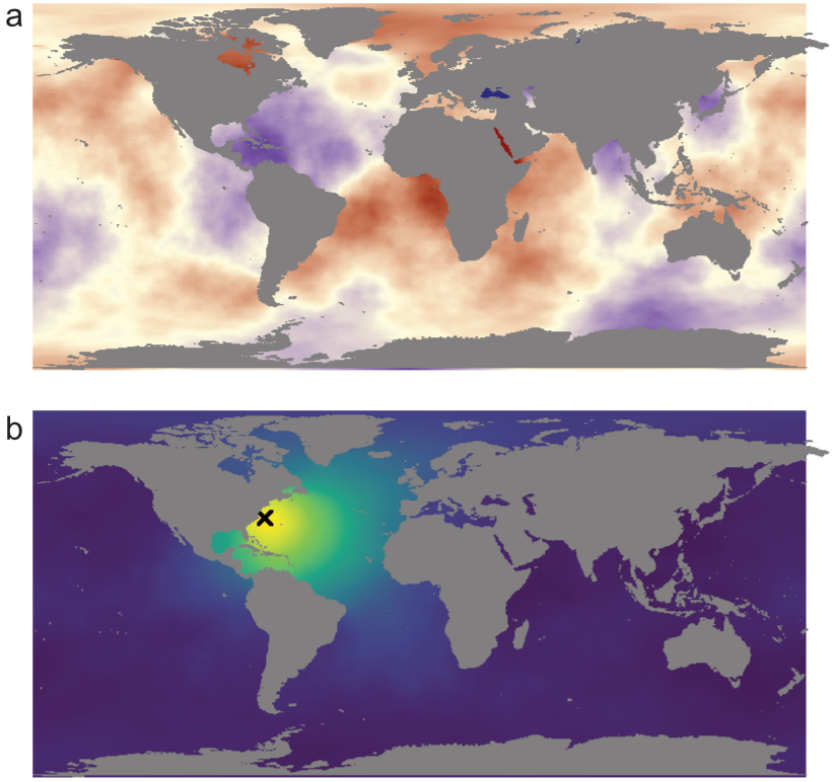

FIGURE 2.9: (a) A sample GMRF on the globe with the spatial correlation length $\kappa = 0.4$ (assuming the Earth's radius is 1) over the ocean and $\kappa = 0.08$ over land. (b) Correlation from 1000 random samples (with the same length scales as above) between a reference location in the northwest Atlantic (marked by the black cross) and the global ocean.

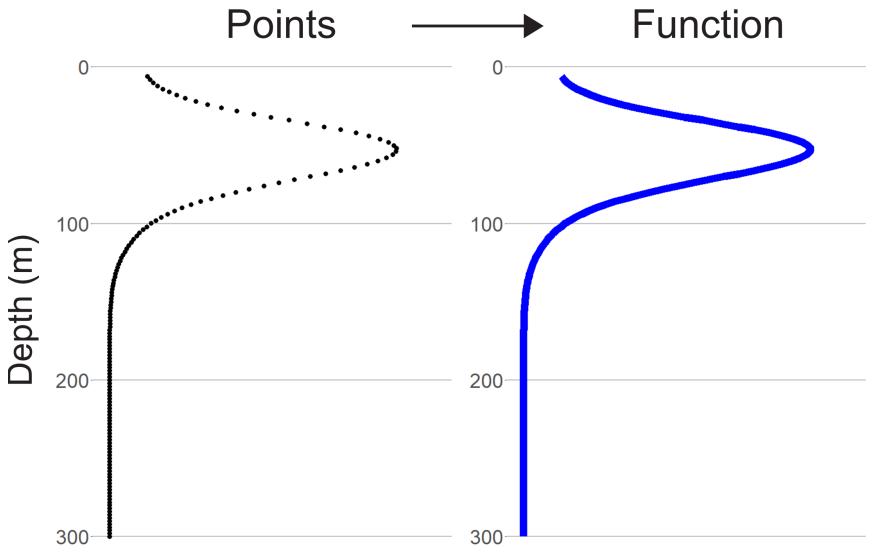

FIGURE 2.10: Here are two statistical perspectives of the measurements from a single Argo float profile. On the left, each measurement is considered one observation, regardless of depth. Consequently, the entire profile consists of multiple observations. Alternatively, on the right the entire profile is viewed as one functional data observation.

squared values are finite $\int_a^b |X(t)|^2 dt < \infty$. Intuitively, this means that the random function $X(t)$ has a finite amount of overall variation.

Suppose that I have n independent and identically distributed realisations $X_i(t) : t \in [a, b]_{i=1}^n$. Analogous to defining a mean and variance for random variables, I can define a mean function and a covariance function to summarise collections of realisations of random functions. The mean function $\mu(t)$ is defined as $\mu(t) = \mathbb{E}[X(t)]$. This implies that the mean function is defined over the same interval from a to b as each of the realisations and that the value of the mean function is simply the mean of all the realisations $X_i(t)$ at index t. The variance function is given by

$$\text{Var}(t) = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i(t) - \mu(t))^2 \tag{2.18}$$

and describes how much variability occurs at each point in the domain of t. The covariance function $\Sigma(s, t)$ is a surface with the domain $[a, b] \times [a, b]$ where $\Sigma(s, t) = \text{Cov}(X(s), X(t))$. In practice, this tells us how values of the function at two different indexes (s and t) tend to covary. The entire surface $\Sigma(s, t)$ summarises these pairwise covariances across the entire domain.

2.2.2.2 Functional regression models

Functional regression models (FRMs) are regression models that contain at least one functional variable with the same objective as traditional scalar regression: to quantify the effect of some covariate(s) on some response variable(s). Function-on-scalar regression aims to describe the effects of scalar-valued covariates on a functional response variable. Mathematically, this is written as

$$Y_i(t) = \beta_0(t) + \int \beta_1(t) X_{1i} dt + \cdots + \epsilon_i(t) \tag{2.19}$$

where the intercept $B_0(t)$ and error $\epsilon(t)$ are both functions. The intercept function can be thought of as a baseline outcome, which describes the outcome when the covariate is zero. The coefficient function $\beta_1(t)$ therefore represents deviations away from the baseline function, such that the integral contributes deviations proportional to the covariate value.

There are other categories of FRMs that are not presented in this thesis but I will mention them for completeness. Scalar-on-function regression describes the effects of function-valued covariates on a scalar response variable. Function-on-function regression describes the effects of functional covariates on functional response variables.

2.2.2.3 Scalar-valued variance for scalar functional data

The classic definition for the variance of functional datasets is also a function (Equation 2.18). This is problematic because functions in the L^2 space do not form an ordered set and consequently comparative statements between the variances of multiple functional datasets cannot be made, meaning correlation between functional datasets cannot be defined under this definition of variance. However, recently Urbano-Leon et al. (2023) developed an approach for defining the variance of functional data as a scalar. This is achieved through the use of basis functions and by defining the variance of a functional dataset as the sum of the variances of the basis coefficients. The idea can simply be extended to defining a scalar-valued covariance and correlation. To avoid repetition, I leave the mathematical details of this method for Chapter 5.

Chapter 3

Mapping Global Subsurface Chlorophyll Maxima Characteristics using Argo Floats and Spatio-temporal Models

This chapter is, at the time of writing, in preparation for publication in *Journal of Geophysical Research: Oceans* as:

Taylor, M., Henson, S., Sahu, S., Hammond, M., Cael, B.B. Mapping Global Subsurface Chlorophyll Maxima Characteristics using Argo Floats and Spatio-temporal Models.

3.1 Abstract

Subsurface chlorophyll maxima (SCMs), a feature in which peak chlorophyll concentrations occur well below the ocean surface, are a ubiquitous feature of the global ocean. The deployment of autonomous biogeochemical Argo (BGC-Argo) floats has significantly increased the availability and spatio-temporal coverage of chlorophyll profiles, enabling global analysis of SCM dynamics. In this study, I estimated two key properties of SCMs, intensity (Chl_{SCM}) and depth (z_{SCM}), using over 5600 BGC-Argo profiles from 2020. I applied a spatio-temporal geostatistical modelling framework utilising the integrated nested Laplace approximation to quantify the effects of physical and biological covariates on SCM properties while accounting for spatial and temporal autocorrelation. My results demonstrated that including a spatio-temporal random effect improves model performance and yields more reliable estimates of environmental drivers. The euphotic depth emerged as a major driver of the SCM structure, positively related to z_{SCM} and negatively related to

Chl_{SCM}. Other influential factors included mixed layer depth, sea surface height anomaly, and zooplankton biomass. I applied my fitted spatio-temporal models to interpolate SCM characteristics across 20 000 locations sampled by core Argo floats, producing global predictions of SCM structure. These maps revealed spatial patterns and seasonal dynamics consistent with previous observational studies, including deep SCMs in subtropical gyres and pronounced seasonality at higher latitudes. My approach demonstrated the utility of formal spatio-temporal models for analysing BGC-Argo float data and provides a framework for future studies of global ocean biogeochemistry that accounts for dependencies between observations, whilst keeping computational costs relatively low.

3.2 Introduction

Chlorophyll concentration is a commonly used proxy for the biomass of phytoplankton, microscopic plant-like organisms that form the base of the marine ecosystem and perform a key role in the global carbon cycle. Subsurface chlorophyll maxima (SCMs) are ubiquitous features of the global ocean, whereby the maximum concentration of chlorophyll is located significantly below the surface of the ocean. They form when the water column is stratified, and nutrients and light are limited from above and below respectively (Cullen, 2015). SCMs account for a significant proportion of the global marine primary production (Silsbe and Malkin, 2016). Although they have been studied for several decades, there is still some uncertainty over which environmental factors control their formation and maintenance. SCMs can form either through an accumulation of phytoplankton (i.e., an increase in their abundance) (Herbland and Voituriez, 1979; Holm-Hansen and Hewes, 2004; Beckmann and Hense, 2007), or through an increase in their intra-cellular chlorophyll content as a proportion of their mass (Fennel and Boss, 2003; Masuda et al., 2021). This second mechanism is called photoacclimation and is a physiological response to light limitation.

The Argo float array is a global network of autonomous platforms that measure properties of sea water throughout the top 2000 m of the water column. The array comprises core floats, which carry sensors for measuring temperature and salinity, and biogeochemical-Argo (BGC-Argo) floats, which are equipped with additional sensors for monitoring several key variables in marine biogeochemistry (Claustre et al., 2020). One sensor measures the fluorescence of chlorophyll, from which chlorophyll concentration is estimated (Schmechtig et al., 2023). The deployment of BGC-Argo floats has hugely increased the number of subsurface observations, as well as the spatio-temporal coverage of observations, particularly in remote environments which were previously inaccessible to ship-borne sampling, one example being the Southern Ocean during winter.

Several studies have identified global patterns in the distribution of SCMs using data from the BGC-Argo float network (Cornec et al., 2021a; Yasunaka et al., 2021; Bock et al., 2022). These studies consistently report the deepest SCMs, often reaching depths of 150 m, in the centres of subtropical oligotrophic gyres. Notably, these deep SCMs are not associated with coincident peaks in particle backscatter, a proxy for particulate organic carbon. Deeper SCMs tend to exhibit lower chlorophyll concentrations both at the maximum and at the surface (Uitz et al., 2006; Xu et al., 2022b; Quartly et al., 2023; Miyares et al., 2024). While SCMs are most prevalent in the subtropics and tropics, they also form in summer at higher latitudes, including polar regions (Baldry et al., 2020; Bouman et al., 2020; Bendtsen et al., 2023; Baldry et al., 2024), emphasising their global significance. As expected, SCM seasonality is more pronounced at higher latitudes due to large seasonal variations in day length and surface irradiance (Cornec et al., 2021a; Yasunaka et al., 2021). SCMs are not observed in the high-latitude winters.

It is well established that the primary determinants of SCM depth (z_{SCM}) are the euphotic depth (z_{eu}) (Cullen, 2015; Xu et al., 2022b) and the nitracline depth (z_{ncline}) (Herbland and Voituriez, 1979; Cullen, 2015; Gong et al., 2015). These represent important depths regarding light availability and nutrient concentration, respectively, and they often constrain the optimal conditions for phytoplankton growth beneath the surface. Thermocline structure, particularly the mixed layer depth (MLD), has also been shown to modulate SCM characteristics by influencing vertical mixing and stratification (Itoh et al., 2015; Zampollo et al., 2023). A shallow mixed layer, for example, may reduce nutrient availability near the surface, and improve conditions for SCM formation. Mesoscale ocean dynamics, such as eddies and fronts, introduce additional variability by changing the vertical structure of both light and nutrients. For instance, anticyclonic eddies are typically associated with a deepening of isopycnals, deepening the z_{eu} and z_{ncline}, and thereby promoting deeper SCMs (Cornec et al., 2021b). In contrast, cyclonic eddies often shoaling of isopycnals, leading to shallower SCMs (Cornec et al., 2021b; Xu et al., 2022b; Strutton et al., 2023). Furthermore, grazing pressure from zooplankton, particularly near the surface, can suppress chlorophyll concentrations in the upper water column and indirectly enhance the formation or deepening of SCMs by shifting the balance between phytoplankton growth and loss processes (Pilati and Wurtsbaugh, 2003). It is worth noting that a similar effect on SCMs can be caused by mixotrophs, organisms that can switch between photosynthesis and grazing as sources of energy depending on light availability (Moeller et al., 2019).

No previous studies of SCMs using BGC-Argo float data have accounted for dependencies between nearby observations through formal spatio-temporal statistical modelling. Spatial and spatio-temporal models have gained considerable attention over the past two decades, with methods tailored to increasingly large and unique datasets. A key category is geostatistical models, where each observation is tied to a

location, and potentially a time in the case of spatio-temporal models. These models typically aim to explain variability within a dataset, and predict values at unobserved locations and times. Furthermore, the Argo dataset poses an interesting statistical challenge: floats move between observations, so no location is ever sampled twice, complicating the modelling of temporal autocorrelation (Stein, 2020). Nonetheless, several studies have explored mapping Argo-derived variables. Sahu and Challenor (2008) used a hierarchical Bayesian model to map temperature and salinity in the North Atlantic. Roemmich and Gilson (2009) developed a local regression model to produce the first global climatologies from core Argo data. Later efforts continued modelling data from specific depths (Kuusela and Stein, 2018; Sahu, 2022). Yarger et al. (2022) took a different approach, treating Argo profiles as functional data to map temperature globally across all depths. While this technique provided several advantages, including 3D prediction, it relied on a relatively uniform and dense coverage of profiles, which is feasible with core Argo floats but not with BGC-Argo floats equipped with chlorophyll sensors.

In this work I apply a novel spatio-temporal modelling method for Argo data, using the stochastic partial differential equation (SPDE) approach first developed by Lindgren et al. (2011). They found that continuous Gaussian random fields (GRFs) (the additional latent effect that makes a model spatial or spatio-temporal) can be discretised into Gauss-Markov random fields (GMRFs). This hugely reduces the computational complexity and allows for larger datasets to be modelled more efficiently, without any loss in accuracy. This technique has previously been used for climate and oceanographic datasets since they are typically sizeable (Dahlén et al., 2020; Fuglstad and Castruccio, 2020; Fioravanti et al., 2023). An additional benefit of this technique is the specification of non-stationary covariance structures (Bakka et al., 2018; Hildeman et al., 2021). I use a method for introducing non-stationary covariance to study ocean regions with hard boundaries (coastlines), inspired by that of (Bakka et al., 2019). This involves reducing the spatial correlation length over land and forcing correlation to pass along a coastline. This has been done on a local scale previously (Chaudhuri et al., 2023) but not on a global scale to my knowledge.

There are two main objectives for this study. Firstly, to identify the relative importance of various environmental conditions on SCM depth and intensity through the use of a spatio-temporal statistical model. Second, to make predictions of SCM characteristics at previously unobserved locations using the fitted models. By leveraging the structure of global biogeochemical datasets, I aim to demonstrate the value of incorporating spatio-temporal dependencies into chlorophyll modelling, by fitting models with and without spatio-temporal latent effects.

3.3 Data

3.3.1 Argo float data

I used a dataset containing measurements from 5602 BGC-Argo and 20 000 core Argo profiles completed in 2020 whose locations are locations are shown in Figure 3.1. This dataset included profiles located in all major ocean basins during each month. This data was accessed from the ifremer index (https://data-argo.ifremer.fr) on 12th April 2023 through the R package **argoFloats** (Kelley et al., 2021b). This package was used to select profiles from 2020 and to perform a quality control check of profiles to remove profiles considered 'Probably bad' or 'Bad' (Argo quality control flags 3 or 4, respectively).

3.3.1.1 Chlorophyll

The chlorophyll variable from BGC-Argo float data was used to estimate the intensity and depth of SCMs. I removed profiles which did not meet the following criteria: (1) at least one observation in the top 15 m, (2) at least one observation below 150 m, (3) and at least 20 observations in total. Each chlorophyll profile was smoothed using a running median over a span of five measurements, similar to Cornec et al. (2021a). Two SCM characteristics were identified for each chlorophyll profile: the maximum concentration (Chl_{SCM}) and the SCM depth (z_{SCM}, i.e. the depth at which the maximum concentration is observed). I chose to estimate the SCM characteristics like this to avoid fitting mathematical curves to profiles as in Carranza et al. (2018) and Xu et al. (2022b), since such a wide range of profile shapes would have required a complicated function, which would have been computationally quite expensive. I applied a $\log_{10}$ transformation of Chl_{SCM}, as chlorophyll has a log normal distribution (Campbell, 1995). This had the added benefit of forcing positive values for predictions of chlorophyll concentration.

3.3.1.2 Temperature and salinity

I calculated potential density profiles using the Gibbs seawater equations (Kelley et al., 2021a), based on temperature and salinity profiles from both BGC and core Argo floats. The same criteria for removing chlorophyll profiles based on the number of measurements were applied to potential density profiles. I then estimated the mixed layer depth (MLD) by identifying the shallowest depth which exceeded the surface potential density by at least 0.03 kg m^{-3} (de Boyer Montégut et al., 2004).

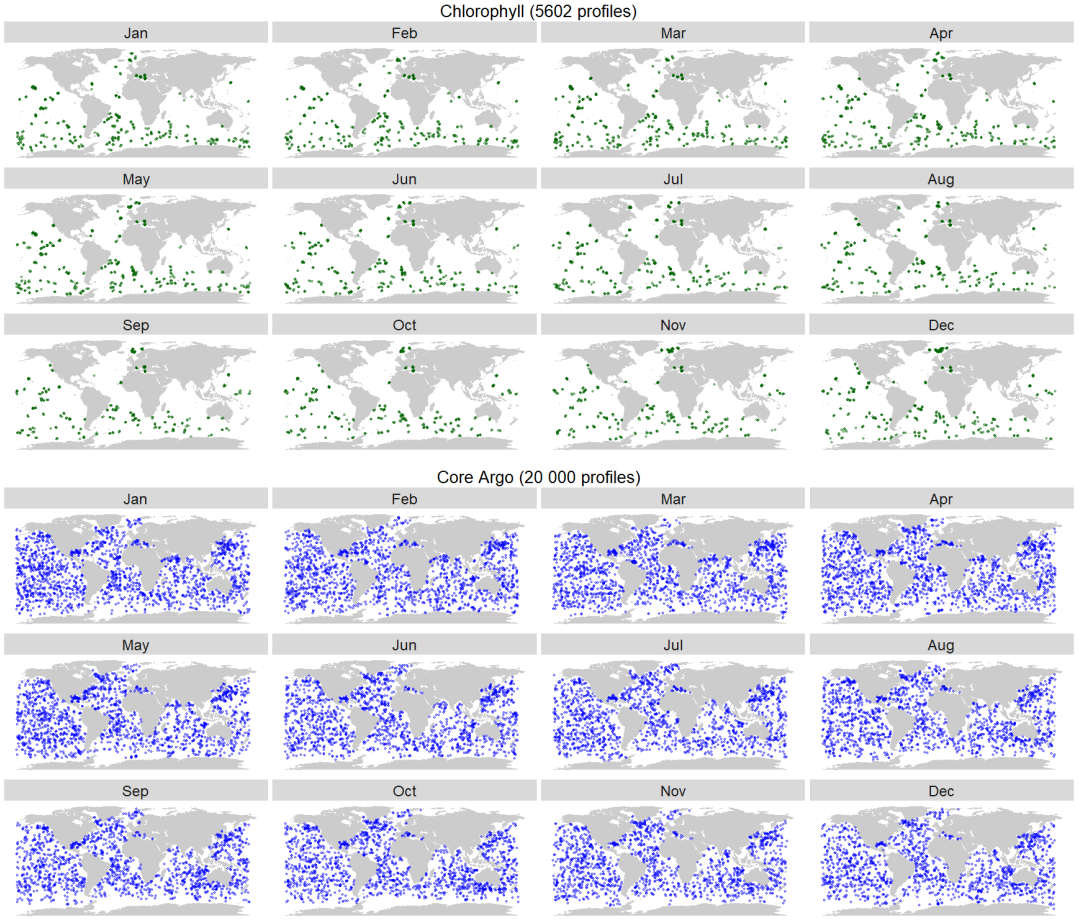

FIGURE 3.1: Locations of all of the BGC-Argo profiles (green) and core Argo profiles (blue) used in this work. Note the sparseness of the BGC-Argo profiles in comparison to the core Argo profiles, and clustering of BGC-Argo profiles in several regions, such as in the Southern Ocean.

3.3.2 Gridded data products

I supplemented the Argo float dataset with additional variables from 3D hindcast products from https://data.marine.copernicus.eu/products. I chose products whose variables had near-global coverage and daily temporal resolution in order to match the time and location of Argo float profiles as closely as possible. Both the following datasets were last accessed on 13th April 2023.

3.3.2.1 Euphotic depth and zooplankton abundance

I obtained estimates for the euphotic depth and zooplankton abundance from the "Global ocean low and mid trophic levels biomass content hindcast" Copernicus product (https://doi.org/10.48670/moi-00020). This product is a hindcast model and has a spatial resolution of 1/12° and daily temporal resolution. The zooplankton abundance is measured as the carbon mass of zooplankton in the water column (g

m^{-2}). The mean absolute difference of the zooplankton biomass is 0.44 gC m^{-2} and is a slight overestimate of observations. This product does not indicate where in the water column the zooplankton are located, although I still include it as it could be a useful indicator for the interactions between different trophic levels. The quality information document for the this gridded product details that uncertainty comes from a variety of sources, firstly from raw data collection and then from the model and its forcings, which is greater than the uncertainty in the euphotic depth variable in this product.

3.3.2.2 Sea surface height anomaly

I used the Copernicus "Global Ocean Gridded L 4 Sea Surface Heights And Derived Variables" product (https://doi.org/10.48670/moi-00148) which has 1/8° spatial resolution and daily temporal resolution. The sea surface height anomaly (SSHA) is calculated with reference to the 20-year average between 1993-2012 and it is estimated through optimal interpolation of along-track measurements. I use SSHA as a proxy for the presence, polarity and intensity of mesoscale eddies (Chelton et al., 2011), which have been shown to affect vertical chlorophyll distribution (Cornec et al., 2021b). The main sources of uncertainty in this product come from the sampling frequency of the altimeters and the interpolation onto a regular grid, but there is good confidence in the SSHA data with altimeter measurements having a root mean square error of around 1 cm.

3.4 Methods

I took a statistical modelling approach to identify the physical and biological influences of SCM properties using data from BGC-Argo profiles. SCM properties have previously been modelled statistically on a regional scale without the inclusion of spatio-temporal latent effects (Xu et al., 2022b) but to my knowledge there are no examples on a global scale. BGC-argo float data has been used to assess global chlorophyll patterns (Cornec et al., 2021a; Bock et al., 2022), however these studies focussed on classifying locations and defining general profile types rather than spatial interpolation between observations across ocean basins. Here I fit statistical models to BGC-Argo float data and then use the results to interpolate across the global ocean at locations of the more widespread core Argo profiles whilst extracting the different sources of profile variability.

3.4.1 Bayesian spatio-temporal models for global oceanographic data

Model structure I fitted two Bayesian models to each of the SCM properties: z_{SCM} and Chl_{SCM}. The first model was a normal linear model with the following covariates as fixed effects: euphotic depth, MLD, sea surface height anomaly (SSHA), day length, and zooplankton biomass. Table 3.1 describes justifications of these covariates. This model was specified as:

$$Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_{ip} + \epsilon_i \tag{3.1}$$

for observations $i = 1, \ldots, N$ and parameters $p = 1, \ldots, P$ and where X is the design matrix and $\epsilon_i \sim \text{N}(0, \sigma^2_{\text{obs}})$ is a random zero-mean error term. The second model was identical to the first other than the addition of a spatio-temporal random effect U. This extra term accounted for the unmeasured sources of variation, which may be specific to particular locations or seasons. This allowed the similarities of neighbouring observations to be incorporated within in the model. Furthermore, the inclusion of latent effects could have potentially offset any biases in fixed effect coefficients introduced by the clustering of BGC-Argo profiles. The model specification became

$$Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_{ip} + U(\mathbf{s}_i, t_i) + \epsilon_i \tag{3.2}$$

where $\mathbf{s}_i$ and t_i are the location and month of observation i respectively.

Spatio-temporal latent field In theory, the spatio-temporal random effect $U(\mathbf{s}, t)$ is a continuous surface that varies smoothly over space and time. Gaussian random fields (GRFs) are commonly used to model such effects, as they allow spatial and temporal dependence to be specified through a covariance function. However, GRFs require the computation of a full covariance matrix, which involves calculating pairwise distances between all observations—a task that becomes computationally prohibitive for large datasets. The Matérn covariance function is frequently used in this context because it provides a flexible way to model spatial autocorrelation based on distance.

To address the computational limitations of GRFs, I instead used Gaussian Markov random fields (GMRFs), which represent a discretised approximation of GRFs. This approach involved constructing a triangular mesh over the study region and assuming that spatial dependence only existed between neighbouring mesh nodes. As a result, the number of necessary distance calculations was drastically reduced, greatly improving computational efficiency. This approximation was justified because the Matérn covariance function can be derived as the solution to a specific stochastic partial differential equation (SPDE) (Equation 2.8), which can be solved numerically using the mesh-based finite element method described in Section 2.2.1.

A further advantage of the SPDE approach is that it allows for non-stationary covariance structures (Bakka et al., 2018). I used this to impose a physical barrier correlation structure (Bakka et al., 2019), where the Matérn correlation length scale was reduced over land such that correlation 'flows' around coastlines. Since this technique has not been used before for large-scale oceanographic data, I chose for the correlation length scale to be five times shorter on land than over ocean. This value was chosen as it matches the example in the original methods paper (Bakka et al., 2019), however it was acknowledged that this arbitrary selection is a limitation of the method. I considered this an important addition for my study given that two locations with land between can have significantly different physical and biogeochemical processes despite being a short physical distance apart (for example, the Arabian Sea and the Bay of Bengal), especially given the clustered spatial distribution of the BGC-Argo floats.

The temporal autocorrelation of the spatio-temporal random field between consecutive months was assumed to be an autoregressive process. This has a single parameter $\alpha \in [-1, 1]$ controlling the autocorrelation. Values of α close to 1 indicate high temporal autocorrelation in which the GMRF associated with each month will be similar to those for the previous and following months.

Priors There were four hyperparameters in this model controlling the spatio-temporal component: the range of U (the distance beyond which spatial correlation is negligible), the marginal variance of U (controlling the amplitude of spatial variation), the temporal autocorrelation coefficient of U and the precision for the observations. Each of these hyperparameters had a prior distribution over their respective valid parameter space. Specifically, I defined weakly informative priors over the positive real numbers for the range, marginal variance, and precision, and over the interval (-1,1) for the temporal autocorrelation coefficient α.

Model fitting From a spatial statistics perspective, my dataset is moderately sized, making estimation of the spatial random effect U computationally intensive through traditional Markov chain Monte Carlo (MCMC) methods. To address this, I utilised an approximate Bayesian inference method called integrated nested Laplace approximations, which provide accurate parameter estimates in a fraction of the time MCMC would take. Since my study region encompasses the global ocean, the SPDE mesh was constructed on the surface of a sphere. I specified a fine mesh resolution over the ocean to improve accuracy, and a coarser resolution over land to provide boundary conditions while minimising computational cost. To preserve hard boundaries in narrow regions such as the Isthmus of Panama, I used a mesh with a maximum edge length of approximately 100 km over the ocean. This resulted in a mesh containing over 36 000 nodes (Figure 3.2). Observation locations are linked to

the mesh through a linear interpolation of the spatial field using the three mesh nodes forming the triangle in which each observation lies (see Lindgren et al. (2011) for details).

The data was randomly split into a training set (80%) to fit the models and a validation set (20%, 1120 observations) to assess model fit. Each spatio-temporal model took approximately 28 minutes on a standard laptop with an Intel core i5 processor whereas the normal linear models took considerably less time. The computation time increased exponentially with each additional month of profile data, which motivated the decision to limit this study to a 12-month period.

3.4.2 Model selection

I used the Watanabe-Akaike information criterion (WAIC) to assess model fit whilst accounting for model complexity. It is a Bayesian extension of the Akaike information criterion that includes information from the entire posterior distribution for each parameter, rather than a point estimate (the mode of the posterior). I also calculated the root mean square error (RMSE), which provides a measure of the differences between observations and predictions. I used these in conjunction with the statistical coverage of the model, which is the proportion of observations whose true value lies within a 95% prediction interval (Sahu, 2022).

3.4.3 Spatial prediction and interpolation

Locations of 20 000 randomly selected core Argo profiles from 2020 were used for prediction using the model output. This meant that all the covariates were available for predictions and were calculated in the same way as the training dataset. Given the spatial distribution of core Argo floats, this provided a relatively uniform distribution of prediction locations over the global ocean. The number of predictions for each month ranged from 1250 to 1842.

3.5 Results

3.5.1 Model comparison

I fitted statistical models to BGC-Argo float data to investigate the drivers of SCM intensity and depth. For each SCM property, I implemented two models: a standard linear model and a spatio-temporal model. Model performance was assessed using a validation dataset by comparing predicted values to observed outcomes. In all cases,

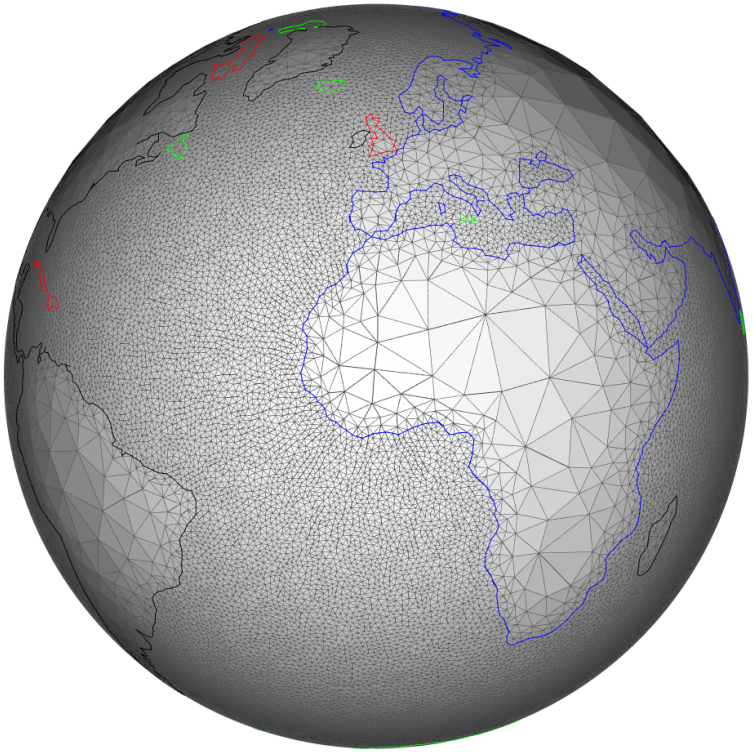

FIGURE 3.2: An example of a high resolution global triangulation mesh for SPDE models. This mesh contains approximately 36 000 vertices, with a very fine mesh over the region of interest (i.e., the ocean) and a low resolution elsewhere (i.e. over land) in order to minimise the computational cost of fitting the spatio-temporal models. Each land mass is shown by a coloured boundary.

Variable	Why include it in the models?	Expectations
MLD	The MLD represents the boundary between well mixed surface water and nutrient-rich deep water.	As the MLD increases, Chl_{SCM} increases and z_{SCM} decreases.
Euphotic depth	The depth to which light can penetrate should affect the growth and abundance of phytoplankton as (Cullen, 2015).	Greater euphotic depths should correspond with deeper and weaker SCM.
Day length	Polar regions experience highly seasonal light levels.	Longer days might encourage more intense SCMs, but I expect their depth will not be affected significantly by daylight hours.
Zooplankton biomass	Zooplankton grazing can influence vertical phytoplankton distribution (Moeller et al., 2019).	I expect that high zooplankton abundance will force weaker and deeper SCMs.
SSHA	Previous studies have shown that SCMs are affected by (sub)mesoscale physics (Cornec et al., 2021b; Xu et al., 2022b).	Positive (negative) SSH anomalies will result in more (less) intense and shallower (deeper) SCMs.

TABLE 3.1: Summary of fixed effects included in the linear model and the spatio-temporal model.

	Chl_{SCM}		z_{SCM}	
	Normal linear	Spatio-temporal	Normal linear	Spatio-temporal
WAIC	1977	-1864	43082	40608
RMSE	0.295	0.195	31.1	24.6
Coverage	9%	76%	6%	52%

TABLE 3.2: Model comparison using WAIC, RMSE, and predictive coverage. For both Chl_{SCM} and z_{SCM}, the spatio-temporal model outperforms the linear model.

the spatio-temporal model outperformed the linear model, even after accounting for the increased complexity introduced by the spatio-temporal component (Table 3.2). The root mean square error (RMSE) for Chl_{SCM} and z_{SCM} was reduced by 33% and 21%, respectively, when using the spatio-temporal model instead of the normal linear model. Additionally, the statistical coverage, which reports the proportion of 95% prediction intervals that contain the corresponding observed value, substantially improved for both SCM properties under the spatio-temporal model. Specifically, it increased from 9% to 76% for Chl_{SCM} and from 6% to 52% for z_{SCM}. Maps showing the spatio-temporal breakdown of statistical coverage of z_{SCM} and Chl_{SCM} in Figures A.1 and A.2, respectively.

3.5.2 Drivers of SCM properties

Figure 3.3 shows the posterior distributions of the five fixed effects for each of the SCM properties. Fixed effects were considered significant if their 95% credible intervals excluded zero. Similarly, a significant change in effect size between models occurred when they overlapped by less than 5%. In the Chl_{SCM} model, all covariates were statistically significant except for sea surface height anomaly. The inclusion of the spatio-temporal effect led to notable changes in the estimated effects of z_{eu}, zooplankton biomass, and MLD. Following the inclusion of the spatio-temporal effect, the z_{eu} had the strongest influence on Chl_{SCM}, with an estimated coefficient of –0.5 per 100 m increase in z_{eu}, corresponding to a three-fold decrease. Incorporating the spatio-temporal effect reduced the effect sizes of z_{eu} and MLD, while increasing the effect of zooplankton biomass. In the z_{SCM} model, all covariates became statistically significant after including the spatial random effect, with the posterior distribution for the MLD effect shifting away from zero. Zooplankton biomass was the only covariate with a negative effect. Euphotic depth and SSHA exhibited the strongest influences on z_{SCM}. The estimated effect of SSHA remained largely unchanged with the addition of the spatio-temporal component, while the uncertainty in the zooplankton biomass effect increased.

The latent spatio-temporal effect for Chl_{SCM} (denoted $U_{Chl_{SCM}}$) displayed a negative effect in both polar regions during their respective winters (Figure 3.4). In these regions, the effect was often as low as -1 (corresponding to an order of magnitude

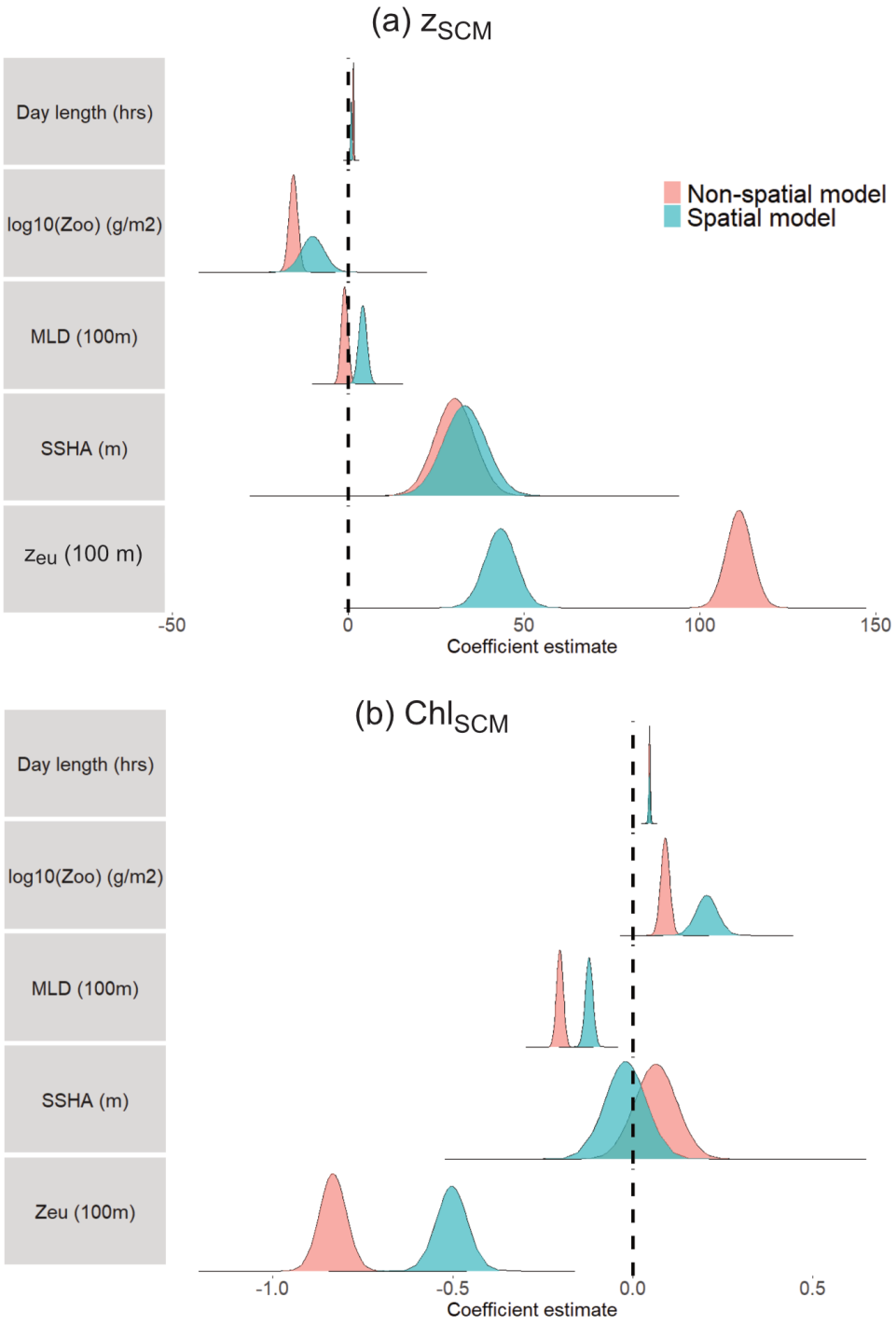

FIGURE 3.3: Comparison of fixed effect coefficient posterior distributions for the spatial models (blue curves) and non-spatial models (red curves) for (a) z_{SCM} and (b) Chl_{SCM}. The dotted line denotes no effect, and a statistically significant effect was defined as one whose 95% credible interval did not contain zero. Note how the magnitude and even the sign of some effects changes when spatio-temporal autocorrelation is included in models.

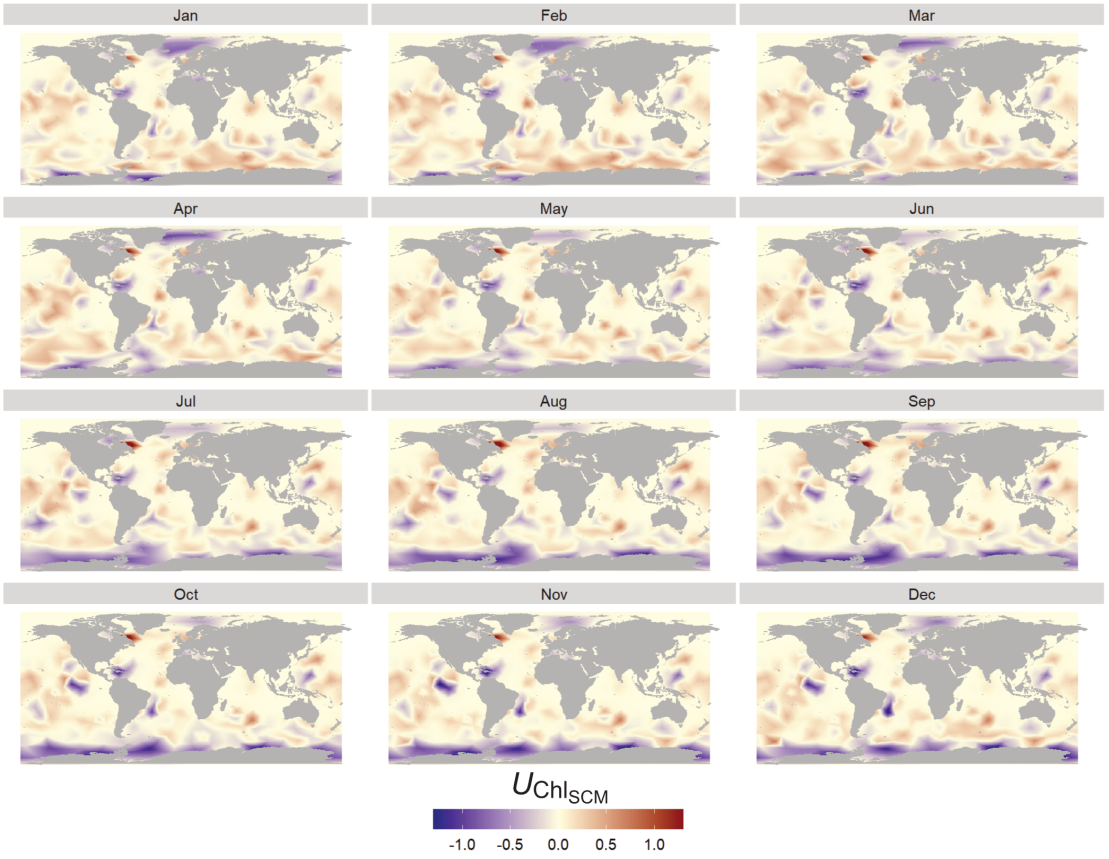

FIGURE 3.4: Monthly estimates of the latent effect included in the spatio-temporal model for $\log_{10}(\mathrm{Chl_{SCM}})$ which accounts for variability unexplained by the fixed effects.

decrease in $\mathrm{Chl_{SCM}}$). This suggests that the primary driver(s) controlling these chlorophyll profiles (e.g. light intensity) was not included in the model. No clear pattern was seen in $U_{\mathrm{Chl_{SCM}}}$ throughout the tropics and subtropics, which indicates that in these regions the fixed effects explained most of the variability and the latent effect had a smaller role over large-scales. In contrast the latent effect for the z_{SCM} model (denoted $U_{z_{\mathrm{SCM}}}$) displayed large positive values in each of the subtropical oligotrophic gyres (expect the Indian Ocean, possibly due to a lack of observations) (Figure 3.5). The effect increased the predicted depth of SCMs by several tens of metres, the most evident being in the South Pacific and, to a lesser extent, the North Atlantic and North Pacific, which were present year-round. A seasonal effect was seen in the polar summers, reducing the depth of predicted SCMs by around 20 - 40 m. In summary, the latent effects pick up some large-scale variability not included in the covariates for $\mathrm{Chl_{SCM}}$ and absorbs unexplained regional variability in z_{SCM}.

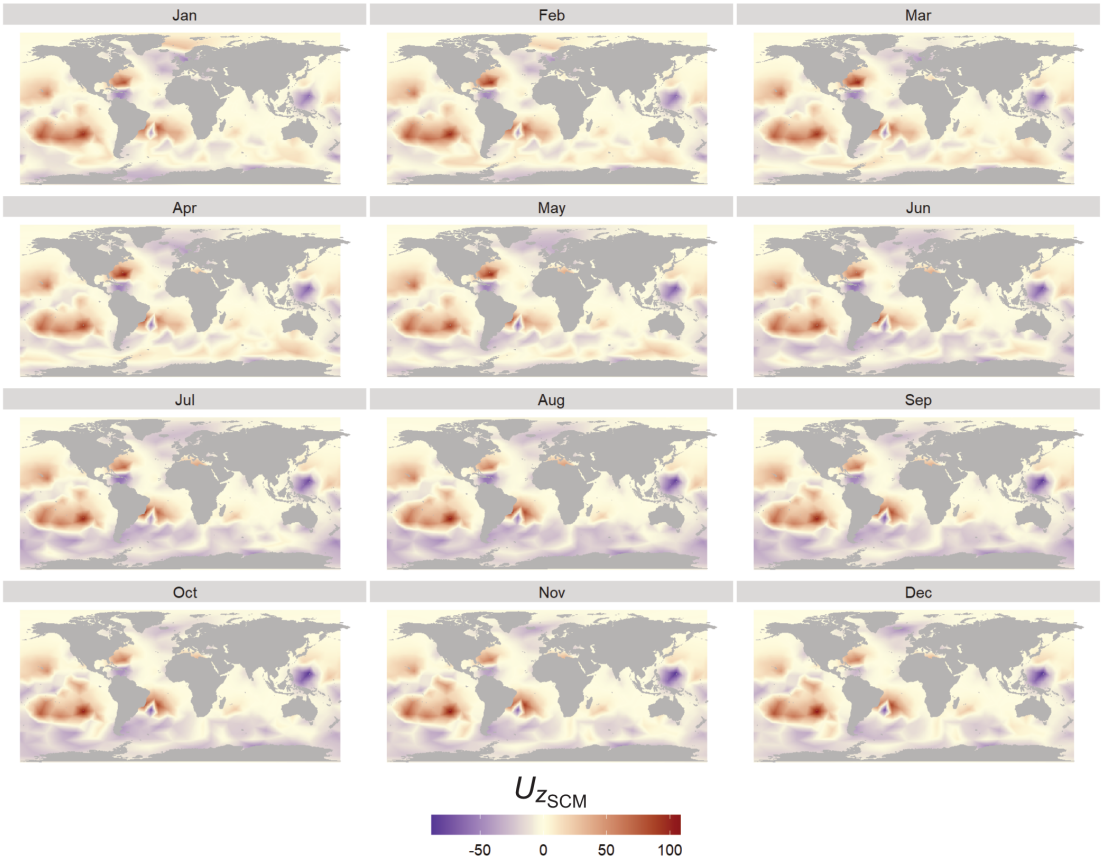

FIGURE 3.5: Monthly estimates of the latent effect included in the spatio-temporal model for z_{SCM} which accounts for variability unexplained by the fixed effects.

3.5.3 Global SCM prediction

I made predictions of SCM properties at 1120 validation locations using the spatio-temporal models. Predictions for Chl_{SCM} were generally in agreement with the true value (r^2 = 0.89) (Figure 3.6a) and the statistical coverage is 76%. The best prediction rate appeared to be for values between 0.1 and 1 mg m^{-3}, however the model did not fit so well at the extremes. Specifically, the highest and lowest values were underestimated and overestimated respectively. The predictive skill for the z_{SCM} spatio-temporal model was slightly inferior to the Chl_{SCM} model ($r^2 = 0.79$). Lower z_{SCM} observations were not well represented and almost no predictions were made for the top 15 m of the water column despite a significant number of observations being in that range. No unphysical predictions were made (negative depths) although the 95% prediction interval for several observations overlaps zero.

Global predictions of Chl_{SCM} showed strong seasonal patterns at mid to high latitudes, with the highest concentrations (> 1 mg m^{-3}) occurring during spring and summer (Figure 3.7). Very low concentrations (< 0.1 mg m^{-3}) were predicted during the polar winter and within the oligotrophic gyres. In contrast, predictions of z_{SCM} exhibited much less seasonal variation. Deep SCMs were predicted within all

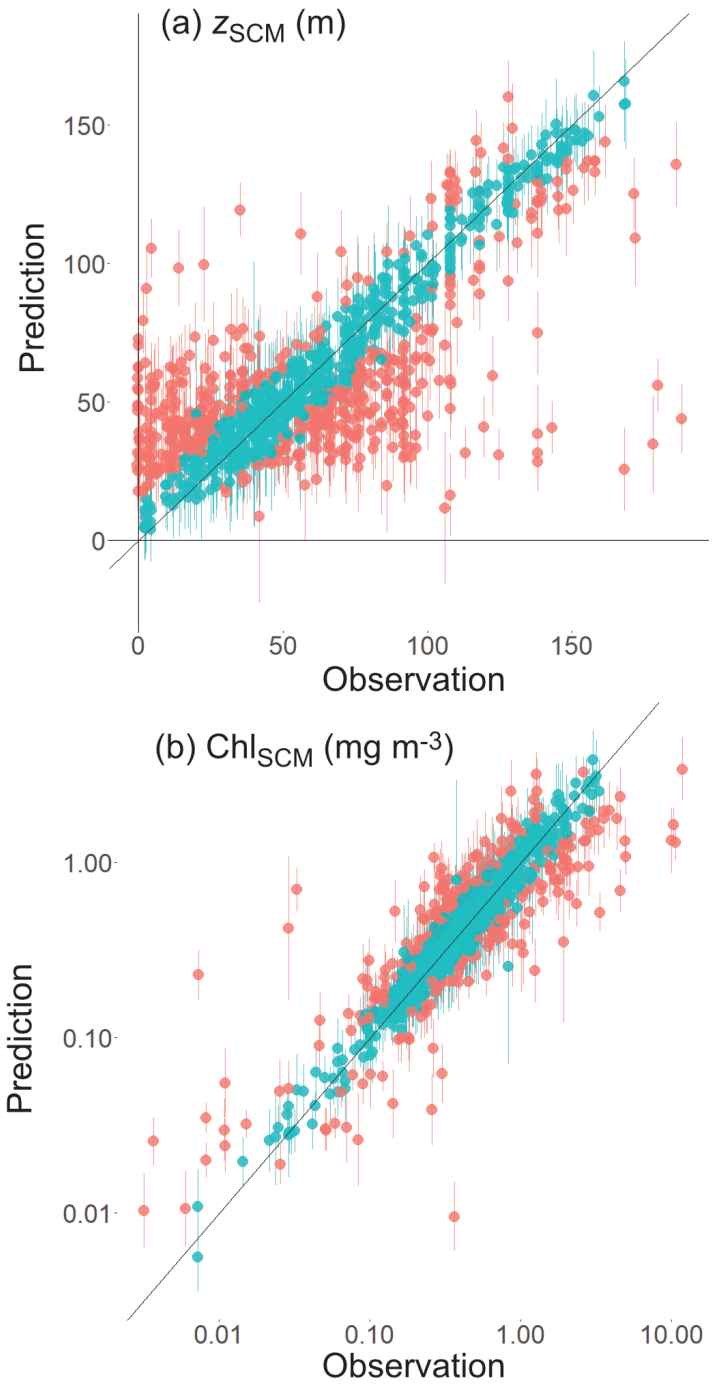

FIGURE 3.6: Comparison of observed values and fitted values (dots) with 95% prediction intervals (vertical lines) for (a) z_{SCM} and (b) Chl_{SCM}. Blue and red dots denote those whose 95% prediction interval did and did not contain the true value, respectively.

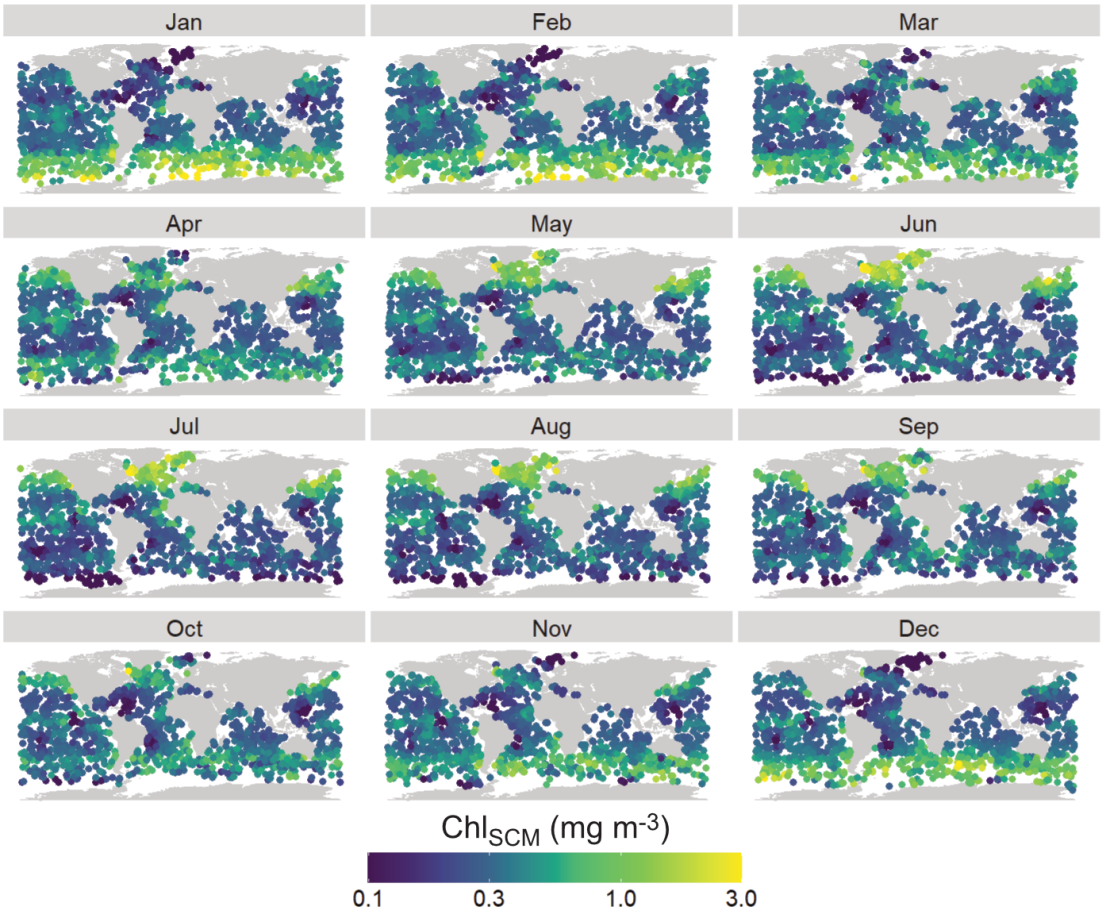

FIGURE 3.7: Monthly predictions of Chl_{SCM} at locations of 20 000 core Argo profiles in 2020 using the covariate estimates and spatial random effects from the spatio-temporal model.

subtropical gyres except the Indian Ocean, which had very few observations (Figure 3.8). In these regions, z_{SCM} typically exceeded 100 m, with the North Pacific gyre showing the greatest seasonal variability. The North Atlantic gyre appeared smaller and shifted westward, potentially reflecting data gaps near the gyre centre. Prediction uncertainty for both z_{SCM} and Chl_{SCM} increased with distance from observations (Figures A.3 and A.4).

Figure 3.9 shows combined predictions of Chl_{SCM} and z_{SCM} as a function of latitude and time. A clear seasonal migration is visible, with peak Chl_{SCM} values — and thus shallower SCMs — shifting between hemispheres. The latitudinal bands between 15° and 35° in both hemispheres correspond to the subtropical regions, which exhibited the deepest SCMs. SCMs in the Southern Hemisphere was slightly deeper on average, largely due to the deepest predictions in the South Pacific gyre. As expected, seasonal variation was minimal near the equator. A small number of unphysical predictions (43 out of 20 000) had predicted depths above the ocean surface (i.e., $z_{SCM} < 0$), which were mainly in the equatorial Atlantic near the South American coast.

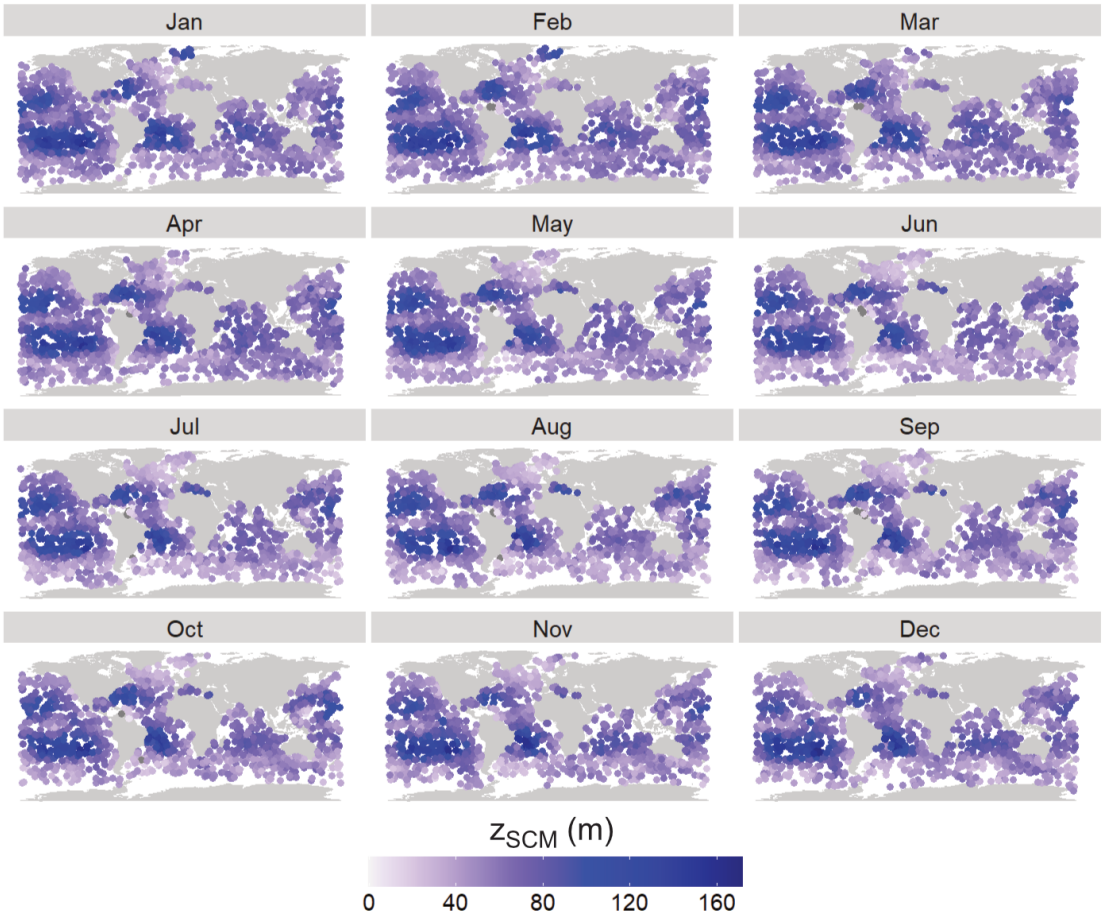

FIGURE 3.8: Monthly predictions of z_{SCM} at locations of 20 000 core Argo profiles in 2020 using the covariate estimates and spatial random effects from the spatio-temporal model.

3.6 Discussion

The vertical distribution of Chl-a in the global ocean varies across space and time (Yasunaka et al., 2021; Bock et al., 2022). In particular, the depth and intensity of SCMs are known to vary due to availability of light and nutrients (see Cullen (2015) and references therein). In this work, I applied a statistical modelling approach to Chl-a data from BGC-Argo floats to quantify the physical and biological influences on SCM characteristics and to produce global maps of these characteristics.

3.6.1 Drivers of SCMs

3.6.1.1 Physical and biological effects

I found euphotic depth to be a significant driver of SCM properties (Figure 3.3), with a deeper z_{eu} corresponding to deeper and weaker SCMs. This supports previous evidence (Xu et al., 2022b) and meets my expectations, since high Chl_{SCM} in surface

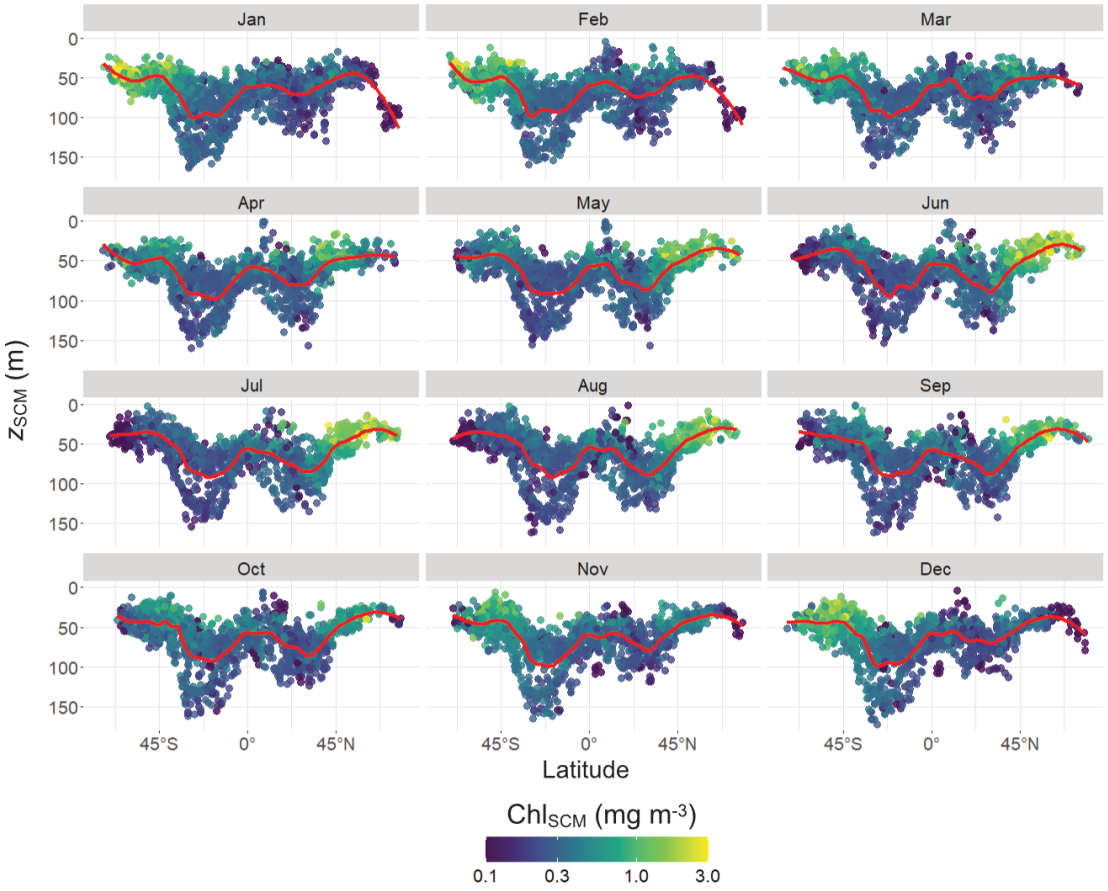

FIGURE 3.9: Monthly predictions of z_{SCM} and Chl_{SCM} using the spatio-temporal model, as functions of latitude. The red curves denote the mean z_{SCM}.

waters increases light attenuation deeper in the water column and limits growing conditions deeper in the water column. The MLD had a relatively small effect, which is surprising given that this controls the depth to which nutrients are entrained, particularly in the Southern Ocean where many of my observation were located (Xu et al., 2022a) and given that previous studies had found links between vertical chlorophyll distribution and the MLD (Carranza et al., 2018; Itoh et al., 2015; Miyares et al., 2024). Bock et al. (2022) saw that a deeper MLD resulted in shallower and more intense SCMs, especially in regions that experience seasonal variation such as the mid- and high-latitudes and the Arabian Sea. SSHA, which I used as a proxy for the presence and intensity of mesoscale eddies, had a large positive effect on z_{SCM} of approximately 40 m per metre of SSHA. This compares to Cornec et al. (2021b), who found that SCMs were around 10% deeper in the core of anticyclonic eddies and attribute this to photoacclimation. Cornec et al. (2021b) also saw that Chl_{SCM} decreased by 5% - 15% in anticyclonic eddies, however I found no significant effect from SSHA on Chl_{SCM}. The effect of SSHA might have been dampened by that of the z_{eu}, as Wang et al. (2023) identified a relationship between the two variables, with (anti-) cyclones promoting shallower (deeper) z_{eu}. Evidence suggests that surface chlorophyll concentration responds to (sub-) mesoscale ocean features like eddies and

fronts (McGillicuddy Jr, 2016; Mahadevan, 2016; Prend et al., 2022; Levy et al., 2023) so it is reasonable to suggest that SCMs are also affected by small-scale processes.

My results suggest that zooplankton biomass accumulates in shallower and more intense SCMs, contradictory to Moeller et al. (2019) who found that the presence of zooplankton near the surface could deepen SCMs (Moeller et al., 2019). Given the zooplankton data I used was integrated over the entire water column, this information may not capture the correct processes occurring across different depths. Including day length as a covariate in the models was intended to account for highly seasonal light availability near the poles, however it had a negligible effect on Chl_{SCM} and z_{SCM}. It is possible that this difference was reflected within the z_{eu} fixed effect already, or that a non-linear effect would have been more appropriate. It could have been beneficial to predict subsurface chlorophyll concentration using the surface concentration like Uitz et al. (2006) did with ship-based measurements and some machine learning studies have done with BGC-Argo float data (Sauzède et al., 2016; Chen et al., 2022). I justify neglecting this potential covariate in my models since my aim was to prioritise quantifying physical and biological drivers of SCMs rather than focussing on constructing the best predictions. The relationship between surface and subsurface chlorophyll will likely be non-linear since low surface concentrations can relate to both negligible subsurface concentrations (in the polar winters) or to low but not insignificant SCMs in the subtropics, whereas in the mid-latitudes the surface concentration could be a stronger predictor for shallower SCMs or blooms.

3.6.1.2 Spatio-temporal effects

The spatio-temporal random effects (Figures 3.4 and 3.5) highlight regions where variability remains unexplained by the fixed covariates. For Chl_{SCM}, the random effect exhibits notable seasonal structure, with strong negative values near the poles during winter. This likely reflects the failure of the fixed effects to fully capture the seasonal absence of phytoplankton, which is plausible given that such profiles represent a minority in the dataset, and the fixed-effect estimates are weighted toward observations with higher concentrations. Localised 'hotspots' in the random effects (e.g., the Labrador Sea for Chl_{SCM}) may result from spatial clustering of observations, suggesting possible local overfitting. The latent effect for z_{SCM} shows a positive effect in the subtropical gyres, indicating a contribution to deeper SCMs. This may be explained by the fact that the dataset from which I obtained the euphotic depth covariate was limited to around 110 m, which would have underestimated the deepest SCMs without the inclusion of the spatio-temporal random effect.

3.6.2 Predictions of SCM characteristics on a global scale

The spatio-temporal model for Chl_{SCM} fitted the observations well (76% statistical coverage, $r^2 = 0.89$) (Figure 3.6), but it did not fit so well for z_{SCM} (52% statistical coverage, $r^2 = 0.79$). SCM intensities between 0.1 and 1 mg m^{-3} were generally well-reproduced, however the highest values were underestimated. These high concentrations are likely triggered by some factor not included in my analysis and on too small a scale to be detected by the spatial effect. Several of the highest values are located in the Southern Ocean which is known to be iron-limited (Hawco et al., 2021). Therefore, it is possible that these blooms have been induced by a local-scale iron supply such as sea ice melt (Behera et al., 2020; Baldry et al., 2020), or sediment from islands (Robinson et al., 2016).

Overall, my model predictions generally agree with the analyses by Cornec et al. (2021a) and Yasunaka et al. (2021), with both SCM characteristics largely dependent on latitude (Figure 3.9). The seasonality of Chl_{SCM} increases towards the poles with highest concentrations during spring and summer in the mid- and high latitudes. The deepest and least intense SCMs are found in the oligotrophic gyres and shallowest are located near continental boundaries and at latitudes greater than 40° in both hemispheres. I found that the SCMs in the North Pacific subtropical gyre are deeper during summer, possibly due to changes in the nitracline depth (Letelier et al., 2004). Spatial or temporal differences in z_{SCM} may indicate variations in phytoplankton community composition, as seen by Sato et al. (2022) and Brewin et al. (2022), however this cannot be determined from my analysis.

3.6.3 Wider implications for SCMs

One of the advantages of this approach is that I was able to interpolate SCM properties across the global ocean. Such information is of use for identifying regions where satellite ocean colour data may be missing for significant subsurface chlorophyll concentration and validating biogeochemical model output (Mignot et al., 2021). If repeated for the entire BGC-Argo dataset, which extends from 2015 to 2025, interannual variability could be identified. My study expands on work by Cornec et al. (2021a) and Yasunaka et al. (2021) in assessing global SCM patterns and drivers but extends their work by interpolating between floats and by quantifying the impact of multiple influences concurrently within a statistical framework. The natural extension is to apply this methodology to the datasets of the aforementioned studies, which collated profiles over a longer time period and acquired concurrent profile covariates, such as nitrate or downwelling irradiance, or supplemented the Argo profiles with observations from other sources such as gliders (Carvalho et al., 2020). My spatio-temporal modelling approach shows the increased uncertainty generated

by more sparsely distributed floats (Figures A.3 and A.4), highlighting the importance of increasing the quantity of floats and improving their spatial coverage across the global ocean.

3.6.4 Spatio-temporal modelling of global BGC-Argo data

In this work, I applied a novel statistical modelling approach to data from BGC-Argo floats. I used the INLA-SPDE method developed by Lindgren et al. (2011) which provided computational benefits and allowed for describing a non-stationary correlation structure. More specifically, I included a physical barrier constraint to dampen the autocorrelation between locations separated by land. SPDE spatio-temporal models have previously been used for global datasets (Dahlén et al. (2020), however to my knowledge this work is the first using marine biogeochemical data on a global scale, where including land barriers might be important. My results suggest that spatio-temporal models are appropriate for research regarding large-scale biogeochemical phenomena like SCMs (Table 3.2). The magnitude (and sign) of some covariate fixed effects were significantly different when a spatio-temporal random effect was included, as in Willems et al. (2022). This work only investigated using fixed effects for covariates, which assumes that the effects are constant in space and time. In practice, SCMs can form due to a variety of environmental factors (Baldry et al., 2020) and these can vary across locations and seasons. Including spatially varying covariates (SVCs) could be more realistic although this would significantly increase the computational expense of the model fitting process. SVCs have not been used often within the R-INLA functionality, however alternative R packages do offer such a feature such as **sdmTMB** (Anderson et al., 2022). Similarly, the significance of some covariates may have been different if smooth non linear effects were used instead of fixed effects, particularly for the zooplankton biomass and day length covariates.

3.6.5 Limitations and future work

The biggest limitation of this work was the estimation of z_{SCM}. Previous studies have used a variety of methods to define and compute the SCM properties from Chl-a profiles (Carranza et al., 2018; Sato et al., 2022; Xu et al., 2022b), although it was difficult to make a rigorous definition for SCMs whilst avoiding computational issues. Many profiles in my dataset did not display a single clear peak, but instead had either no peak, a non-Gaussian shape such as a sigmoid, or even multiple peaks. Consequently, my method for defining SCM characteristics was likely sub-optimal. Another potential drawback is that I did not include the nitracline depth or absolute light intensity as covariates, which in some regions may be as important (or more important according to some studies) than the z_{eu} in determining z_{SCM} (Herbland and

Voituriez, 1979; Gong et al., 2017; Miyares et al., 2024), since it controls the availability of nutrients in the upper ocean.

My spatio-temporal modelling approach considered data that comprised a single value for each observation. This meant that the SCM properties needed to be estimated prior to model fitting. In particular, estimating z_{SCM} poorly may have introduced some added bias into my results. One could address this issue by using a statistical approach called functional data analysis (FDA) that views each profile as a single observation. Yarger et al. (2022) recently developed a functional data methodology for Argo float data to interpolate temperature and salinity profiles across the global ocean. Not only would using this method utilise all measurements within a profile, but it would avoid estimating the SCM properties so a wide range of profile shapes could be included without assuming they have clear peaks. Perhaps the bias of sampling location affected model fitting, especially when identifying spatial patterns. I explore the application of FDA techniques on Argo float profiles in Chapter 4 and Chapter 5.

I make several recommendations following this work: (1) Use spatio-temporal modelling when using data from BGC-Argo profiles across a large area and spanning at least several months. Including a spatio-temporal latent effect accounts for unobserved effects as well as autocorrelation between neighbouring observations in space and time. Furthermore, I suggest considering a model with spatially varying covariates because the limiting factors of phytoplankton growth vary with location and season, although I acknowledge that this requires further computational power. (2) Use chlorophyll data from multiple years; the BGC-Argo array is unevenly distributed across the global ocean so increasing the span of observations could improve model fit especially in under sampled areas. This might aid in distinguishing between interannual variability and more permanent spatial differences over smaller scales. (3) If the aim of the work is to produce maps, then I recommend using covariates from spatially complete datasets. This would allow a gridded product to be produced without further interpolation of the model predictions. (4) Using a functional data approach such as Yarger et al. (2022) would be beneficial for several reasons. First, all measurements from a profile are included in the model-fitting process so the need to define and estimate the SCM is avoided. Second, this method allows for a variety of profile shapes, and it accounts for the fact that many profiles do not have a single, clear peak but rather multiple peaks or no peak at all. (5) Repeating the work using both chlorophyll and backscatter data could help address questions regarding the role of photoacclimation in SCM formation on a global scale (Cornec et al., 2021a; Masuda et al., 2021).

3.7 Conclusion

This research aimed to use chlorophyll data from BGC-Argo floats to quantify the effects of different environmental factors on the vertical distribution of chlorophyll throughout the global ocean during 2020. In particular, I investigated how the intensity and depth of SCMs vary over space and time using a spatio-temporal modelling approach. My approach was computationally efficient as it used the SPDE-INLA technique developed by Lindgren et al. (2011) and accounted for reduced spatial autocorrelation when two locations were separated by land. I found that a combination of biological and physical factors affected both the intensity and the depth of SCMs, with the z_{eu} having a significant influence on both SCM properties. MLD had a negligible effect on both characteristics and SSHA was not found to be a significant driver for Chl_{SCM}, in contrast to previous research (Cornec et al., 2021b), although it was for z_{SCM}. Fitted models were used to make predictions of SCM properties across the global ocean using data from core Argo floats. This revealed global patterns seen by previous studies (Cornec et al., 2021a; Yasunaka et al., 2021), with the deepest and least intense SCMs occurring in the oligotrophic subtropical gyres and the most intense occurring closer to the surface, especially at higher latitudes. My approach worked well for modelling SCM intensity, but less well for SCM depth. This was likely due to many profiles not having a clearly defined peak. Future statistical models of vertical chlorophyll distribution could benefit from using functional data analysis. Such an approach could use all measurements in a profile, eliminating the need to identify SCMs and accounting for a variety of profile shapes, including those that do not contain a single, distinct peak.

Chapter 4

Assessing Environmental Influences on Subsurface Chlorophyll Maxima with Functional Regression Models

This chapter is, at the time of writing, in preparation for publication in *Journal of Geophysical Research: Oceans* as:

Taylor, M., Cornec, M., Henson, S., Sahu, S., Hammond, M., Cael, B.B. Assessing Environmental Influences on Subsurface Chlorophyll Maxima with Functional Regression Models.

4.1 Abstract

Subsurface chlorophyll maxima (SCMs) are a common phenomenon in the global ocean, characterised by a subsurface peak of chlorophyll concentration, reflecting active phytoplankton layers well below the ocean surface. SCMs form when nutrients and light are limited from above and below respectively. It is well established that the vertical distribution of chlorophyll, including the depth of SCMs, varies spatially and temporally, due to changing environmental conditions. In this study, I used profiling data from 26 biogeochemical-Argo floats to identify relationships between environmental conditions and SCMs in tropical and subtropical ocean regions. I utilised functional regression, a statistical technique in which each profile can be considered as one datum, in order to identify relationships between profile shape and environmental conditions. I fitted functional regression models to profiles of chlorophyll and particle backscatter (b_{bp}), a proxy for particulate organic carbon. I found that bio-optical profile shape was better reproduced using an additive model with non-linear scalar effects than with linear scalar effects. My results suggest that

the SCM closely follows the euphotic depth, whereas the peak b_{bp} is located at or above the nitracline. Predictions chlorophyll and b_{bp} profiles were made over the permanently stratified regions of the global ocean. This revealed distinct latitudinal bands where the ordering of the euphotic depth and nitracline depth determined the depth and intensity of the SCM. I also found evidence suggesting that photoacclimation is a dominant characteristic of SCMs in both subtropical and equatorial regions.

4.2 Introduction

Chlorophyll-a is the primary pigment in photosynthetic organisms, and its concentration is widely used as a proxy for phytoplankton biomass in marine ecosystems. Chlorophyll concentration in the ocean varies by several orders of magnitude over time, geographic space, and depth. One common feature of the vertical distribution is a local maximum concentration located significantly below the surface due to limitations of nutrients and light from above and below, respectively (Cullen, 2015). This phenomenon is termed a subsurface chlorophyll maximum (SCM) and can be located as deep as 200 m below the surface (Mignot et al., 2014). SCMs can form through two mechanisms: biomass accumulation and photoacclimation. The former refers to cases where an actual increase in phytoplankton biomass is observed at depth (more precisely an increase in phytoplankton carbon, denoted C_{phyto}) (Beckmann and Hense, 2007; Herbland and Voituriez, 1979; Hodges and Rudnick, 2004), whereas the latter describes a physiological response of phytoplankton to low light levels by increasing the amount of intracellular chlorophyll per unit biomass (Fennel and Boss, 2003; Letelier et al., 2004; Masuda et al., 2021). The cases have been termed subsurface biomass maxima (SBMs) and subsurface photoacclimation maxima (SAMs) respectively. Over longer time periods, adaption to low-light may also be a driver, rather than acclimation SCMs are ubiquitous features across the tropics and subtropics, where nearly-permanently stable environmental conditions allow for their formation and maintenance, particularly in oligotrophic regions where surface chlorophyll concentration is low (Uitz et al., 2006). It has been found that DCMs contribute around half of depth-integrated NPP using ocean models (Silsbe and Malkin, 2016) and BGC-Argo float data (Vives et al., 2024a), although this fraction varies regionally and seasonally. Consequently SCMs play an important role in the global carbon cycle and marine ecosystem.

Due to their depth, SCMs cannot be directly observed by satellite and their monitoring requires the use of ship-based sampling or autonomous platforms. The biogeochemical-Argo float array is a global network of robotic profiling floats equipped with bio-optical sensors capable of measuring a range of physical and biogeochemical variables throughout the top 2000 m of the water column, typically

every ten days (Claustre et al., 2020). Many floats carry bio-optical sensors for measuring chlorophyll fluorescence, from which chlorophyll concentration can be estimated (Roesler et al., 2017), and particle backscatter (b_{bp}), a proxy for particulate organic carbon (POC) (Loisel and Morel, 1998; Cetinić et al., 2012). Coincident chlorophyll and b_{bp} profiles from BGC-Argo floats have been used to describe the global distribution and classification of SCMs between SBMs and SAMs (Cornec et al., 2021a) as well as identifying distinct biogeographical regimes based on seasonal variability of the two bio-optical parameters (Bock et al., 2022). Deeper SCMs are typically found in the subtropical oligotrophic gyres (Cornec et al., 2021a; Yasunaka et al., 2021) and have a lower maximum chlorophyll concentration and are thicker (Uitz et al., 2006) and are more likely to be formed through photoacclimation rather than biomass accumulation (Cornec et al., 2021a).

It is well-established that SCMs form in strongly stratified water columns (Beckmann and Hense, 2007; Cullen, 2015; Garg et al., 2024), where the upper layer is nutrient-depleted and stable conditions allow for the persistence of these features. The depth of SCMs (z_{SCM}) has been shown to be associated with a range of water column features including the euphotic depth (z_{eu}) (Agustí and Duarte, 1999; Gong et al., 2015; Xu et al., 2022b; Garg et al., 2024; Miyares et al., 2024), the nitracline depth (z_{ncline}) (Herbland and Voituriez, 1979; Richardson and Bendtsen, 2019; Garg et al., 2024; Miyares et al., 2024), isopycnals (Xu et al., 2022b) and isotherms (Chowdhury et al., 2021). Moreover, SCM intensity ($\mathrm{Chl}_{\mathrm{SCM}}$), defined as the chlorophyll concentration at the SCM peak, is typically lower for deeper euphotic depths (Xu et al., 2022b), whereas Gong et al. (2015) showed that the nitracline gradient determined the $\mathrm{Chl}_{\mathrm{SCM}}$. Seasonal variability in light intensity in subtropical regions can shift the z_{SCM} by several tens of metres (Letelier et al., 2004) and influence phytoplankton community composition, with deeper SCMs favouring smaller species due to their ability to grow in low light levels (Latasa et al., 2016; Garg et al., 2024). Bock et al. (2022) compared chlorophyll and b_{bp} seasonal cycles on a global scale to the z_{eu}, the z_{ncline} and the mixed layer depth (MLD). They found that in tropical regions the SCM was located around the nitracline and above the z_{eu}, whereas in subtropical regions the z_{eu} and z_{ncline} were much more similar and were both located deeper than the SCM. In seasonally stratified regions SCMs only formed during summer in response to the substantial variability in the MLD, in contrast to the z_{eu} and z_{ncline} remaining nearly constant year-round (Bock et al., 2022).

Accurately identifying and quantifying the mechanisms behind the formation and maintenance of SCMs is important in order to fully understand how they vary in space and time, what their global significance to the marine carbon cycle is and anticipate how that might be impacted by anthropogenic climate change. However, uncertainty remains over precisely how the influences of light and nutrients affect the shape of chlorophyll and b_{bp} profiles pertaining to SCMs. Previous modelling studies assessed

how each limiting factor affects the profiles of phytoplankton carbon and chlorophyll respectively, which shed some light on the importance of photoacclimation (Fennel and Boss, 2003; Masuda et al., 2021). Several attempts have been made to fit mathematical curves such as sigmoids and Gaussians to chlorophyll profiles (Gong et al., 2015; Carranza et al., 2018; Xu et al., 2022b; Brewin et al., 2022). This approach reduces the complexity of entire profiles into a few well understood parameters, allowing for simple analysis of how profile characteristics (such as the z_{SCM}) vary under different circumstances. Alternatively, profiles have been clustered based on their shape before analysis (Cornec et al., 2021a). In recent years, neural networks have become increasingly popular when reconstructing bio-optical profiles (Sauzède et al., 2015; Chen et al., 2022; Yu et al., 2024). Although such methods provide an excellent opportunity for spatio-temporal interpolation at unsampled locations, they do not explicitly identify the relationships between entire profiles and environmental conditions. Consequently, a gap in the literature persists for a statistical approach which identifies the effects of environmental conditions on bio-optical profiles without prior fitting of profiles to mathematical curves or extraction of profile characteristics.

In this work, I aim to use profiling data from BGC-Argo floats to better understand drivers of variability in chlorophyll vertical distribution through the use of functional data analysis (FDA). FDA is a branch of statistics focussed on data that take the form of curves or surfaces, where the variable of interest is a function of at least one other variable. A well-established and growing literature exists for FDA, encompassing standard statistical techniques such as regression, clustering, and principal component analysis (Ramsay and Silverman, 2005; Wang et al., 2016) as well as more advanced topics like geostatistics (Mateu and Giraldo, 2021). FDA methods have recently been applied to the Argo float profiling data, with Yarger et al. (2022) and Korte-Stapff et al. (2022) modelling temperature and salinity, and oxygen, respectively, all as functions of pressure. These studies showed the utility of functional data representations for oceanographic variables, especially in understanding the dependencies between measurements across depth. I apply a similar treatment to profiles from BGC-Argo profiling data in the top 250 m of the water column to assess causes of variability in shape amongst bio-optical profiles. This work highlights the opportunity to analyse the variability of entire profiles through FDA rather than focussing on profile characteristics, whilst avoiding predefining possible theoretical curves and thus allowing the data to speak for itself.

4.3 Materials and methods

4.3.1 Study region

I used profiling data from BGC-Argo floats located in low latitude biomes, where SCMs are reported as dominant features of chlorophyll profiles (Cullen, 2015; Cornec et al., 2021a). Profiles were selected if they were located within either the subtropical or equatorial biomes defined by Fay and McKinley (2014), both being nearly-permanently stratified systems. I use the mean biomes of Fay and McKinley (2014), which are defined by clustering locations into contiguous regions, based on similarities in surface chlorophyll concentration, sea surface temperature (SST), sea ice coverage and MLD. Any profile that was not assigned a biome but was located nearest to one of my desired biomes, and was within 500 km of it, was also included in my dataset. This was done for two reasons, the first being to increase the number of profiles assigned to biomes, and the second being to slightly widen the range of environmental conditions, which might aid the identification of relationships between covariates and response variables. I only used profiles with measurements of chlorophyll, b_{bp} (a measure of suspended particles and a proxy for particulate organic carbon (Cetinić et al., 2012; Loisel and Morel, 1998)), nitrate and photosynthetically available radiation (PAR) alongside the standard CTD (conductivity, temperature and depth) measurements. Furthermore, I restricted myself to only using profiles completed at most two hours either side of local noon. This was done so that PAR profiles were not affected significantly by diurnal light variations. In total, 1323 profiles from 26 BGC-Argo floats met these criteria, collectively spanning a period from 24/10/2012 to 05/11/2023, with each float completing a profile approximately once every 10 days. The locations of profiles used in this analysis are shown in Figure 4.1.

4.3.2 Bio-optical profiles

Several quality control procedures were applied to the chlorophyll and b_{bp} profiles to remove bad data and prepare the profile for modelling, as detailed in Cornec et al. (2021a). Measurements with an Argo quality control flag of 3 (“Probably bad”) or 4 (”Bad”) were removed. I only used measurements from the top 250 m of the water column, as chlorophyll is typically negligible below that depth, and consequently is not of interest in the study of SCMs. BGC-Argo floats do not all sample the water column at the same depths so first I regridded and interpolated the chlorophyll and b_{bp} profiles to 1 m intervals. Next, I applied a $\log_{10}$ transformation to the chlorophyll data given that chlorophyll has an approximately log-normal distribution (Campbell, 1995). This helped reduce the magnitude of spikes in the profiles and ensured that no unphysical predictions (negative chlorophyll concentrations) were made. Finally I

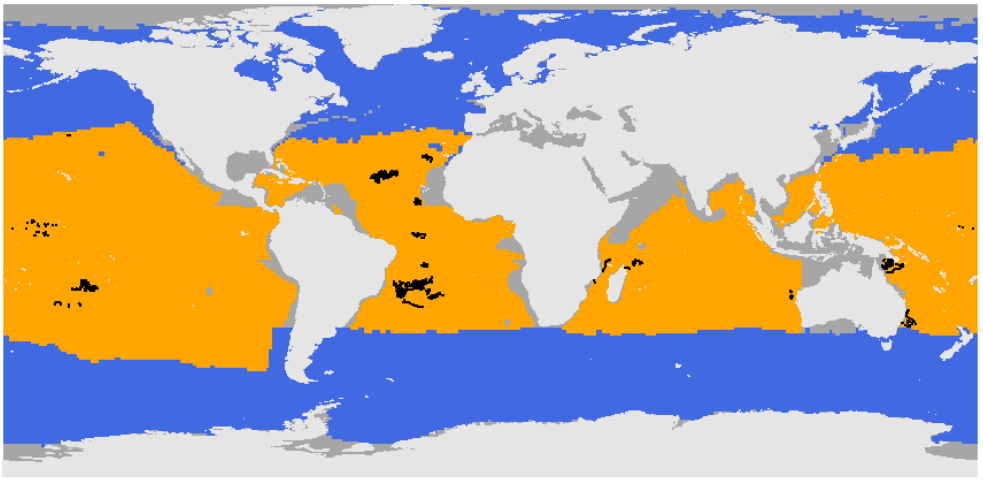

FIGURE 4.1: My study region was the combined area of subtropical permanently stratified biome and the equatorial biome defined by Fay and McKinley (2014) (shown in orange). Other biomes are shown in blue and regions in grey denote cases where no biome could be assigned. Within this region, I used data from BGC-Argo floats which had profiles of chlorophyll, b_{bp}, potential density (derived from temperature, salinity and pressure), nitrate and PAR. My dataset comprised 1323 profiles (black dots) from 26 BGC-Argo floats. Specifically, I used 423 profiles in the Pacific Ocean, 850 profiles in the Atlantic Ocean and 50 profiles in the Indian Ocean.

smoothed the chlorophyll and b_{bp} profiles to reduce the relative magnitude of spikes using a running median over a window of 10 m. Note that throughout this work when I refer to optical backscatter, I assume measurements at a wavelength of 700 nm, unless clearly stated otherwise.

4.3.3 Covariate quality control and preparation

Following Argo protocols, hydrological data collected by the SBE 41 seabird CTD sensors were processed and quality-controlled as described by Wong et al. (2020). The MLD was estimated as the minimum depth at which the potential density increased by at least 0.03 kg m^{-3} relative to its value at 10 m (de Boyer Montégut et al., 2004). The z_{eu} was defined by the minimum depth where PAR was smaller than 1% of its median value in the top 15 m of the water column. This was done to remove variability in near-surface measurements caused by waves. The z_{ncline} was defined by a 1 μmol L^{-1} threshold above the surface nitrate concentration as in Cornec et al. (2021a).

4.3.4 Analysis

Model structure I used a method from a branch of statistics called functional data analysis (FDA), where each observation is a continuous curve or surface with respect to some other variable (Ramsay and Silverman, 2005). In practice, functional data observations are a finite set of measurements and, as the gaps between measurements decrease, the approximation to a continuous function improves. In many applications, the indexing variable is time, but in an oceanographic context, it is natural to view profiling data as functions of pressure. Functional regression models (FRMs) aim to infer relationships between variables, where at least one of which is a functional variable. Refer to Table B.1 for a comparison of concepts in functional regression and their analogues in scalar-valued regression. In this work, I treated profiles from BGC-Argo floats as functions of pressure similar to Yarger et al. (2022) and Korte-Stapff et al. (2022). The following equation shows an example model formula which includes a linear effect $\beta_1(p)$ and a non-linear effect $f(z_2, p)$,

$$y(p) = \mu(p) + z_1\beta_1(p) + f(z_2, p) + \cdots + \epsilon(p) \tag{4.1}$$

where the response variable $y(p)$ is a function of p, $\mu(p)$ is the intercept function, z_1 and z_2 are scalar covariates and $\epsilon_i(p)$ is a normally distributed functional error term. The dots in Equation 4.1 signify that there could be additional covariates (of either type shown in Equation 4.1). Note that in this work I ignore functional covariates since they add considerable complexity to the model and could make results more difficult to interpret from a mechanistic perspective.

I fitted two FRMs to bio-optical profiles to compare the predictive ability of models using linear scalar covariates or non-linear scalar covariates to compare their effectiveness. The scalar covariates I used were MLD, z_{eu}, and z_{ncline}, where each of these was derived from a profile, namely potential density, PAR or nitrate respectively. The mean functions of the response variable $y(p)$ of specific models I fitted are shown below.

$$y(p) = \mu(p) + \text{MLD}\beta_1(p) + z_{\text{eu}}\beta_2(p) + z_{\text{ncline}}\beta_3(p) + \epsilon(p) \tag{4.2}$$

$$y(p) = \mu(p) + f_1(\text{MLD}, p) + f_2(z_{\text{eu}}, p) + f_3(z_{\text{ncline}}, p) + \epsilon(p) \tag{4.3}$$

Equations 4.2 and 4.3 will be referred to from here on as the linear model and the non-linear model respectively. Each of the model structures was fitted using chlorophyll and b_{bp} profiles as the response variable.

Computation In practice, the functional coefficient parts of the model are described by basis functions. In this work, I use cubic splines for functional coefficients, given they guarantee a smooth function whilst combining computational efficiency and

model flexibility. The splines had knots at regular 5 m intervals vertically in the water column. The models were fitted using the **pffr** function within the **refund** package in R (R Core Team, 2023). Model parameters were estimated through restricted maximum likelihood (REML), in which redundant parameters are removed by applying a transformation to the dataset, thereby reducing the computational complexity. After fitting the models, I used the **maxd** function from the **castr** R package to identify the z_{SCM} of observed and fitted chlorophyll profiles from each model. I also retrieved the chlorophyll concentration at the SCM (i.e., Chl_{SCM}) to compare the different models.

4.3.5 Prediction

I expanded the method to a wider spatio-temporal scales (with greater coverage) by leveraging time and depth resolved gridded products and the output from the best fitting model (the non-linear scalar model). Gridded products for each covariate were obtained at a 1° spatial resolution and a 1-month temporal resolution. The MLD field was sourced from the Global Ocean Surface Mixed Layer (GOSML) monthly climatology, with estimates derived from Argo float CTD profiles. Nitrate profiles were estimated for each grid cell using the CANYON-B machine learning model (Bittig et al., 2018; Sauzède et al., 2017), which relied on CTD and oxygen profiles collected by BGC-Argo floats, as oxygen sensors are more widely available than nitrate sensors. Using these profiles, the same method was applied to estimate nitracline depth. These three covariates at each grid cell were then used to predict chlorophyll and backscatter profiles from 5 m to 250 m at a 1 m vertical resolution.

The fitted b_{bp} profiles were used to derive the phytoplankton carbon mass (C_{phyto}) through the same approach as Estapa et al. (2019). This first involved estimating the backscatter at 470 nm from the backscatter at 700 nm using the power law found by Morel and Maritorena (2001)

$$b_{bp}(470) = b_{bp}(700)\left(\frac{470}{700}\right)^{-1}. \tag{4.4}$$

Note that estimates of the exponent in Equation 4.4 vary significantly. Despite this, the analysis is continued to provide an indication of the mechanisms that affect C_{phyto}. The conversion developed by Graff et al. (2015) was then used to estimate the C_{phyto} concentration which took the following form

$$\text{C}_{\text{phyto}} = 12128 \times b_{bp}(470) + 0.59. \tag{4.5}$$

4.4 Results

4.4.1 Evaluation of FRMs for bio-optical profiles

Both model structures reproduced chlorophyll profiles well when z_{SCM} was located between 110 m and 150 m (e.g., Figures 4.2a-c), a range which contains 56% of profiles. However, the concentration of shallower SCMs was not characterised quite as well as deeper SCMs by either model (Figures 4.2d-e). The thickness of SCMs in predicted profiles was overestimated for observations with narrower peaks (Figure 4.2e). Some unusual features from my dataset, such as chlorophyll profiles multiple peaks were not captured by either of the FRMs (e.g. Figure 4.2f). However, those chlorophyll profiles formed a small minority ($< 1\%$) of my dataset, so this is unlikely to be reflected in predictions. The linear model did not effectively capture the variability throughout the top 250 m of the water column (Figure 4.3a). Its fitted profiles appeared to give the same prediction at a depth of approximately 110 m, suggesting there was no variability. The relationship between z_{SCM} and Chl_{SCM} in the observed profiles was better reproduced by the non-linear model (Figure 4.3b), with deeper peaks having a smaller concentration until a depth of 100 m. In contrast the predicted profiles from the linear model indicated that, below a depth of 115 m, the Chl_{SCM} increases with z_{SCM}, which disagrees with both the observations and previous research. The lower AIC and RMSE of the non-linear model suggested it explains a significantly greater amount of variability in chlorophyll profiles, without including unreasonably many parameters (Table 4.1). Consequently, the non-linear model was considered to fit chlorophyll profiles better than the linear model and was used in later analyses.

The mean fitted b_{bp} profile of each model closely resembled the mean observed b_{bp} profile (Figure 4.4), but the variability differed, with the non-linear model displaying slightly more variability around the mean, which was more similar to the observed profiles. Both models identified that variability was highest in the top 50 m. The linear model produced predictions with very little variability at a depth of around 115 m, possibly reflecting a situation where the linear effects are changing sign. An inspection of the characteristics of b_{bp} peaks (depth and concentration) did not aid the comparison between models as both models gave similar results. The AIC and RMSE were again smaller for the non-linear model (Table 4.1), indicating that the non-linear model better explained the variability in b_{bp} profiles.

4.4.2 Effects of environmental conditions on chlorophyll and b_{bp} profiles

I explored the relationships between the covariates and the bio-optical profiles identified by the non-linear scalar model (Figure 4.5). I found that the z_{SCM} was

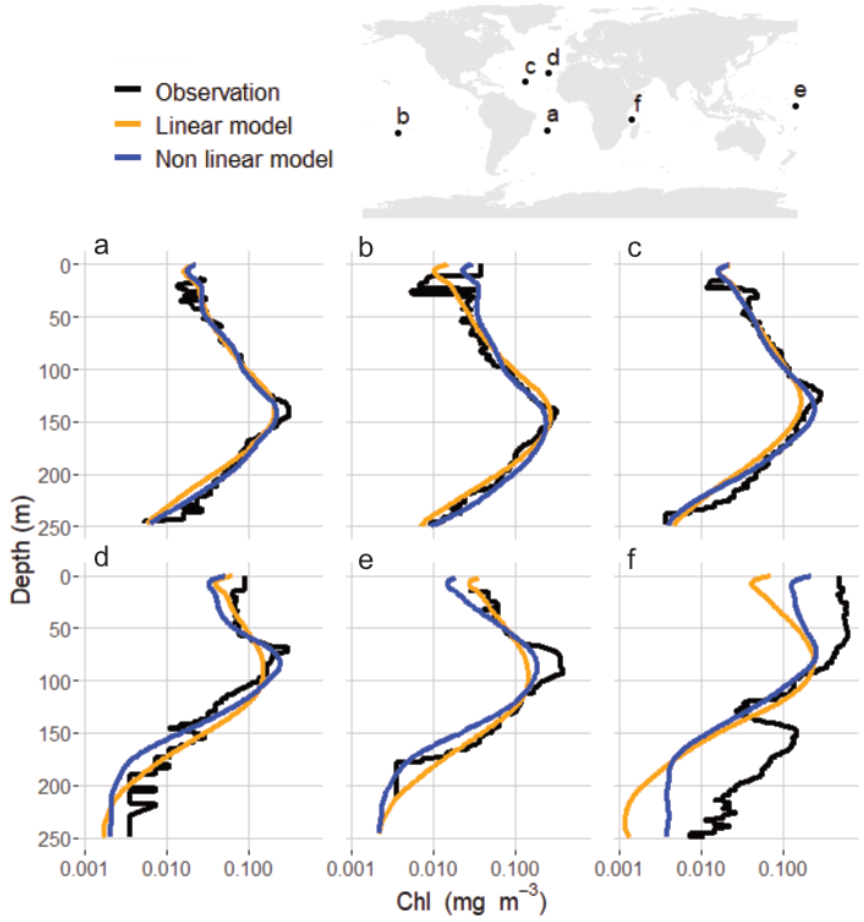

FIGURE 4.2: Six examples of chlorophyll profile observations (black curves) and fitted profiles from the two models. (a-c) show deeper SCMs. (d-e) show slightly shallower SCMs. (f) shows a more complicated profile shape with two peaks in chlorophyll.

Response	FRM type	RMSE	AIC	Explained Deviance (%)
Chlorophyll	Linear	0.448	468768.2	57.0
	Non-linear	**0.407**	**396362.8**	**64.5**
b_{bp}	Linear	1.217×10^{-4}	-5568338	48.8
	Non-linear	$\mathbf{1.161}\times10^{-4}$	**-5596971**	**53.4**

TABLE 4.1: A comparison of model performance using three different measures of goodness-of-fit. The best fitting model type according to each measure is shown in bold.

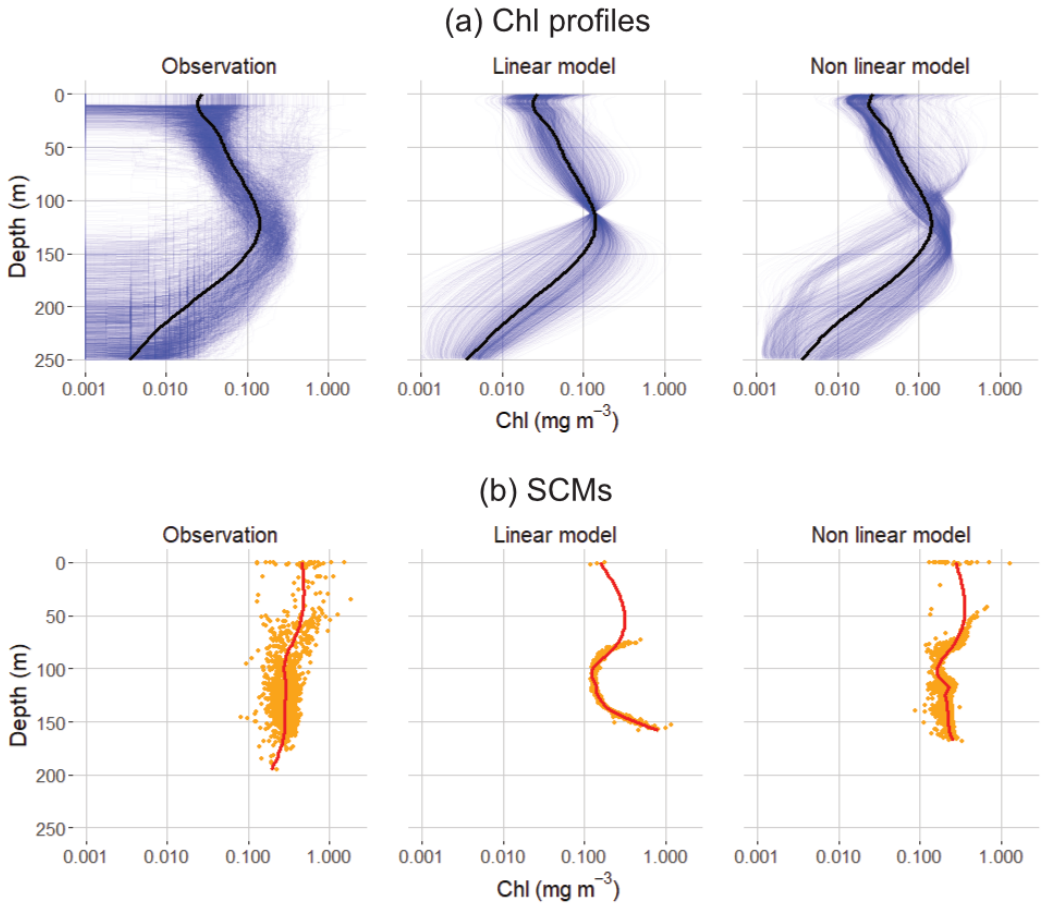

FIGURE 4.3: (a) Comparison of variability among all observed chlorophyll profiles and the fitted profiles from the linear model and the non-linear model, respectively. The black curve in each panel represents the mean chlorophyll profile. (b) The orange dots display the z_{SCM} and Chl_{SCM} of each profile. The red line shows the relationship between the depth and concentration of the SCM across models and the observations.

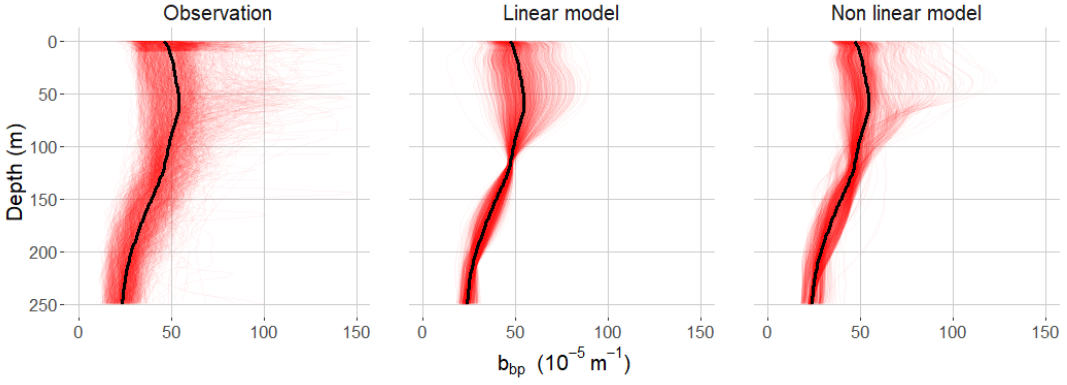

FIGURE 4.4: Comparison of variability among all observed b_{bp} profiles and the fitted profiles from the linear model and non-linear model, respectively. The black curve in each panel indicates the mean profile.

located at the euphotic depth, which ranged from 50 m to 160 m, with the SCM peak spanning around 20 m either side of the z_{eu}. High concentrations of b_{bp} were restricted to above the z_{eu}, especially for $z_{SCM} < 100$ m, where there was a sharp decrease at the z_{eu}. SCMs were typically found below shallow (< 90 m) nitraclines and above deep ($>$ 140 m) nitraclines. The peak in biomass often occurred at the nitracline (for $z_{ncline} < 90$ m). Deeper than this, the biomass peak was considerably thicker (up to 100 m), but of a lower intensity. Biomass concentrations were significantly lower at depths more than 20 m below the nitracline compared to the same distance above it, whereas chlorophyll exhibited a more symmetric peak on both sides of the nitracline. Although the effects of MLD on chlorophyll and b_{bp} were statistically significant, their magnitudes were substantially smaller than that of the other covariates. The intercept functions for the non-linear model of chlorophyll and b_{bp} are shown in Figures B.1a and B.1b respectively. The non-linear effects for chlorophyll and b_{bp} are shown in Figures B.2 and B.3 respectively. The standard errors of the non-linear effects for chlorophyll and b_{bp} are shown in Figures B.4 and B.5 respectively. Summaries of the non-linear model effects for chlorophyll and b_{bp} are given in Tables B.2 and B.3 respectively.

4.4.3 Predictions of SCM characteristics

Predictions of z_{SCM} and Chl_{SCM} exhibited notable spatio-temporal variability (Figure 4.6). The deepest SCMs, typically around 120 m and reaching depths of up to 157 m in the South Pacific, were predicted in the central regions of the subtropical gyres with a chlorophyll concentration of around 0.2 mg m^{-3}. In these locations, the deepest SCMs were predicted during summer, with a seasonal range of approximately 15 m. Shallower SCMs (around 60 m) were predicted in upwelling regions along the equator, near continental shelves and in mid-latitudes where the mixed layer was deeper. The highest Chl_{SCM} (0.8 mg m^{-3}) were predicted in the eastern equatorial Pacific and eastern Central Atlantic, with minimal seasonal variability.

Predicted profiles located 15° either side of the equator had prominent peaks in chlorophyll around 100 m without a corresponding peak in b_{bp} (Figure 4.7). These profiles had a peak in Chl:C_{phyto} (typically around 0.016 mg Chl mg $C_{phyto}{}^{-1}$) located between 10 m and 20 m below the SCM. There was a small increase in b_{bp} from the surface to the SCM, before it started decreasing with depth. SCM thickness was considerably larger for profiles at 15° and 30° (in both hemispheres) than at the equator. SCMs around the equator appeared as SBMs, with peaks in b_{bp} occurring just above the SCM. Despite the increased b_{bp}, the highest Chl:C_{phyto} ratios (0.025 mg Chl mg $C_{phyto}{}^{-1}$) were found in the mid-latitudes during spring and summer, and at the equator. Seasonal variability in chlorophyll, b_{bp}, and Chl:C_{phyto} profile shape increased with distance from the equator (Figures 4.7 and 4.8). In the subtropics, the overall shape of chlorophyll profiles below 40 m resembled Gaussian curves (Figure 4.7).

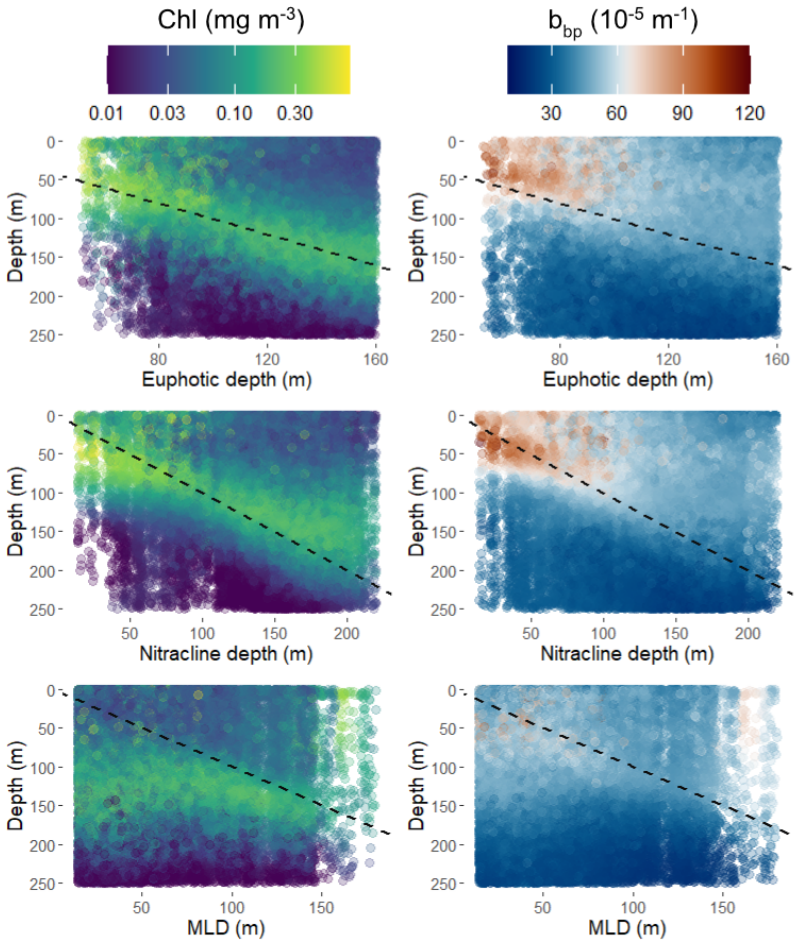

FIGURE 4.5: Fitted profiles of chlorophyll and b_{bp} compared to the covariates from the non-linear model. The dashed line represents the one-to-one line of the covariate depth. Each row shows the effect of a different covariate. Note that I use points rather than lines here to avoid overlapping results.

Chlorophyll profiles without a prominent SCM and slightly elevated near-surface concentrations were predicted during winter 30° from the equator in both hemispheres (Figure 4.7).

Zonal averages of the predictions revealed five latitudinal regimes, identified by changes in the sign of the difference between z_{eu} and z_{ncline} (Figure 4.8). These bands roughly corresponded to the north and south mid-latitudes, the north and south subtropics, and the equatorial region respectively. Note that the precise boundaries of these regimes changed seasonally. Positive values of z_{eu}- z_{ncline} were associated with SCMs and higher Chl_{SCM}, while negative values corresponded to deeper SCMs and lower Chl_{SCM}. Elevated b_{bp} in the top 80 m was restricted to the three latitudinal bands where $z_{\text{eu}} < 100$ m (the equator, and the mid latitudes). In the subtropical gyres the elevated Chl:C_{phyto} was typically located between the z_{eu} and z_{ncline}, whereas in equatorial regions, it extended to about 20 m below the z_{eu}. SCMs with high

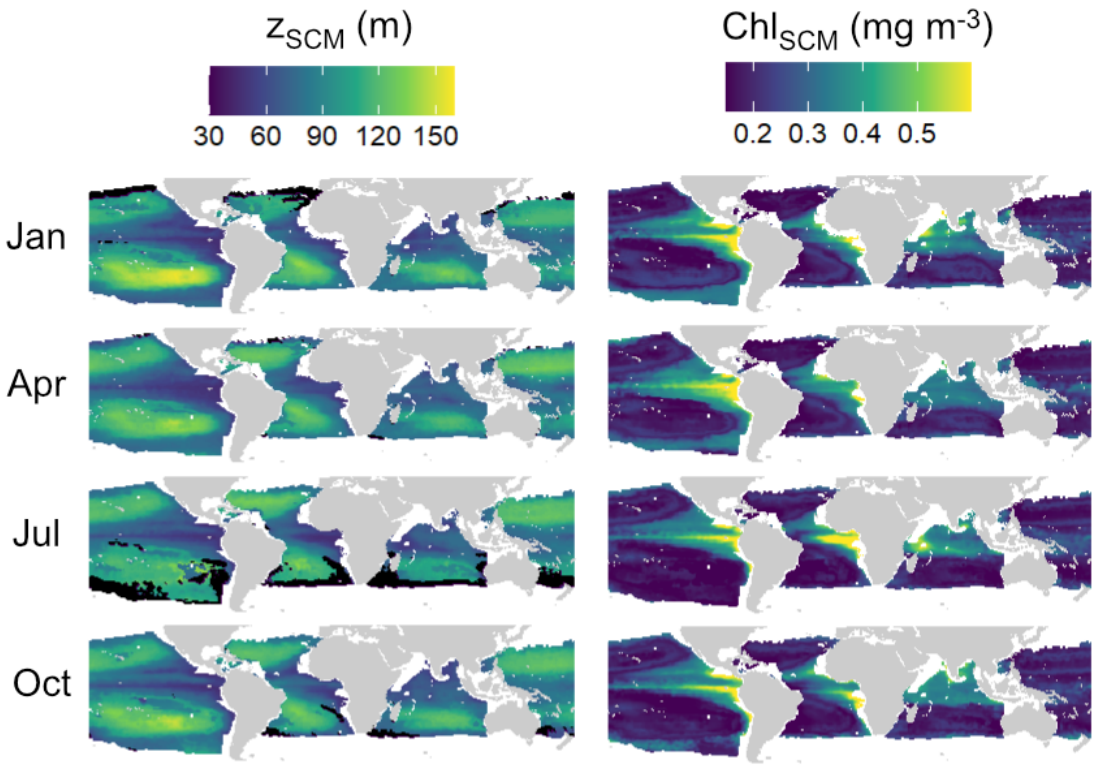

FIGURE 4.6: Predicted climatologies of z_{SCM} and Chl_{SCM} for January, April, July and October. The black grid cells indicate locations where the non-linear model did not predict an SCM but instead predicted the maximum chlorophyll concentration at the shallowest prediction depth of 5 m (2.2% of predicted profiles).

Chl:C_{phyto} rarely overlapped the depths of high b_{bp}. Instead its distribution with respect to latitude more closely resembled that of chlorophyll.

4.5 Discussion

4.5.1 Comparing FRMs for bio-optical profiles

I present a novel method for analysing biogeochemical profiling data, with the aim of describing relationships between environmental conditions and vertical profiles of bio-optical variables in SCM environments. Specifically, FRMs were employed to identify how profiles of chlorophyll and b_{bp} were affected by light, nitrate and stratification. Two different model structures were compared with linear effects and non-linear effects respectively. The model with non-linear effects performed substantially better than the model with scalar effects (Table 4.1). In particular the relationship between z_{SCM} and Chl_{SCM} was best reproduced by the non-linear scalar model (Figure 4.3). My approach allows for the study of vertical profile variability and effects of environmental conditions without prior extraction of specific SCM

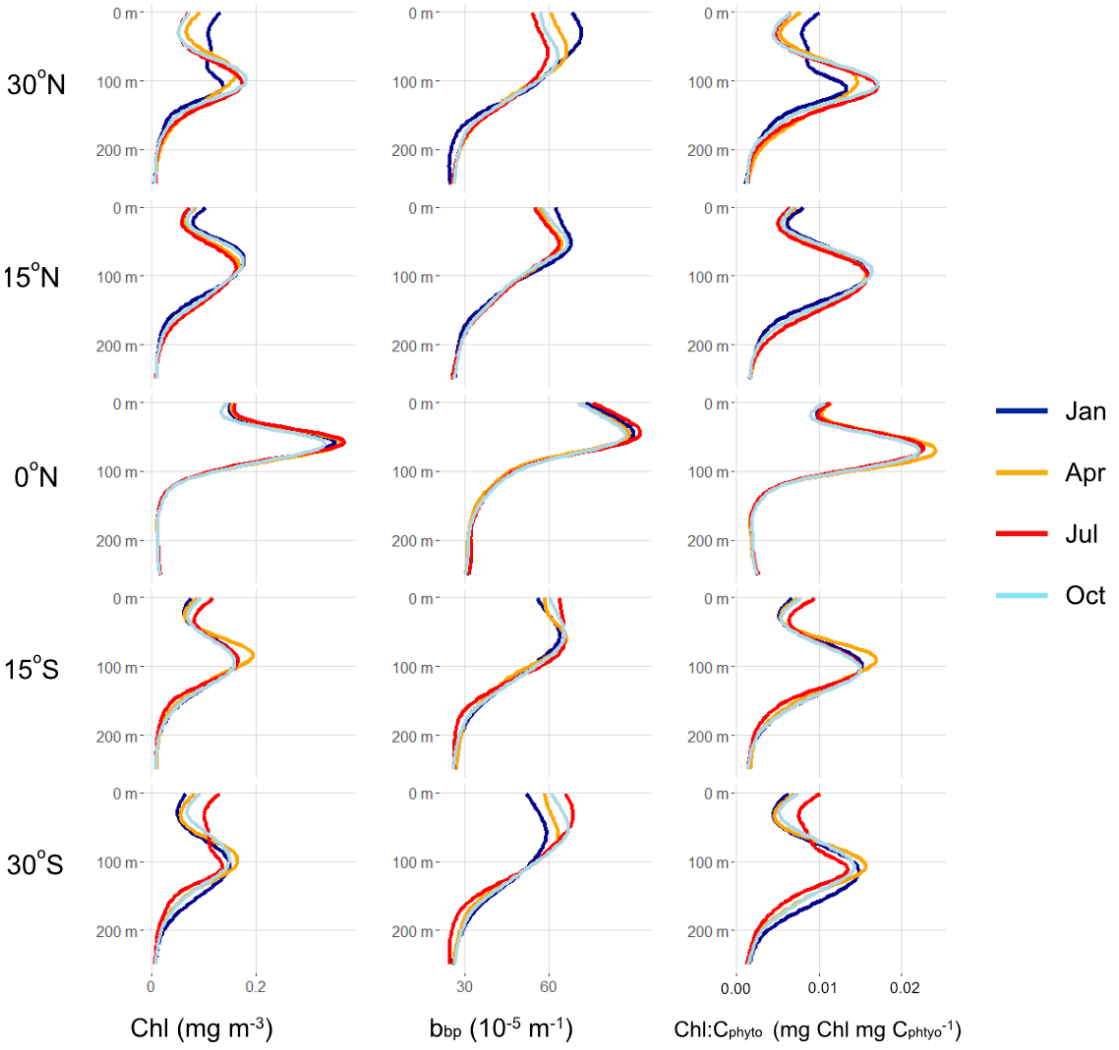

FIGURE 4.7: Predicted zonally-averaged climatological profiles of chlorophyll, b_{bp} and Chl:C_{phyto} for January, April, July and October at a selection of latitudes in the low latitudes.

characteristics or fitting a profile to a family of mathematical curves (Gong et al., 2015; Carranza et al., 2018; Brewin et al., 2022; Xu et al., 2022b). This increases the flexibility of the model and allows the data to speak for itself, although it does require greater computational power and more advanced statistical understanding to interpret the output. None of the models recreated very thin phytoplankton layers (< 5 m), which have been highlighted by Durham and Stocker (2012), or chlorophyll profiles with two peaks (Muñoz-Anderson et al., 2015), which might be better modelled with a reduced dataset containing only profiles with such characteristics. It is worth mentioning that I did not investigate using functional covariates, i.e., profiles as covariates, which intuitively might provide more information for predicting bio-optical profiles. However, I did not decide to try this as inferring real world mechanisms might have been more difficult to identify from model output. This could be worth attempting in future work.

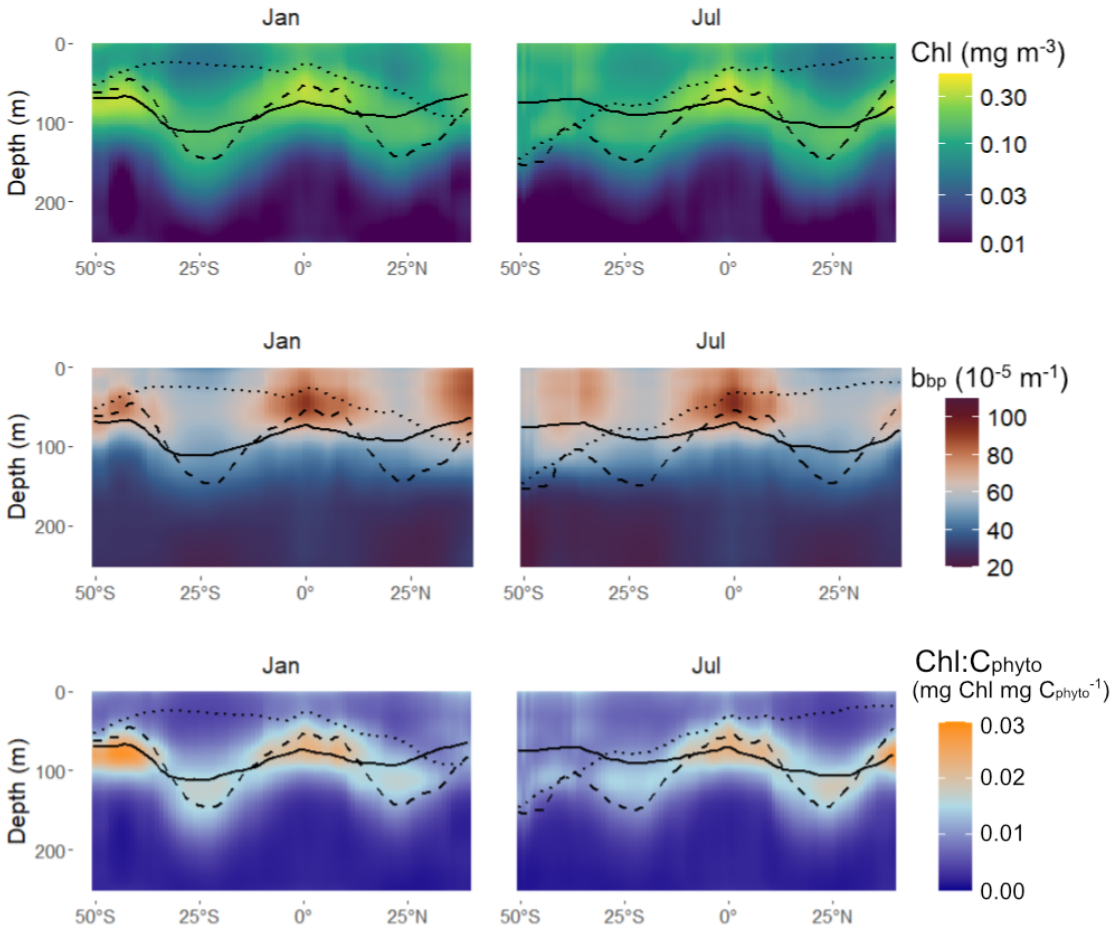

FIGURE 4.8: Climatological zonal averages for January and July of predicted chlorophyll, b_{bp} and Chl:C_{phyto}. The solid, dashed and dotted black curves denote the z_{eu}, the z_{ncline}, and the MLD respectively.

4.5.2 Relationships between bio-optical profiles and environmental conditions

In contrast to previous studies (Uitz et al., 2006; Cornec et al., 2021a; Xu et al., 2022b), I obtained statistical relationships between environmental conditions and continuous bio-optical profiles. My results suggest that the SCMs form at the z_{eu}, whereas peaks in backscatter sit closer to the z_{ncline}. The prediction of elevated b_{bp} at and above the z_{eu} within the top 80 m in equatorial regions aligns with Bock et al. (2022). The coupling between z_{eu} and z_{SCM} (Figure 4.5) supports previous findings (Letelier et al., 2004; Mignot et al., 2014; Xu et al., 2022b; Xing et al., 2023; Garg et al., 2024). My estimated effect of the MLD on the chlorophyll was negligible compared to those of z_{eu} and z_{ncline}, which suggests the MLD plays a smaller role in determining SCM characteristics. This would agree with previous studies who found that the MLD only affected SCMs in seasonally stratified regions (Bock et al., 2022; Dai et al., 2023). The strong positive correlation between the z_{SCM} and the z_{ncline} described by Herbland and Voituriez (1979) was less evident in my study. I identified large-scale relationships

between environmental conditions and bio-optical profiles (Figures 4.5 and 4.8). Latitudinal patterns in bio-optical profiles have been observed before but here I provide a bridge between small-scale studies that used in situ data (Xu et al., 2022b; Garg et al., 2024) and theoretical modelling studies (Fennel and Boss, 2003; Hodges and Rudnick, 2004; Gong et al., 2015). One notable inconsistency in my results is the reduction in Chl_{SCM} for SCMs located at a depth of around 100 m (Figure 4.3). This occurs in locations where both z_{eu} and z_{ncline} are both around 100 m (Figure B.6). This might be explained by the effects from z_{eu} and z_{ncline} cancelling each other out, resulting in an unusual chlorophyll profile shape compared to those of slightly different covariate values. Three-dimensional scatter plots showing the combinations of the covariates for the observed and predicted profiles are available in Figures B.7 and B.8.

4.5.3 Large-scale patterns in bio-optical profiles

I observed that the deepest and least intense SCMs occur in the subtropical gyres, while shallower and more intense SCMs are found near the equator (Figures 4.7 and 4.8), consistent with previous studies (Cornec et al., 2021a; Masuda et al., 2021; Bock et al., 2022; Yasunaka et al., 2021). My results also indicate that, in general, SCMs in the subtropics are not only deeper but also exhibit the thickest vertical structures (Figure 4.7). Additionally, I found that seasonal variability in chlorophyll and b_{bp} increases with distance from the equator (Figures 4.7 and 4.8), aligning with the findings of Cornec et al. (2021a) and Bock et al. (2022). At the equator, seasonal changes in the z_{SCM}, b_{bp}, and $\text{Chl:C}_{\text{phyto}}$ are minimal (typically less than 10 m across the year) supporting observations by Bock et al. (2022).

The zonally averaged predictions suggest the $\text{Chl:C}_{\text{phyto}}$ ratio does increase in SCMs across a range of latitudes, with a larger proportion seen during summer than winter (Figure 4.8). The $\text{Chl:C}_{\text{phyto}}$ ratio is very high (> 0.02 mg Chl mg $\text{C}_{\text{phyto}}{}^{-1}$) even as far south as 50°S during summer, suggesting that nutrients have been depleted and light is now a limiting factor. Steele (1964) found that SCM depth is not related to the maximum possible value of the $\text{Chl:C}_{\text{phyto}}$ ratio. However, it is possible that increases in $\text{Chl:C}_{\text{phyto}}$ might not only be due to photoacclimation but also represent a change in species composition (Letelier et al., 2004) which may be adapted to have different maximum $\text{Chl:C}_{\text{phyto}}$ ratios, indicating a longer term selection rather than a physiological change within individual organisms.

4.5.4 Limitations and future work

A notable caveat of this study was the prediction of bio-optical profiles in locations distant from the observations used for model fitting. This was due to the sparseness

and clustering of floats equipped with all of the sensors and restricting the dataset to those profiles completed within two hours of local noon. Most of the available profiles were concentrated in the subtropical gyres, which may have biased covariate effect estimates. This issue may be alleviated in the future as more fully equipped floats are deployed more evenly throughout the global ocean. I did not include any interaction effects in the additive non-linear model, which may be important since phytoplankton growth depends on the availability of both nutrients and light simultaneously. However, extracting useful information from such a model may have been difficult. Additionally, top-down controls on phytoplankton, which have been shown to be significant (Longhurst, 1976; Prowe et al., 2012; Moeller et al., 2019; Rodríguez-Gálvez et al., 2023), were not considered due to the lack of coincident measurements of zooplankton with BGC-Argo profiles. The method used to infer Chl:C_{phyto} may not be applicable at all prediction locations, particularly in regions with the deepest SCMs. However, its application across a broad range of latitudes by Arteaga et al. (2022) suggests it may still be appropriate in many cases. Finally, it may not be entirely appropriate to define the z_{eu} based on a fraction of surface irradiance (Banse, 2004), since this does not account for the actual amount of light reaching a given depth. Instead, perhaps a definition based on absolute light intensity would be more informative from a physiological perspective.

This work does not include nutrients other than nitrate, which may also be limiting factors in phytoplankton growth, as the BGC-Argo floats do not carry sensors for phosphate, silicate, etc. It could be interesting to apply this methodology to a dataset containing measurements of a wider range of potentially limiting nutrients, for example the GEOTRACES program (Anderson, 2020). Alternatively, this approach could be used to analyse the vertical distribution of different phytoplankton species in various environments, such as the datasets analysed by Sato et al. (2022) and Miyares et al. (2024). Although this study focussed on SCMs located in permanently stratified biomes, several studies have highlighted their formation at higher latitudes in summer (Bouman et al., 2020; Cornec et al., 2021a; Baldry et al., 2020, 2024). Moreover, the study of SCMs could be applied to quantify the effect of mesoscale ocean physics on SCMs (Cornec et al., 2021b; Wang and Liu, 2024). From a statistical perspective, a bivariate FRM may more effectively capture the covariance between chlorophyll and b_{bp}, potentially leading to more accurate estimates of the Chl:C_{phyto} ratio, although this may significantly increase the computational cost of model fitting. Alternatively, approaches that incorporate spatio-temporal latent effects, such as the one developed by Yarger et al. (2022), could offer further insights. As the number of BGC-Argo floats equipped with the full array of sensors increases both in number and global coverage (Owens et al., 2022; Thierry et al., 2025), there may be opportunities to apply the present methodology to address further questions from marine biogeochemistry.

4.6 Conclusions

Over the past decade, SCMs have received growing attention, partly due to the increasing availability of subsurface measurements of bio-optical parameters from the BGC-Argo float array. Plenty of research has focussed on establishing relationships between characteristics of SCMs (e.g. z_{SCM} or Chl_{SCM}) and environmental conditions, however these have usually involved prior identification of SCM characteristics (Herbland and Voituriez, 1979; Cornec et al., 2021a) or the fitting of profiles to convenient mathematical curves (Gong et al., 2015; Xu et al., 2022b). In this study I present a functional regression analysis, in which data takes the form of curves, of chlorophyll and b_{bp} profiles, using data from 26 BGC-Argo floats (comprising 1323 profiles) across tropical and subtropical ocean regions. Through an additive model, I identified non-linear relationships between bio-optical profiles and the z_{eu}, the z_{ncline}, and the MLD. Notably, I found that the depth of the z_{SCM} closely follows z_{eu}, while peaks in b_{bp} were concentrated at the z_{ncline}. This suggests that light availability is main control of z_{SCM} whereas biomass accumulates at the nitracline. These findings align with theoretical results by Fennel and Boss (2003) and Gong et al. (2015) by indicating that photoacclimation is a primary driver of most SCMs, not only for the deepest SCMs. Using my model, I produced a climatology of predictions of chlorophyll, b_{bp} and, through subsequent calculations, Chl:C_{phyto}. This revealed the large-scale differences in subsurface bio-optical profiles previously identified by Bock et al. (2022). Overall, my findings demonstrate that functional data analysis is a viable and insightful approach for investigating biogeochemical processes, and it is possible to reconstruct large-scale interpolations through the use of a few features of the water column as covariates.

The methodology shown in this work could be extended in several ways including assessing the vertical distributions of phytoplankton functional groups (Brewin et al., 2022; Sato et al., 2022), or the influence of mesoscale physics (Cornec et al., 2021b; Xu et al., 2022b) on bio-optical profiles. Moreover, this work was restricted to regions with strongly stratified water columns so it could be interesting to repeat it for mid-latitude or polar environments, where SCMs do form during summer (Cornec et al., 2021a; Baldry et al., 2020). Here I have demonstrated the usefulness of utilising FRMs for profiling data and hope that others in marine biogeochemistry find it an attractive alternative to previous methods, especially as the number of BGC-Argo floats equipped with a full selection of sensors increases (Owens et al., 2022; Thierry et al., 2025).

Chapter 5

Scalar Variance and Correlation for Oceanographic Profiles: an Argo Float Application

This chapter is, at the time of writing, in preparation for publication in *Global Biogeochemical Cycles* as:

Taylor, M., Henson, S., Sahu, S., Hammond, M., Cael, B.B. Scalar Variance and Correlation for Oceanographic Profiles: an Argo Float Application.

5.1 Abstract

Depth profiles are a commonly used observation in oceanography and their number has grown rapidly in recent years with the deployment of autonomous platforms such as biogeochemical-Argo (BGC-Argo) floats. Functional data analysis (FDA) allows us to treat profiles as single datums, enabling the analysis of profile shape within a convenient and coherent framework. Consequently, FDA has recently been utilised in oceanographic studies. However, previous analyses have typically assessed variability only as a function of depth, without providing a single summary measure for variance across entire profiles. Here, I applied a new technique that calculates scalar-valued measures of variance and correlation for groups of curves. This enabled the assessment of profile variability and correlation in ways directly analogous to traditional scalar-valued data analysis. I used this method to assess the variability of chlorophyll and temperature profiles (between 5 m and 250 m) from a global dataset of 98 413 BGC-Argo floats. Chlorophyll profiles had significantly higher variance in the high latitudes during spring and summer than in the tropics. The variability of temperature profiles was greatest between 30° and 40° either side of the equator, in

regions with fronts between water masses. The strongest correlation between chlorophyll and temperature profiles occurred at the fronts, as well as in the Mediterranean Sea. From a semi-Lagrangian perspective, I found that the seasonal autocorrelation of temperature profiles is stronger than that of chlorophyll profiles, assuming that a float remains in a relatively small area. From an Eulerian perspective, I showed that seasonal variation dominates spatial variation towards the poles when compared over large spatial scales. This work provides an indication that scalar variance and correlation of oceanographic profiles could be useful in a variety of contexts, including observing system optimisation and calibrating sensors.

5.2 Introduction

Amongst the most common forms of oceanographic observation is the depth profile since it aids the assessment of depth-dependent variation of properties of seawater. In many scenarios, the overall shape of a depth profile is as important as the exact numerical values at any specific depth as this can provide information about the vertical structure of the ocean and how phenomena are formed and maintained, e.g. the distribution of water masses or the structure of eddies. In recent decades, the quantity of depth profiles has increased substantially due to the widespread deployment of autonomous observing platforms, which are capable of measuring a variety of variables concurrently in remote locations and on a global scale, previously unattainable through traditional ship-based research (Chai et al., 2020). Consequently, the development of statistical tools appropriate and, if possible, dedicated for the analysis of depth profiles will help reveal features of the vertical structure of the ocean.

Functional data take the form of curves and surfaces, whereby the variable of interest is considered a function of another so-called indexing variable. Recent work has demonstrated the benefits of treating oceanographic profiles as continuous functional data objects, where depth (or equally pressure) is used as an indexing variable (Yarger et al., 2022; Korte-Stapff et al., 2022). This allows for the essence of the shape of an oceanographic profile to be captured within each datum, alongside the numerical values. Although previous studies have utilised this approach in the context of spatio-temporal modelling, there are opportunities to explore more fundamental analyses through the lens of functional data analysis. For example, Urbano-Leon et al. (2023) developed a methodology for quantifying the variance of a set of functional data as a single value (a scalar), when previously only a variance function was possible (Ramsay and Silverman, 2005). This approach involves representing each function as the linear combination of a set of basis functions and summing the variances of basis coefficients. Moreover, they demonstrated a simple extension to calculate a scalar correlation coefficient between paired sets of functional data, allowing them to quantify the similarity of annual temperature patterns across several

regions in Canada. The purpose of this work is to present the first calculation of scalar variance and correlation for oceanographic profiles, using autonomous platform data as an example. This provides oceanographers with a new tool to quantify and compare profile variability, potentially aiding interpretation of physical and biogeochemical patterns in the ocean.

The biogeochemical-Argo (BGC-Argo) float array is a global network of profiling floats that carry a range of sensors capable of measuring biogeochemical and physical parameters of the top 2000 m of the water column (Claustre et al., 2020). Around 520 BGC-Argo floats can measure chlorophyll, a common proxy for the biomass of phytoplankton, which form the base of the marine food web and play an important role in the ocean carbon cycle (Falkowski, 1994). The understanding of subsurface chlorophyll distribution has improved significantly due to the abundance of measurements by BGC-Argo floats over a range of spatial and temporal scales from global (Cornec et al., 2021a; Yasunaka et al., 2021) to mesoscale and sub-seasonal scales (Cornec et al., 2021b; McKee et al., 2023; Strutton et al., 2023).

Identifying spatial and temporal scales of variation of chlorophyll is useful in determining the mechanisms that promote or hinder the growth of phytoplankton. Several studies have revealed a connection between mesoscale dynamics (such as eddies) and chlorophyll concentration (Cornec et al., 2021b; McKee et al., 2022; Strutton et al., 2023). Some analyses have found that the difference between length scales of surface chlorophyll when viewed from Eulerian and Lagrangian time scales are negligible (Kuhn et al., 2023), whereas McKee et al. (2022) utilised both frameworks to find that profile anomalies were connected to mesoscale stirring. Temperature fields have also been characterised by distinct length scales through a variety of methods (Storto et al., 2018; Gille and Kelly, 1996; Mirouze et al., 2016; Song et al., 2022). However, a gap in the literature remains for the global assessment of the spatial and temporal length scales of chlorophyll profiles and how they compare to those of temperature profiles. In this work, I calculated the variance of chlorophyll and temperature profiles as well as their correlation from 98 413 BGC-Argo floats on a variety of spatial scales and from Eulerian and Lagrangian perspectives. My results suggest that the spatial scales over which profiles are paired when calculating correlation is important when comparing chlorophyll and temperature. Furthermore, I found that temporal length scales in a Lagrangian framework are less useful when a float trajectory passes from one biogeographical region to another.

The development of specific statistical techniques for oceanographic profiles is important due to the increased deployments of autonomous profiling platforms. Benefits of utilising the variance and correlation for profiles could range from identifying relationships between the shapes of profiles to the calibration of multiple platforms and optimisation of observing system deployments (Chamberlain et al., 2023; Chu et al., 2024). Given the trajectory of oceanography towards increased

autonomous subsurface measurements, suitable statistical analysis methodologies are necessary to extract maximum value from the data. My results provide an example of treating profiles as functional data allows for the calculation of a novel measure of variance and correlation, and how these metrics vary over space and time.

5.3 Materials and Methods

5.3.1 Scalar variance and correlation of functional data

Functional datasets are those in which each datum takes the form of a continuous curve or surface which is a function of at least one other variable. Here, I apply this framework to chlorophyll and temperature as a function of pressure. Recent developments in the field of functional data analysis have produced a method for calculating scalar-valued summary statistics, specifically the variance and correlation, for functional datasets (Urbano-Leon et al., 2023). The method will be described briefly here but refer to the original paper for full details and proofs. Suppose that I have two sets $\mathcal{X}$ and $\mathcal{Y}$ each containing n curves (for example, paired chlorophyll and temperature profiles) and then decompose each curve from $\mathcal{X}$ and $\mathcal{Y}$ into p orthogonal basis functions with $a_{i,j}$ and $b_{i,j}$ denoting the coefficient of the jth basis component of the ith curve in sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. The mean basis coefficients for each set are calculated as follows

$$\overline{A_j} = \frac{1}{n}\sum_{i=1}^{n} a_{i,j}, \quad \overline{B_j} = \frac{1}{n}\sum_{i=1}^{n} b_{i,j}. \tag{5.1}$$

The mean basis coefficients represent a mean function for each set, describing a characteristic curve shape. Using these mean functions, I can calculate the variance of the coefficients for each of the basis functions (denoted V_{a_j} and V_{b_j} respectively). The variances of sets $\mathcal{X}$ and $\mathcal{Y}$ are then simply the sums of each of the basis variances V_{a_j} and V_{b_j} respectively.

$$Var(\mathcal{X}) = \sum_{j=1}^{p} V_{a_j} = \sum_{j=1}^{p} \frac{1}{n}\sum_{i=1}^{n} (a_{i,j} - \overline{A_j})^2 \tag{5.2}$$

$$Var(\mathcal{Y}) = \sum_{j=1}^{p} V_{b_j} = \sum_{j=1}^{p} \frac{1}{n}\sum_{i=1}^{n} (b_{i,j} - \overline{B_j})^2 \tag{5.3}$$

Similarly, the covariance between the sets for the jth basis function is C_j and the total covariance between the sets $\mathcal{X}$ and $\mathcal{Y}$ is the sum of the covariances of each basis coefficient C_j.

$$Cov(\mathcal{X}, \mathcal{Y}) = \sum_{j=1}^{p} C_j = \sum_{j=1}^{p} \frac{1}{n}\sum_{i=1}^{n} (a_{i,j} - \overline{A_j})(b_{i,j} - \overline{B_j}) \tag{5.4}$$

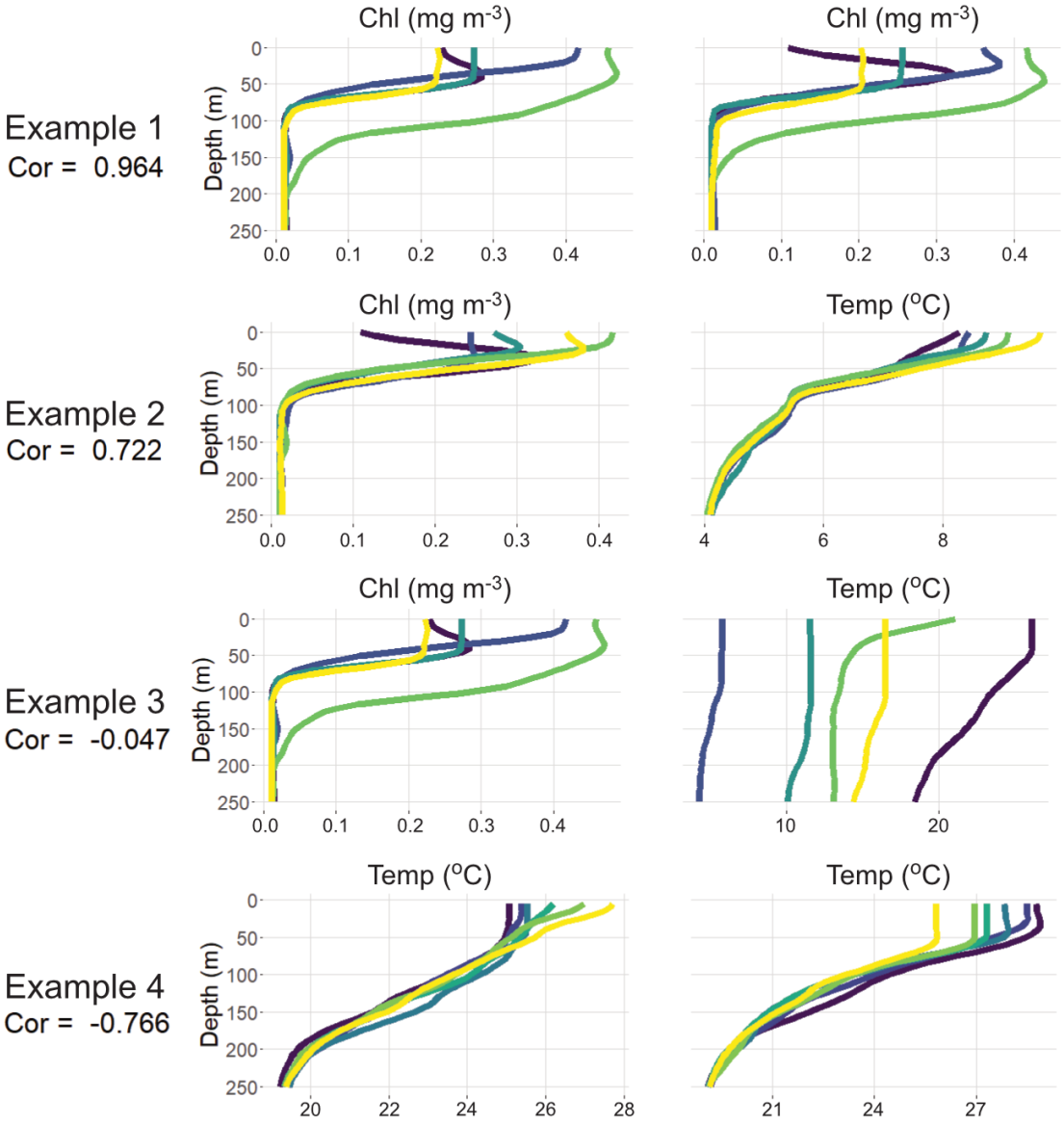

FIGURE 5.1: Four examples of paired sets of ocean profiles as functional data and their respective correlations. Colours indicate matching pairs across datasets. The specific numerical values within each profile are not as important as the direction and relative magnitude of any deviations from a typical profile shape.

The equation to calculate the correlation between the sets $\mathcal{X}$ and $\mathcal{Y}$ is the same for scalar data.

$$Cor(\mathcal{X},\mathcal{Y}) = \frac{Cov(\mathcal{X},\mathcal{Y})}{\sqrt{Var(\mathcal{X},\mathcal{Y})Var(\mathcal{X},\mathcal{Y})}} \tag{5.5}$$

The standard deviation of functional data is simply the square root of the variance, in an identical way to scalar-valued statistics.

Mathematically, a correlation of +1 would represent a case where each basis coefficient is perfectly linear and positively correlated. A deviation (from the mean function) in a specific basis component in set $\mathcal{X}$ corresponds to a deviation in the same direction in the equivalent component in the set $\mathcal{Y}$. In contrast, a correlation of -1 represents a case

where each basis is perfectly linear and negatively correlated. A correlation of +1 implies the functional shape is maintained perfectly whereas a correlation of -1 suggests the sets are exactly out of phase, meaning any deviation in a particular component in set $\mathcal{X}$ corresponds to a deviation in the opposite direction from the mean function in set $\mathcal{Y}$. Note that this does not imply that each pair of functions is a pointwise negative of the other, but instead they vary in opposite directions along the structural features captured by the basis coefficient. Figure 5.1 shows several examples of datasets containing six pairs of oceanographic profiles, and their correlation. The first example displays a near-perfect dependency, meaning almost all deviations from the mean in one set are replicated in the second set. The second shows a slightly weaker positive correlation but a clear relationship is visible by visual inspection. The third shows almost zero correlation and the fourth shows a case with strong negative correlation where deviations from a mean function in the positive direction near the surface in one set corresponds to negative deviations from the mean function near the surface in the other set.

5.3.2 BGC-Argo float data

In this study I used profiles collected by BGC-Argo floats with chlorophyll and temperature measurements. Observations were unevenly distributed across the global ocean, increasing in frequency over time between 30/05/2010 and 30/12/2024. Profiles were partitioned into groups according to the mean biomes defined by Fay and McKinley (2014). The Mediterranean Sea (excluding the Black Sea) was used as an additional region. Any profiles without a timestamp were excluded from the analysis. Measurements assigned a quality control flag by the Argo data centre of 3 (”Probably bad”) or 4 (”Bad”) were removed. Profiles with fewer than 20 measurements between 5 m and 250 m were also removed, as were those whose range of measurements did not span at least from 20 m to 230 m. In total, 98 413 profiles from 890 BGC-Argo floats remained and were used in subsequent analyses. Figure 5.2 shows the distributions of profiles in space and time.

The R package **castr** was used to prepare chlorophyll and temperature profiles for analysis. The profiles were regridded to 5 m across a range from 5 m to 250 m using linear interpolation. In cases where the edge values were NA, the nearest non-NA value was interpolated until either 5 m or 250 m. Each profile was then smoothed using a sliding window of width 15 m using a moving median. The chlorophyll profiles had been corrected for non-photochemical quenching prior to download (Schmechtig et al., 2023). A $\log_{10}$ transformation was not applied to the chlorophyll profiles in this work (as in previous chapters) because that would have reduced the variability of large values (which are of considerable interest) whilst increasing the relative variability of the small values. This could have led to poorly identifying

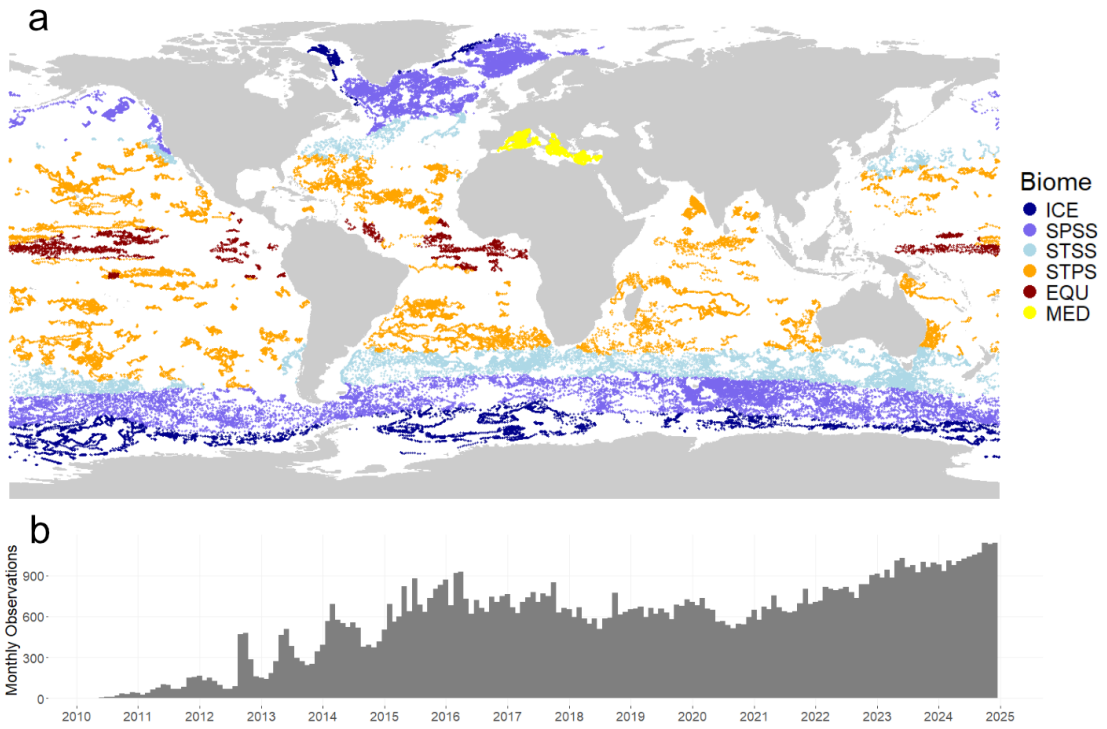

FIGURE 5.2: (**a**) Global distribution of the 98 413 BGC-Argo profiles used in this work. Points are coloured by the mean biome in which they are located according to the classification by Fay and McKinley (2014). Profiles were partitioned into the ice biome (ICE, n = 10 173), the subpolar seasonally stratified biome (SPSS, 30 958), the subtropical seasonally stratified biome (STSS, 14 483), the subtropical permanently stratified biome (STPS, 29011) and the equatorial biome (EQU, 4454). The Mediterranean Sea was also included as an additional biogeographical region (MED, 9366). (**b**) Timeline showing the number of completed profiles globally each month from January 2010 to December 2024.

strong correlation given that the overall shape of the profile has been altered and amplified uninteresting sections of profiles. Moreover, in contrast to the analysis in Chapter 4, there was no prediction and consequently there was no requirement to enforce positivity of chlorophyll measurements later. Profiles with either no recorded location or an unreasonable location (detected by a float trajectory having a speed greater than 0.5 ms^{-1}) were identified by interpolating along the great-circle path between the two adjacent observations, using the relative time of the missing location.

5.3.3 Application of scalar variance and correlation on BGC-Argo profiles

In a similar way to Yarger et al. (2022) and Korte-Stapff et al. (2022), I consider the profiles from Argo floats as functional data objects. Specifically, I treat chlorophyll and temperature profiles as functions with respect to pressure. As described in section 5.3.1, the profiles are represented using basis functions. Due to the continuous nature of oceanographic profiles and for computational efficiency, I use Fourier basis functions with 50 components (one for each of the regridded profile measurements).

From here on, the analysis is divided into three parts depending on how the theoretical sets $\mathcal{X}$ and $\mathcal{Y}$ are comprised with the BGC-Argo profiles prior to calculations of variance and correlation.

5.3.3.1 Variance and correlation between chlorophyll and temperature profiles

Consider the variance and correlation of concurrent profiles of chlorophyll and temperature. The sets $\mathcal{X}$ and $\mathcal{Y}$ are simply the chlorophyll and temperature profiles respectively (the labelling of each set is unimportant, but for consistency chlorophyll will be set $\mathcal{X}$ if both variables are involved). The respective variances of each variable and their correlation were calculated for the entire dataset (98 413 profiles) as well as after extracting profiles for individual biomes. Then the profiles were binned into a 5° longitude-latitude grid. In each grid cell, the variances of chlorophyll and temperature, and their correlation was calculated, in cells with at least 10 profiles in an individual bin. The same was done for meteorological seasons and monthly bins on a 10° grid.

5.3.3.2 Semi-Lagrangian autocorrelation of chlorophyll and temperature profiles

Suppose instead of grouping profiles by location, I group them by which float they were collected. Argo floats are free to drift with the horizontal advection of the ocean and consequently can follow the same water mass over time (a Lagrangian framework). Given that Argo floats spend the majority of their time at their parking depth of 1000 m, and my depths of interest are shallower than 250 m, the float will experience a slightly different horizontal drift to the surface. Hence, the Lagrangian perspective of chlorophyll is only semi-Lagrangian although previous application of this has yielded effective results (McKee et al., 2022).

I analysed seven floats whose lifespans were longer than two years and had regular sampling frequencies namely (WMOs 1902385, 4903365, 6901767, 5905107, 5906204, 5904021 and 6901585). These floats represented a range of biogeographical regions. Sets $\mathcal{X}$ and $\mathcal{Y}$ are filled with corresponding chlorophyll and temperature profiles. The respective variances of chlorophyll and temperature profiles were calculated for each float, as well as their correlation.

In addition, the temporal autocorrelation functions (ACFs) of chlorophyll and temperature profiles were calculated. This was done by iteratively extracting pairs of profiles with a certain time lag. Sets $\mathcal{X}$ and $\mathcal{Y}$ were filled with the earlier and later profiles from pairs respectively, but crucially here they are the same variable but from another cycle. The correlation between $\mathcal{X}$ and $\mathcal{Y}$ was then calculated for daily time

lags from 1 to 730 days. If there were fewer than 10 pairs of profiles for a given time lag, then no correlation was calculated.

5.3.3.3 Eulerian autocorrelation of chlorophyll and temperature profiles

The Eulerian spatio-temporal autocorrelation structure (STAS) of chlorophyll and temperature profiles was computed separately for each biome. For each combination of spatial radius r and temporal lag l, I identified pairs of profiles $\{i, j\}$ that satisfied the conditions $|d_{GC}(\mathbf{s}_i, \mathbf{s}_j) - r| < r_{\text{error}}$ and $|t_i - t_j| = l$, where $\mathbf{s}$ denotes the geographic coordinates (longitude and latitude) of an observation and t denotes time. The great circle distance (the shortest distance between two points) d_{GC} was computed using the Haversine formula (Robusto, 1957). Note that the distances cross land for a minority of profile pairs. I then calculated the correlation between all profile pairs meeting these criteria. I considered search radii ranging from 50 km to 2000 km in 50 km increments and temporal lags from 1 to 365 days, resulting in a total of 29 200 spatial-temporal combinations. Given that distances between profiles are highly unlikely to be an exact multiple of 50 km apart, I allowed a window of distances, centred about r with maximum difference of r_{error}, which I set as 50 km. To improve the interpretability of the STAS, a smoothing kernel was applied to the resulting autocorrelation grid. This kernel was weighted by the number of profile pairs contributing to each (r, l) combination, and had a span of 150 km in space and 10 days in time. Due to the smaller number of profiles in the ICE and EQU biomes, these were combined with SPSS and STPS biomes respectively for this part of the analysis.

Finally, the previous analysis was repeated using chlorophyll and temperature values from individual depths rather than full vertical profiles. This was done by creating false profiles with the surface value repeated throughout, which is equivalent in practice to using a single depth with one basis function. This resulted in separate STAS calculations for each depth bin from 5 m to 250 m. The resulting depth-specific STASs were then compared to those obtained using the full-profile data.

5.3.3.4 Computation

All analyses were conducted using the statistical software R version 4.4.1 R Core Team (2023) on a computer with an intel Core i5 processor. Calculating the Eulerian spatio-temporal autocorrelation was the computationally most expensive result, which took around 30 minutes. The majority of this time is spent calculating the distance matrix between all profiles a certain time lag apart. The search for viable pairs in terms of temporal lag was optimised using a function written in C++, which was considerably faster than the equivalent function written in R.

	Variance		Covariance	Correlation
	Chlorophyll	Temperature		
Global	119	152185	-958	-0.22
EQU	18	27756	57	0.08
STPS	18	34481	-182.2	-0.23
STSS	56	32199	-370	-0.27
SPSS	190	18859	-104	-0.05
ICE	306	1983	69	0.09
MED	31	6360	-152	-0.34

TABLE 5.1: Summary of the variance and the correlation of annual chlorophyll and temperature profiles (5 m - 250 m), by region.

5.4 Results

5.4.1 Variance and correlation of chlorophyll and temperature coincident profiles

The variances of chlorophyll and temperature profiles varied substantially across biomes (Table 5.1). Chlorophyll profiles exhibited the highest variances in high-latitude regions, particularly in the ICE and SPSS biomes, with variances up to 17 times greater than those observed in the EQU and STPS biomes. In contrast, the highest temperature variances were found in the EQU, STPS, and STSS biomes, which may partly reflect the broader latitudinal ranges encompassed by these regions. The ICE biome, by comparison, exhibited relatively low temperature variance. The global correlation between chlorophyll and temperature profiles (measured between 5 m and 250 m depth) was slightly negative (-0.22). No strong correlation between the two variables was found within any biome, with the largest magnitude observed in the Mediterranean (-0.34), a region that had relatively low variance for each variable. These results suggest that, at biome scales and annual timescales, chlorophyll and temperature profiles are not strongly correlated. It is noteworthy that the global variance of chlorophyll profiles was not greater than in any single biome, whereas the global variance of temperature profiles was approximately an order of magnitude higher than within any individual biome. This means that temperature profiles vary mostly over large spatial scales, compared to seasonally within a single region. In contrast, chlorophyll varies widely within some high-latitude regions, potentially over both space and time. A seasonal breakdown of profile variance is given in Table C.1.

Figure 5.3 shows how the variability of chlorophyll and temperature changed over smaller spatial (10° longitude–latitude grid) and seasonal scales. Chlorophyll profiles exhibited the greatest variability in the North Atlantic and across the Southern Ocean during their respective spring and summer. Within the Southern Ocean, variance increased towards the Antarctic continent. The standard deviation of chlorophyll profiles near the equator was approximately an order of magnitude lower than in

high-latitude regions. Temperature profile variability also varied by an order of magnitude, although a clear latitudinal pattern was less apparent. A persistent band of higher temperature variance occurred between 30°S and 40°S throughout the year - similar to the transition zone identified by Henson et al. (2009). The northwest Atlantic and the eastern equatorial Pacific also exhibited elevated variance. In contrast, the Southern Ocean south of the Antarctic Circumpolar Current showed consistently low temperature variability year-round, as did the extreme northern region of Baffin Bay. Grid cells with the highest temperature variability tended to coincide with frontal zones between subtropical and subpolar water masses, where the generation of eddies may have contributed to greater variability in temperature profile structure. Monthly maps of the standard deviation of chlorophyll and temperature profiles are shown in Figure C.1.

The year-round correlation between chlorophyll profiles and temperature profiles was weak (a magnitude less than 0.3) at most locations (Figure 5.4). Latitudinal bands in each hemisphere between 30° and 40° from the equator showed weak to moderate negative correlation ($-0.6 < \text{Cor} < -0.2$). This relationship was clearest in the North Atlantic, Mediterranean, central South Pacific, and South Atlantic. This may indicate that the variable supply of nutrients (due to changes in the MLD) influences chlorophyll profile shape and consequently that these are regions of nutrient limitation. The region south of 60° showed weak to moderate positive correlation. This likely reflects greater light availability during summer, which contributes to higher temperatures and alleviates light-limited growth. When partitioned into seasonal correlations, the spatial pattern from Figure 5.4 became less clear although the high-latitude regions showed a stronger positive correlation during winter (Figure 5.5). The correlation in the high latitude Atlantic varied considerably throughout the year, even changing sign from negative to positive from summer to autumn. The correlation in the Mediterranean did not vary significantly throughout the year. Monthly maps of the correlation between chlorophyll and temperature profiles are shown in Figure C.2.

5.4.2 Autocorrelation of chlorophyll and temperature profiles from a semi-Lagrangian perspective

From a semi-Lagrangian perspective, floats located at lower latitudes exhibited lower variance among chlorophyll profiles, with differences spanning two orders of magnitude, whereas floats with the highest temperature variance were located in mid- to high-latitude regions (Table 5.2). Notably, float 5906204 travelled a substantial distance during its lifetime, moving from relatively warm waters in the Indian Ocean to cooler waters near the Cape of Good Hope. This movement induced a change in profile structure from persistent deep chlorophyll maxima to near-surface blooms. In

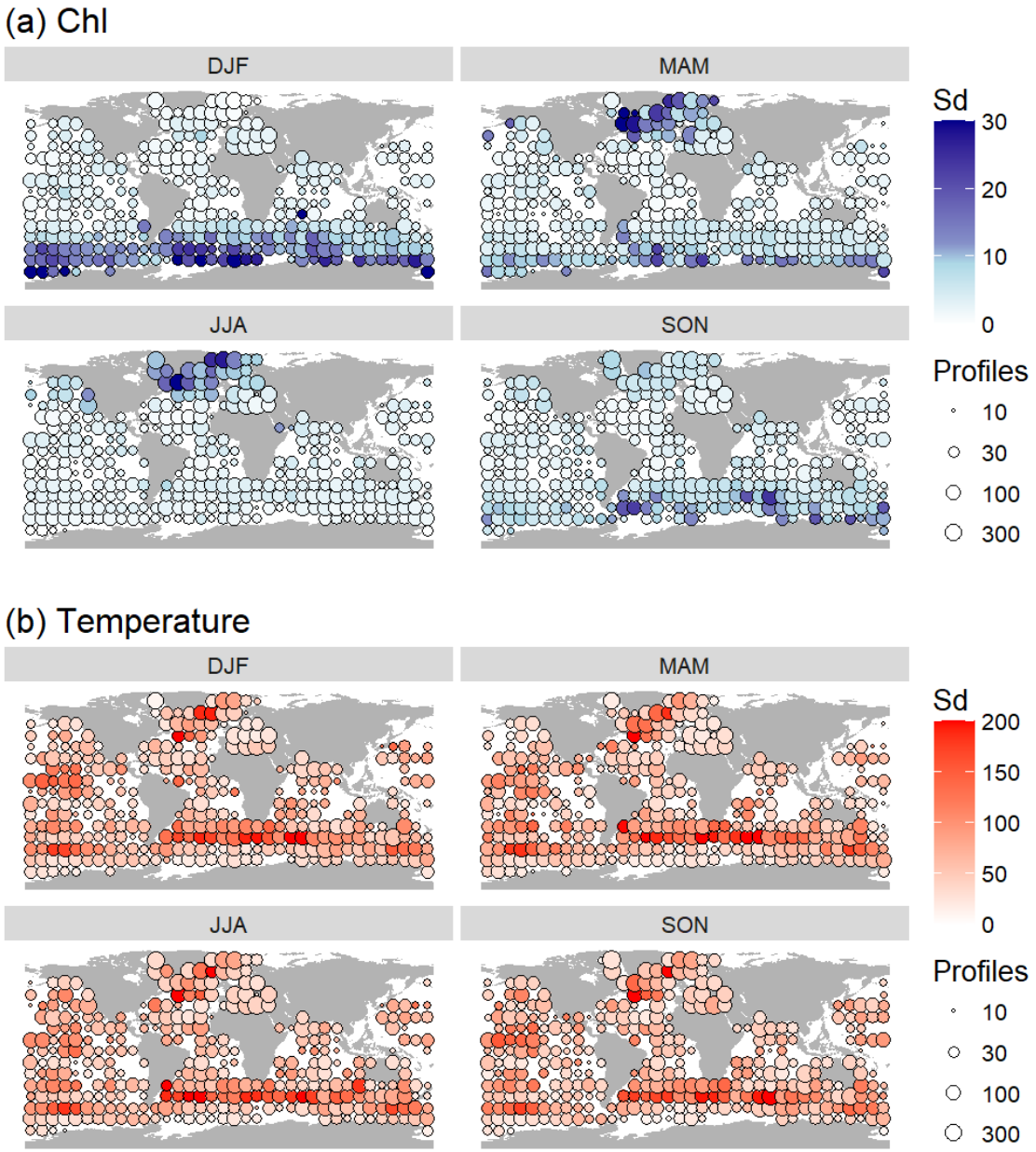

FIGURE 5.3: Maps showing the standard deviation of (**a**) chlorophyll profiles and (**b**) temperature profiles within each 10° grid cell for each season. Note that profiles were restricted to between 5 m - 250 m.

contrast, the float with the lowest variance in both chlorophyll and temperature profiles (WMO 1902385) was located in the subtropical North Atlantic Ocean and remained within a single biogeographical region.

The temporal autocorrelation of chlorophyll and temperature profiles from individual floats exhibited clear patterns (Figure 5.6). Autocorrelation functions (ACFs) from four floats (WMOs 1902385, 4903365, 6901767, and 5905107) displayed a clear annual cycle, with a sinusoidal pattern for both temperature and chlorophyll. WMO 1902385, located in the North Atlantic gyre, showed the ACF with the lowest amplitude, consistent with the weaker seasonality in temperature and stratification in this region.

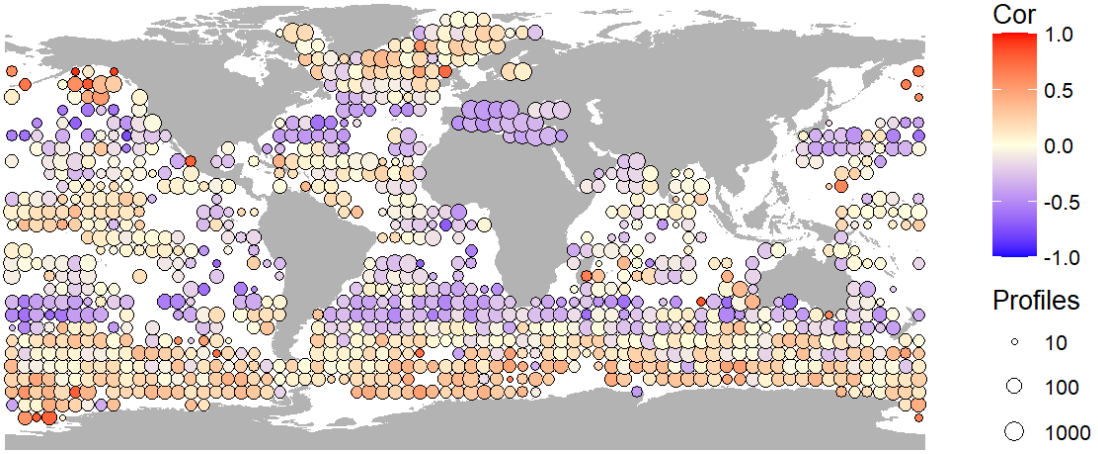

FIGURE 5.4: Map showing the correlation between chlorophyll and temperature profiles (5 m - 250 m) within a 5° grid cell. Larger dot sizes indicate a greater number of profiles in a grid cell. Grid cells with fewer than 10 profiles were ignored. The dots to the right of the map are correlations per latitudinal band.

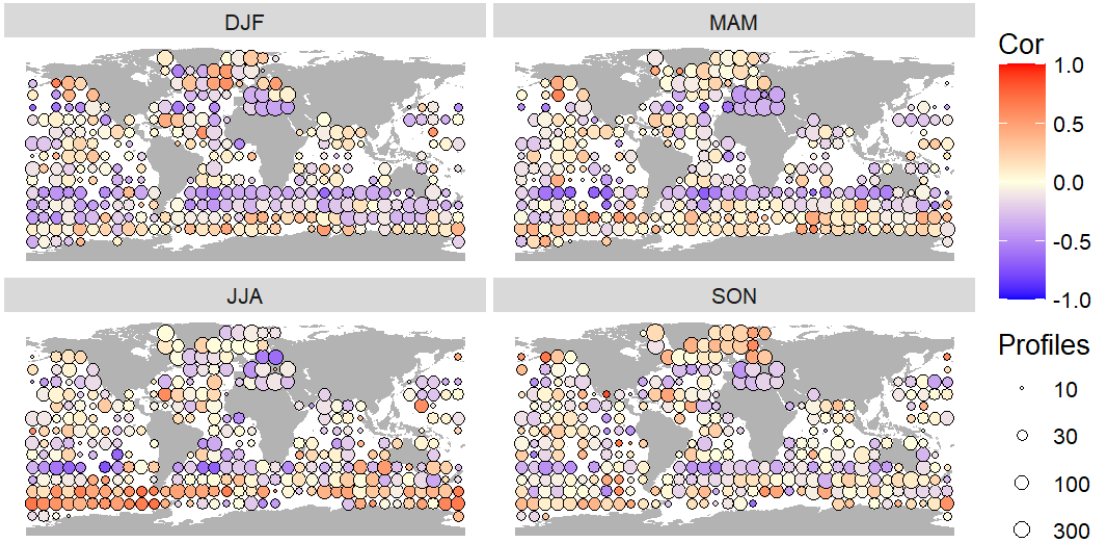

FIGURE 5.5: Maps showing the correlation between chlorophyll and temperature profiles (5 m-250 m) in each 10° grid for each season. Larger dot sizes indicate a greater number of profiles in a grid cell. Grid cells with fewer than 10 profiles were ignored. The dots to the right of the map are correlations per latitudinal band.

After a time lag of only 10 days, the autocorrelation of chlorophyll was only 0.4 compared to 0.81 for temperature, and generally, the amplitude of the temperature ACF exceeded that of chlorophyll, indicating that the annual cycle of temperature was more distinct. This suggests that chlorophyll profiles may respond more significantly to sub-seasonal or mesoscale effects. The ACFs of three floats (WMOs 5906204, 5904021, 6901585) did not exhibit a sinusoidal pattern. These floats travelled a substantial distance, sometimes over a range of latitudes and between contrasting marine environments. Consequently, in these cases the temperature ACFs remained above 0.5 for considerably longer than the seasonal cycle (up to a year for WMO 6901585). The time lag required to reach an autocorrelation of zero for chlorophyll and

Float ID	Mean lat (°N)	Lifespan (yrs)	No. of profiles	Var(Chl)	Var(Temp)	Cor(Chl,Temp)
1902385	25.6	2.93	108	0.5	1421	-0.04
4903365	57.1	3.37	124	102.7	2389	0.43
6901767	40.6	3.07	206	12.8	3567	-0.43
5905107	-34.0	4.20	139	9.0	1679	-0.40
5906204	-31.0	4.98	175	16.9	17564	-0.39
5904021	33.8	3.79	267	70.8	7602	-0.25
6901585	-54.8	2.83	225	120.1	4856	0.17

TABLE 5.2: Variance and correlation between chlorophyll and temperature profiles along a selection of seven BGC-Argo float trajectories.

temperature was typically around 90 days (for floats that stayed in the same region). Float 6901585 had the shortest decorrelation time scale for chlorophyll as it reached zero after only around 45 days.

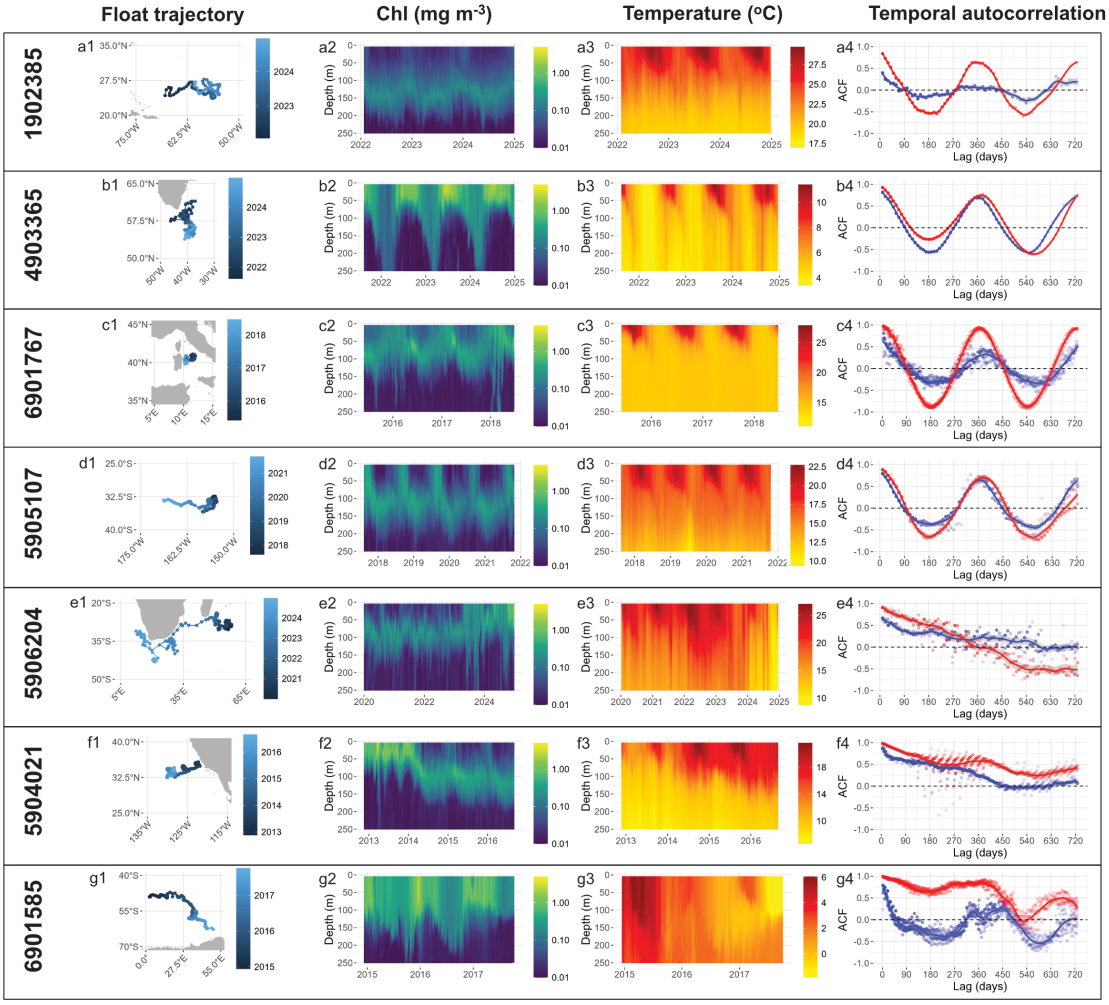

FIGURE 5.6: Temporal ACFs of chlorophyll and temperature profiles from a selection of seven BGC-Argo floats. **First column**: map of the trajectory of each float. **Second and third columns**: semi-Lagrangian sections over the floats' lifespan of chlorophyll, and temperature, respectively. **Fourth column**: smooth curves showing the temporal ACFs for temperature (red) and chlorophyll (blue). Point opacity is lower for lags with fewer pairings and the ACFs are weighted towards points with more pairs. Scattered points indicates more irregular sampling in time, indicating a higher quantity of less common lag times.

5.4.3 Autocorrelation of chlorophyll and temperature profiles from an Eulerian perspective

The Eulerian spatio-temporal autocorrelation structure (STAS) of chlorophyll profiles exhibits a substantial seasonal signal in all regions, except the STPS and EQU biomes (Figure 5.7). Figures 5.7a, d, g and j show substantially different STAS patterns, with temporal variability disappearing in the STPS and EQU biomes. In contrast, a seasonal cycle and spatial decay were observed in the Mediterranean. The spatial autocorrelation function is moderately similar across all biomes, initially decreasing with distance until approximately 500 km, after which it either decreases more slowly or fluctuates. There is a slight trend for the spatial autocorrelation function to decay

more rapidly in higher-latitude biomes. Temporal autocorrelation after six months and one year is approximately 0 and 0.5, respectively. In comparison, the STAS of temperature profiles shows less temporal variability relative to spatial decay (Figure 5.8), with the Mediterranean being the only exception. The spatial autocorrelation decay for temperature is longer than that for chlorophyll (Figures 5.8a, d, g and j). From Figures 5.7 and 5.8, I infer that, at the biome scale, chlorophyll profile shapes vary over shorter spatial and temporal distances than temperature profiles. Additionally, temperature variability across the biome is greater than that experienced at most locations over the annual cycle. In contrast, chlorophyll profiles at individual locations exhibit variability throughout the year that is more comparable to the variability observed across the biome at a given time.

Regardless of biome, the STAS of chlorophyll when using entire profiles matched most closely to those when using only a single depth near the surface (Figure 5.9a). The similarity to entire profiles decreased with depth, as did the disparities between biomes. In contrast, the STAS of temperature profiles was very closely related to the STAS of any single depth across all biomes except the Mediterranean (Figure 5.9b). These results suggested that most of the spatio-temporal variability of chlorophyll was confined to the top 50 m, whereas the spatio-temporal variability of temperature spanned a wider range of depths.

5.5 Discussion

5.5.1 Scalar variance and correlation of chlorophyll and temperature profiles

Prior to the first uses of functional data analysis on oceanographic datasets (Yarger et al., 2022; Korte-Stapff et al., 2022), the shape of chlorophyll and temperature profiles have been analysed by fitting various mathematical curves to observations (Carranza et al., 2018; Xu et al., 2022b), or from theoretical studies (Fennel and Boss, 2003; Beckmann and Hense, 2007). A benefit of using parameterised functions is the interpretation of variation between profiles, however those approaches lack an overall value of profile variability analogous to scalar-valued data. My work proposes a solution to this issue, which could be used in conjunction with parameterised functions to fully utilise the available data. The use of empirical orthogonal functions has allowed for the identification of the dominant modes of variability within profiles (Bock et al., 2022; Kuhn et al., 2025) which has the advantage that it conveys a sense of the overall variability and some interpretation of the main sources.

The regions with the highest chlorophyll variation were the high latitudes (Figure 5.3). Although surface blooms are common in these regions (Cornec et al., 2021a),

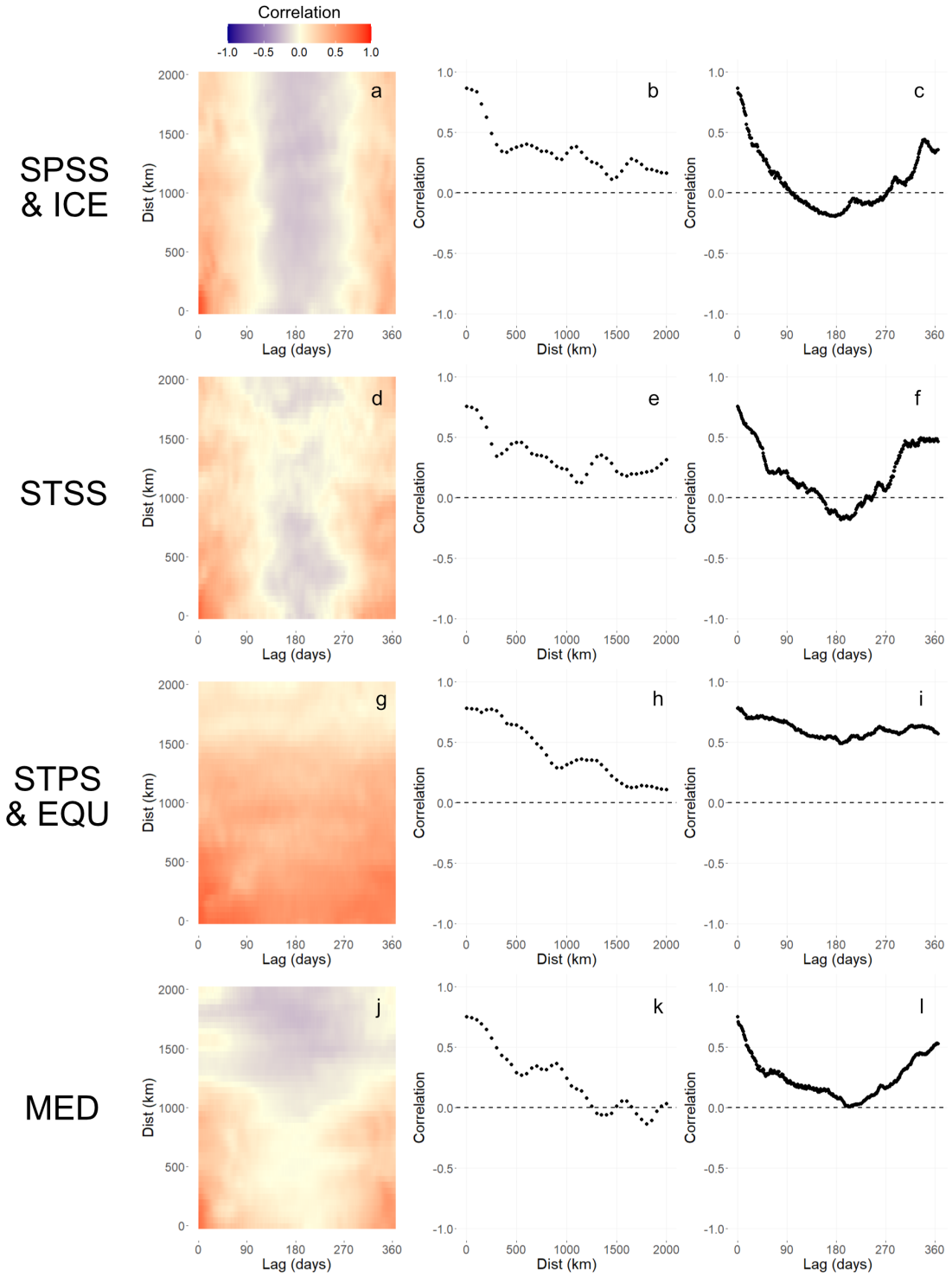

FIGURE 5.7: **(a, d, g, j)** Eulerian spatio-temporal autocorrelation structure (STAS) of chlorophyll profiles. **(b, e, h, k)** Spatial autocorrelation ($l = 0$) of chlorophyll profiles. **(c, f, i, l)** Temporal autocorrelation of chlorophyll profiles ($d_{GC} < 50$km).

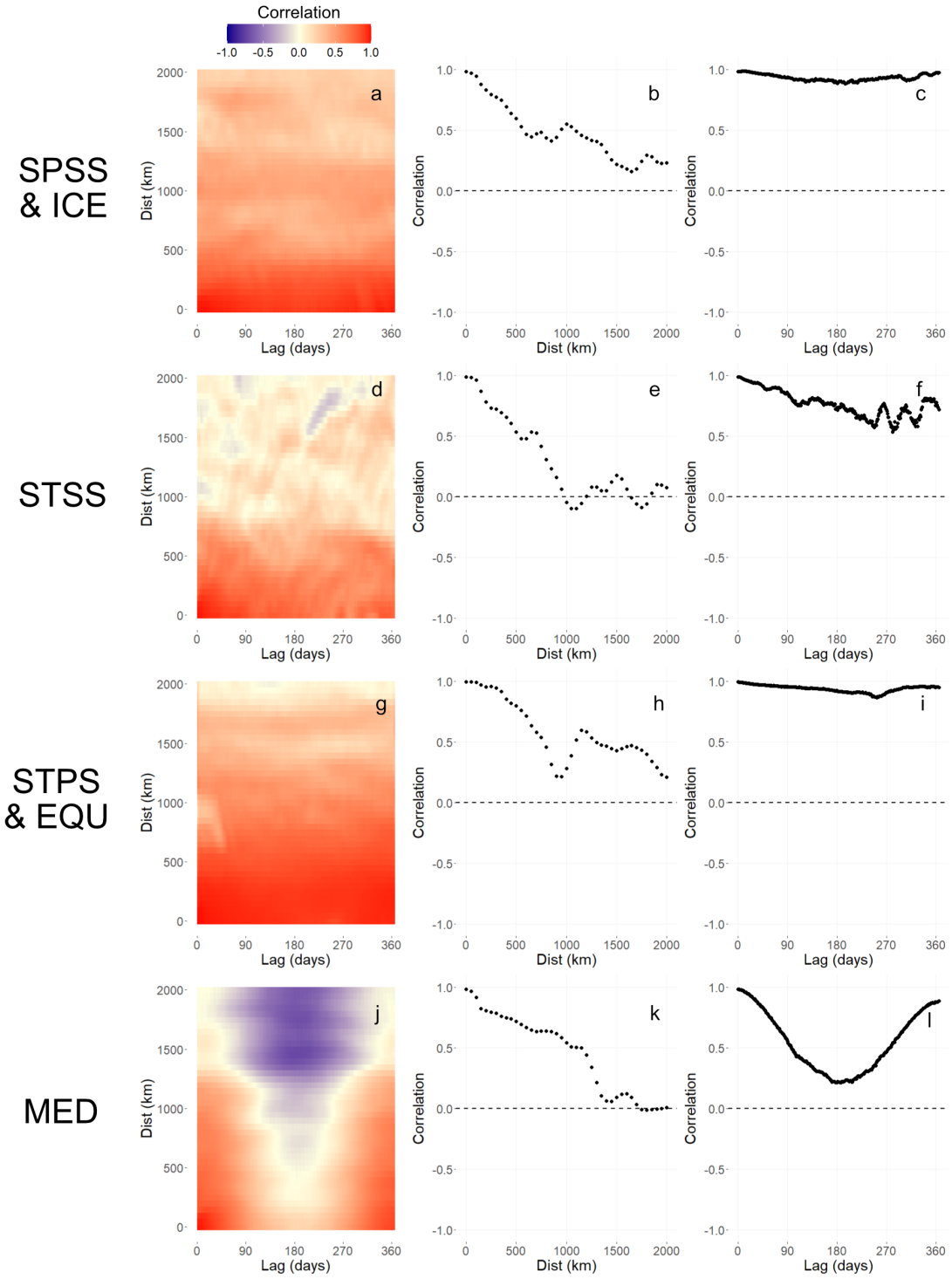

FIGURE 5.8: **(a, d, g, j)** Eulerian spatio-temporal autocorrelation of temperature profiles. **(b, e, h, k)** Spatial autocorrelation ($l = 0$) of temperature profiles. **(c, f, i, l)** Temporal autocorrelation of temperature profiles ($d_{GC} < 50$km).

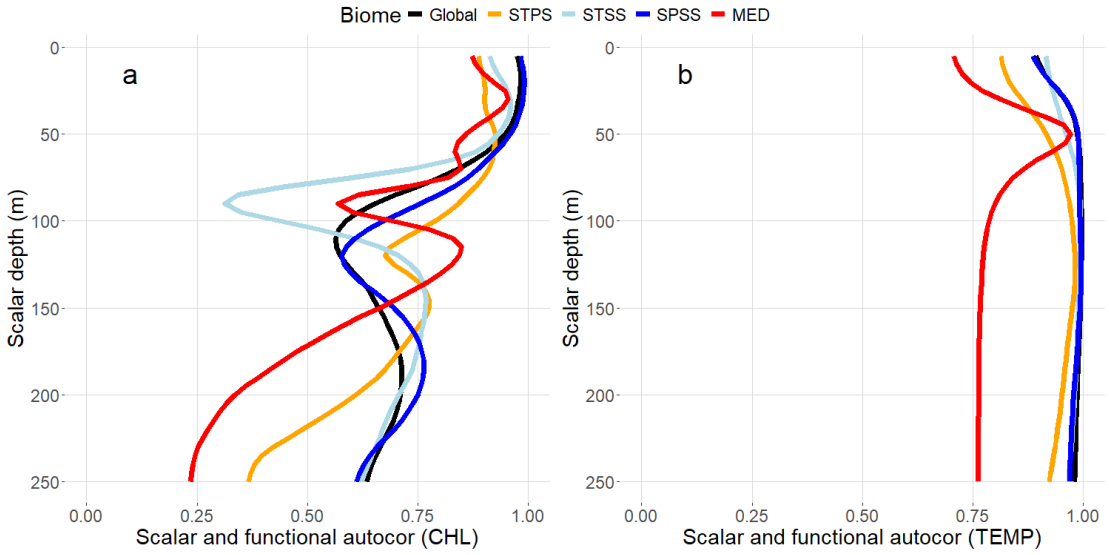

FIGURE 5.9: Correlation between spatio-temporal autocorrelation based on entire profiles and that derived from individual depths, for (**a**) chlorophyll and (**b**) temperature.

subsurface chlorophyll peaks have been regularly identified (Baldry et al., 2020; Boyd et al., 2024; Bouman et al., 2020). Therefore, it is possible that the variance could come from a variety of depths, depending on the small-scale mechanisms such as the presence of sea ice. I observed high variability in temperature profiles along the boundaries between the subpolar and subtropical regions (the transition zone in Henson et al. (2009)), which are areas in which ocean fronts and mesoscale eddies are common (Chapman et al., 2020). It is well understood that eddies alter the mixed layer depth which in turn has the effect of aiding or hindering (depending on eddy polarity) the injection of nutrients into the euphotic zone (Brannigan, 2016). Therefore, it might not be surprising that these are some of the areas with strongest correlations, although I expected a stronger relationship given the results from previous studies (Cornec et al., 2021b; Wang and Liu, 2024). This may be elucidated more easily using chlorophyll profile anomalies instead of the absolute profiles. Feng et al. (2015) performed a global comparison of SST amongst other variables with surface chlorophyll using satellite data and found that the effect of temperature on chlorophyll did vary by region. The BGC-Argo program has not been operational for long enough to establish climate-scale changes to chlorophyll profiles as a response to warming oceans.

5.5.2 Scales of variability

My study did not fully resolve any questions regarding the spatial and temporal length scales of chlorophyll and temperature profiles. I calculated the temporal autocorrelation of profiles in a semi-Lagrangian perspective but could only quantify variation over a seasonal temporal scale. In contrast, McKee et al. (2022) were able to

identify variation over mesoscale length scales by combining BGC-Argo and satellite data, as well as using chlorophyll anomalies. My restriction of using only BGC-Argo float locations meant that I could not estimate spatial length scales in this framework. However, Kuhn et al. (2023) showed that the differences in spatial scales of surface chlorophyll between Eulerian and Lagrangian frameworks are negligible so it could be possible to estimate the length scales sufficiently well using chlorophyll profile anomalies from multiple floats. My results address slightly different questions, specifically, which of the two variables has a higher seasonal predictability along a given trajectory? I found that, if a float remains in the same biogeographical region, temperature profiles display a stronger seasonal signal than chlorophyll profiles (Figure 5.6). In the case of float 6901585, the float did move a significant distance over its lifespan and consequently the temperature ACF did not show a seasonal cycle, but instead remained above 0.8 up to lags of a year before decreasing to zero. This was in stark contrast to the chlorophyll ACF, which had a rapid decay to zero. This may indicate that within a biome the seasonal variability in chlorophyll across the entire region is greater than at a particular location, whereas temperature varies more substantially across the region, rather than seasonally at any single location. This is only one example so no strong conclusion should be made, however it is a noteworthy result. I identified differences in the relative contributions of spatial and temporal variability of chlorophyll and temperature between biogeographical regions (Figures 5.7 and 5.8), with temporal variability increasing towards the poles. By treating the Mediterranean Sea as a separate biome, I found that to be the only region which had significant variability in both space and time, with similar patterns in temperature and chlorophyll (although the latter had a weaker signal). This combination of spatial and seasonal variation probably derives from the differences in physical oceanography (Schroeder et al., 2023) and biogeography (Lavigne et al., 2015) at either end of the Mediterranean.

Given that I found the STAS of chlorophyll and temperature are very similar regardless of whether profiles or surface values are used, this supports the use of neural networks that combine spatio-temporal data from satellite observations with Argo profiling data (Sauzède et al., 2016; Meng et al., 2021; Hu et al., 2022a; García-Jimenez et al., 2025). My results suggest that the spatio-temporal variability captured by the satellite should be a fair representation of the variation at depth, which is of significance in particular for oligotrophic regions, where the highest chlorophyll concentrations are well beyond the view of satellites. However, this does not mean that satellites are detecting variability in SCMs, but rather that they measure differences in the corresponding surface concentrations.

5.5.3 Limitations

One of the drawbacks of using the scalar variance and correlation is that the coefficient does not provide any information about the variations in shape or the depths at which they occur. For example, although I have found moderate correlations between chlorophyll and temperature profiles, my output does not indicate whether deviations from a mean profile are near the surface or deeper in the water column. Consequently, fully understanding these correlation coefficients might require some prior understanding of the processes potentially underpinning the variation or further investigation. Another limitation is that, as with correlation for scalar data, the correlation coefficient is a measure of linear dependence. Therefore, this method will not detect a non-linear dependency between variables. In this scenario, a functional regression model with a non-linear covariate effect (similar to the methodology of Chapter 4) might be a viable alternative. Additionally, the BGC-Argo array is relatively sparse and clustered in space, which may lead to some biases in autocorrelation length scales. Finally, when calculating temporal autocorrelation functions from a Lagrangian perspective, it is important to have regular temporal sampling in order to avoid adding bias into the autocorrelation of a specific lag. Out of 890 BGC-Argo floats in this dataset, relatively few (< 100) had regular enough sampling gaps for easily interpretable ACFs.

5.5.4 Future work

There might be potential for utilising this approach across a range of applications in oceanography. Firstly, incorporating the variance of profile shapes within methods for optimising observing system design (Chamberlain et al., 2023; Chu et al., 2024) could be beneficial since differences in profile shape may be associated with underlying mechanisms. These approaches typically involve identifying combinations of Lagrangian trajectories from which the maximum amount of spatio-temporal variability is explained. It is reasonable that these methods could absorb the scalar valued variance from entire profiles. A second application could be the calibration of sensors onboard multiple platforms. For example, comparing measurements between high frequency glider profiles and long-term observations from moorings or Argo floats. I did not explore spatial autocorrelation in a Lagrangian framework. The calculation of this could allow for a dimensional analysis (similar to McKee et al. (2022)), although that would require the calculation of chlorophyll anomalies. As the abundance of oceanographic profiles increases, a range of functional data techniques including the one presented here could form a new statistical toolkit for oceanographers, particularly in the context of exploratory data analysis. Alternatively, this approach could be trialled on profiles from ocean models to assess how well they reproduce real world variability.

5.6 Conclusions

To my knowledge, this is the first application of scalar variance and correlation for functional data in oceanography. By treating ocean profiles as functional data, I can capture the essence of the shape of profiles, which often provides insights into the processes affecting the water column. This technique enables us to investigate variability across a range of depths measured by observing systems such as the BGC-Argo float array in the same way as scalar valued data. Here I applied this approach to 98 413 chlorophyll and temperature profiles across the global ocean and analysed their variance and correlation over different scales.

My results highlight the importance of considering spatio-temporal scales when assessing variability and correlation. I provide the first global maps of the variance of chlorophyll and temperature profiles (Figure 5.3), and their correlation (Figure 5.4). These results confirm that chlorophyll profile variability increases towards higher latitudes, whereas maximum temperature profile variability occurs near fronts between water masses, particularly in the Southern Ocean and North Atlantic. At the biome scale, from an Eulerian perspective, temperature profiles vary more spatially than seasonally, whereas for chlorophyll profiles, temporal variability dominates (Figures 5.7 and 5.8). This suggests that variation among chlorophyll profiles within a single biome is primarily driven by time (increasingly with distance from the equator), whereas temperature profile variation is more strongly influenced by space, largely through latitudinal gradients. Conversely, when changing to a semi-Lagrangian perspective and following individual floats along their trajectories, temperature profiles exhibit a stronger seasonal autocorrelation signal than chlorophyll profiles (Figure 5.6). However, this analysis also indicated that floats drifting between biogeographical regions do not produce sinusoidal autocorrelation functions, making the estimation of temporal decorrelation scales ineffective with this approach. Finally, I found that chlorophyll and temperature profiles are typically weakly positively correlated over small spatial scales. Stronger positive correlations occurred at high latitudes during winter, while moderate negative correlations were observed in the latitudinal band between 30° and 40° in each hemisphere.

As the quantity of oceanographic depth profiles increases, developments in the field of functional data analysis offer an opportunity for statistical analyses from an alternative perspective that focusses on the shape of profiles. I see potential for further use in a variety of applications, including observing system optimisation and the calibration of sensors across multiple platforms.

Chapter 6

Synthesis

6.1 Summary of results

6.1.1 Vertical chlorophyll distribution

The spatio-temporal distribution of SCMs has garnered renewed attention in recent years (Mignot et al., 2014; Cornec et al., 2021a; Yasunaka et al., 2021) due to the widespread deployment of BGC-Argo floats. In Chapter 3, I identified large-scale spatial and seasonal patterns in SCM characteristics using a spatio-temporal modelling approach. This provided evidence that the z_{eu} was positively correlated with z_{SCM} and negatively correlated with Chl_{SCM} (Figure 3.3). This also indicated that MLD and zooplankton biomass had small but significant effects on both SCM depth and intensity. Only z_{SCM} was affected by sea surface height anomaly (SSHA) (i.e., downwelling eddy features) which suggested that positive SSHAs resulted in deeper SCMs which was the opposite to previous studies (Cornec et al., 2021b; Xu et al., 2022b). Nevertheless, maps of predicted Chl_{SCM} (Figure 3.7) and z_{SCM} (Figure 3.8) closely resembled previous studies (Cornec et al., 2021a; Yasunaka et al., 2021; Masuda et al., 2021), with the deepest and least intense SCMs occurring in the subtropical oligotrophic gyres, especially the South Pacific subtropical gyre.

In Chapter 4, I found that light availability (or more specifically, the z_{eu}) was the primary driver of z_{SCM}, whereas the z_{ncline} was coupled to the peak in phytoplankton biomass. Year-round predictions of chlorophyll and b_{bp} profiles were produced for the subtropical and tropical ocean. This revealed that photoacclimation (represented by the $Chl:C_{phyto}$ ratio) was an important mechanism across a range of latitudes, with peaks in photoacclimation coinciding with peaks in chlorophyll (Figure 4.7). Particularly at the equator, where the most intense SCMs were predicted, photoacclimation was highest. Seasonal variability in photoacclimation (Figure 4.8) was predicted to have elevated $Chl:C_{phyto}$ in the summer at latitudes more than 30°

from the equator. This analysis provides evidence that phytoplankton biomass (represented by b_{bp}) is typically restricted to above the nitracline and reduces significantly for deeper nitraclines (Figure 4.8).

In Chapter 5, the focus shifted from assessing the drivers of profile variability to demonstrating a new measure of variability and correlation for profiling data. This highlighted the seasonal differences in the variability of chlorophyll profiles when compared to temperature profiles (Figure 5.3). Higher latitudes (more than 40° from the equator) displayed the highest variability in chlorophyll profiles, especially during the summer, possibly reflecting the presence of elevated chlorophyll in some profiles. In contrast, temperature profiles varied most along the boundaries of water masses such as the Antarctic Circumpolar Current. I identified a weak to moderate correlation between temperature and chlorophyll profiles (Figure 5.4), although this varied substantially by region and season. For example, the two variables were negatively correlated year-round in the mid-latitudes, whereas the correlation strength varied seasonally at high latitudes (Figure 5.5).

In summary, results from this work indicate that light availability is the primary driver for vertical chlorophyll distribution, although nutrient availability determines the depth of biomass accumulation, which is typically located shallower than the z_{SCM}. Consequently, photoacclimation is an important mechanism through which SCMs form and are maintained. The relationship between chlorophyll and temperature profiles varies regionally, suggesting that stratification is more important in some locations than others in determining phytoplankton abundance.

6.1.2 Statistical methods for BGC-Argo float data

In Chapter 3, the spatio-temporal modelling method demonstrated that using a spatio-temporal latent variable improved the predictive ability of models for SCM characteristics. Specifically, the RSMEs for z_{SCM} and Chl_{SCM} reduced by 21% and 33% respectively. Furthermore, several of the covariate coefficient values changed substantially after including the latent effect, with some changing sign. This suggests that the latent effect explained a significant portion of the variability. The model for predicting z_{SCM} was less successful than for Chl_{SCM}. This may be explained by the sensitivity in estimates of z_{SCM} for profiles without a clear peak, as this can lead to a situation where two very similar profiles could have significantly different z_{SCM} estimates. Such profiles could take the form of sigmoids as described by several previous studies (Carranza et al., 2018; Brewin et al., 2022) or with very low chlorophyll concentrations throughout the water column.

In light of this, I utilised methods from functional data analysis (FDA) for the remainder of the thesis. In Chapter 4, I fitted two functional regression models (FRMs)

to assess how environmental conditions affected chlorophyll and b_{bp} profiles. I found that the models were better at reproducing SCM characteristics when the covariate effects were implemented as non-linear scalars (i.e., each covariate value had a smooth functional effect associated with it), rather than linear scalar (in which a function was added after some scaling) (Table 4.1). In Chapter 5, I demonstrated the first application of scalar variance and correlation of oceanographic profiles treated as functional data. My results also showed the importance of study region scale, as length scales of chlorophyll where shorter from the semi-Lagrangian perspective of a single float (Figure 5.6), whereas the opposite was true on biome scale in an Eulerian framework (Figures 5.7 and 5.8).

Overall, I found that FDA techniques aid the analysis of Argo float profiles. The major benefit of these approaches is that they bypass the identification of profile features (such as SCMs) prior to analysis and do not ignore the remainder of the profile in later analyses. Another key motivation is communicating the idea of connectedness between measurements within a single profile. For example, the presence of particles higher in the water column reduces light availability below, and similarly nutrient consumption deeper in the water column reduces the availability higher up. Given that this dependency between all depths is fundamental to the theoretical modelling of phytoplankton (Fennel and Boss, 2003; Beckmann and Hense, 2007; Gong et al., 2015), it seems appropriate to treat profiles in a similar way during statistical analyses.

6.2 Wider implications

6.2.1 Predictability of chlorophyll profiles

Given the relatively sparse and clustered nature of subsurface chlorophyll profiles, there is motivation to interpolate between observations using other oceanographic variables as predictors. Previous work used a variety of predictor variables to predict characteristics of SCMs primarily the z_{SCM} (Xu et al., 2022b; Miyares et al., 2024) but also the entire profile (Uitz et al., 2006). In this thesis, I investigated which environmental conditions affect the shape of chlorophyll profiles and found that the z_{eu} was the biggest influence. The non-linear model tested in Chapter 4 shows that large-scale chlorophyll profile variability can be reproduced using functional effects of only three covariates: z_{eu}, z_{ncline} and MLD. Even here, the MLD effect was relatively small compared to the other two effects so it is reasonable to suggest that MLD could be removed from the model without losing much predictability.

Machine learning approaches can utilise satellite and BGC-Argo float data to interpolate full-depth bio-optical profiles across over large spatio-temporal scales and at high resolution (Sauzède et al., 2016). My results suggest that these approaches

could benefit from using light and nitrate to predict profiles as well as location and time. I also found that the spatial and temporal scales of variability of chlorophyll profiles are very similar to those of chlorophyll concentration at the surface (Figure 5.9a). This provides further support for the machine learning techniques combining satellite and BGC-Argo float data. Another option is to implement machine learning with functional data as inputs and outputs.

6.2.2 Community composition and the marine ecosystem

Previous research has identified that different phytoplankton taxonomic groups favour specific environmental conditions (Latasa et al., 2017; Sato et al., 2022; Latasa et al., 2023). This partitioning of the water column allows multiple groups to exploit conditions for which they are better adapted. Although the BGC-Argo float dataset does not contain information about community composition, Brewin et al. (2022) highlighted this phenomenon by detecting periods when two communities thrived, and how this occurred at different depths. The results in this thesis support the idea that light drives the SCM depth, and potentially different phytoplankton species' vertical distribution (Sato et al., 2022; Latasa et al., 2023). Some phytoplankton functional groups favour low light conditions, by having a high Chl:C_{phyto} ratio. Sato et al. (2022) note that the relative abundances of phytoplankton species do not vary considerably over large-scales and this may reflect the importance of photoacclimation across a range of latitudes, as identified in Chapter 4.

Phytoplankton form the foundation of the marine ecosystem and consequently their abundance and distribution affect higher trophic levels, including their consumers, zooplankton, whose biomass was included as a predictor variable in the models in Chapter 3, although the results suggest that it actually had little predictive power. The initial rationale was that the presence of zooplankton near the surface could deepen and weaken the SCM, as modelled theoretically by Moeller et al. (2019). In practice, however, this was not easy to infer from the model since the zooplankton biomass data I had access to were integrated over the entire water column. Instead, my results indicated that high zooplankton biomass was associated with high phytoplankton abundance, as expected. It is worth noting that the z_{SCM} was deeper in the presence of elevated zooplankton biomass, although this cannot be interpreted as a causal effect due to the aforementioned issues.

6.3 Potential improvements to statistical methods

This thesis has presented a variety of statistical methods to apply to BGC-Argo data, though each approach has scope for change and improvement. I will highlight some

potential modifications to these methods for future applications.

6.3.1 Spatio-temporal modelling

The stochastic partial differential equation (SPDE) methodology presented in Chapter 3 incorporated a non-stationary correlation structure in space through the hard barrier condition similar to the one proposed by Bakka et al. (2019). This was one of several potential ways to include non-stationarity, with anisotropy another option, in which the correlation length scales at a location are larger in one direction, meaning curves of constant correlation are ellipses rather than circles, and this direction can change with location or time. There are several examples of this being applied in two dimensions (Fuglstad et al., 2015; Tomasetto et al., 2024) and three dimensions (Pereira et al., 2022; Berild and Fuglstad, 2023). Anisotropic correlation structures allow for the idea of flow to be added into a covariance structure (Tomasetto et al., 2024; Berild and Fuglstad, 2023). This is highly relevant for the ocean as there are distinct features of global ocean dynamics, such as gyres, boundary currents, and fronts, that create regions of similarity that align along different directions, often following latitudinal bands or oceanographic structures. As a result, spatial covariance of chlorophyll is often anisotropic and non-stationary, meaning that the correlation structure changes in location or time. This has been applied to SST satellite data (Hu et al., 2022b). The 3D version has been applied to oceanographic data on a local scale (Berild and Fuglstad, 2023), however it may not be as useful when horizontal length scales exceed vertical length scales by several orders of magnitude, as in the case of global Argo float data.

An alternative extension could involve treating depth as an autoregressive component (like I did for time) so that data over multiple depths can be included whilst maintaining their connectivity although this may be computationally challenging. Covariate values can also be used to introduce non-stationarity into the correlation structure (Ingebrigtsen et al., 2014), however this may have fewer benefits than anisotropy in the context of the work in this thesis.

I only conducted a brief preliminary investigation into the effect of the mesh size (in terms of the number of vertices). It is possible that the computation speed may improve with a reduction in mesh size without substantially affecting the model fit, which might also have been improved by fitting a bivariate model with z_{SCM} and Chl_{SCM} as concurrent response variables. However, this would also have increased the computational cost. This may be possible using the **INLA** R package, but it is not guaranteed to improve predictive performance. In the future quantum computing may allow for more complex models to be fitted without the drawback of long computation times, further strengthening the case for using spatio-temporal modelling on large oceanographic datasets.

6.3.2 Functional regression models

The work presented in Chapter 4 did not consider using profiles of other biogeochemical parameters as functional covariates (e.g. nitrate or PAR profiles), which could have aided the prediction of bio-optical profiles. However, this might reduce the interpretability of the model. In addition, the proposed FRMs had no spatio-temporal component within the model structure, which could capture dependencies between observations, as was done by Yarger et al. (2022) and Korte-Stapff et al. (2022). However, including such a component for an FRM would significantly add to the computational cost and would require more advanced statistical coding to implement, especially with non-linear effects as I believe there are currently no publicly available methods to fit such a model. Alternatively, using a different set of smaller biogeographical regions could have yielded different results, for example, the Longhurst regions (Longhurst et al., 1995). However, this was not appropriate due to the relatively small number of profiles completed within two hours of noon and equipped with bio-optical, nitrate and irradiance sensors. A possible extension could involve fitting a multivariate FRM to concurrent chlorophyll and b_{bp} profiles, rather than fitting two separate models with the same structure. Could this capture the dependence between the two variables? It may be worth considering a functional regression model with satellite-derived surface chlorophyll concentration as a covariate. This might be more beneficial for mapping (i.e., spatial prediction), rather than for a mechanistic understanding. It is also possible that the best combination of covariates was not tested, i.e., potentially a subset of the covariates would have been an improvement. Therefore, a more comprehensive assessment of different combinations of covariates and covariate types might yield better results.

6.3.3 Scalar variance and correlation of oceanographic profiles

As mentioned before, the shape of an oceanographic profile (how many peaks and troughs occur, and at which depths) is sometimes the primary interest of researchers. In this work, I provided the first use of scalar variance and correlation for oceanographic profiles as functional data objects. Urbano-Leon et al. (2023) suggested that there was potential for their approach to be developed further in order to create a range of novel statistical tools for functional data, analogous to standard scalar-valued techniques. The usefulness of statistical tests and suchlike in the context of oceanographic profiles would be exciting to explore. For example, an analysis-of-variance for oceanographic profiles - to partition variance between natural (or residual) variability and the effect of an external factor. Given that so many oceanographic profiles are collected each year, it would be beneficial for the community to develop statistical methods suitable for functional profiling data in order to better understand the scale and causes of profile variability.

Combining the Argo data with geostrophic velocity data could have allowed the estimation of spatial decorrelation length scales between profiles in a similar way to Kuhn et al. (2023). However, those authors did not find a significant difference between Eulerian and Lagrangian perspectives, so this might not yield any additional insights. I also highlighted the fact that the sampling frequency of BGC-Argo floats affects the estimation of temporal autocorrelation functions. In the case of irregular sampling, estimates are very sensitive and produce noisy ACFs. I recommend performing analyses after filtering profiles such that only the most regularly sampled time lags are used in the ACF calculation.

6.3.4 Combining the three approaches

There are aspects of each method that could be adopted within a subsequent piece of work assessing spatio-temporal patterns in subsurface chlorophyll from the BGC-Argo float dataset. Such an approach could contain the following characteristics:

- A functional spatio-temporal model building on the work by Yarger et al. (2022) and Korte-Stapff et al. (2022).
- Covariates describing the light and nutrient fields (either scalar or functional) measured simultaneously with chlorophyll.
- Non-stationary correlation structure displaying anisotropic structures as well as the barrier condition. This structure could be estimated using the scalar correlation approach described in Chapter 5.
- Potential for extending to bivariate model by including b_{bp} as a second response variable in order to study the importance of photoacclimation.

Developing an approach of this structure may have taken significant time and was not attempted in this thesis.

6.4 Future applications in marine biogeochemistry

6.4.1 Subseasonal variability in subsurface chlorophyll distribution

In this thesis I only used profiles of chlorophyll from BGC-Argo floats, which typically sample the water column once every ten days so the primary timescale under consideration was seasonal. However, other platforms are capable of sampling at higher temporal resolutions, including gliders (Thomalla et al., 2017; Carvalho et al., 2020) and biologging with sensors carried by marine mammals (Carranza et al., 2018;

Le Ster et al., 2023). It could be interesting to repeat the analysis carried out in Chapter 4 using profiling datasets collected over smaller spatial and temporal scales. This could help extract the effects of environmental conditions, especially if they are short-lived or difficult to detect with infrequent measurements. Variability in SCMs related to small-scale physical processes, such as fronts and eddies have already been studied (Cornec et al., 2021b; Xu et al., 2022b; Strutton et al., 2023), although it may be beneficial to apply a formal statistical model to assess changes to profile shape.

6.4.2 Detecting climate trends in subsurface chlorophyll

Monitoring global trends in chlorophyll concentration is important for understanding the impact warming oceans may have on phytoplankton abundance, both for biogeochemical processes and for marine ecosystems. Previous research suggests that several decades of observations are required to identify trends in surface chlorophyll (Henson et al., 2010, 2016). Hammond et al. (2020) proposed a spatio-temporal model for assessing trends in satellite-derived chlorophyll data which pooled observations over space and time to extract seasonal and decadal variability. This approach was found to be more sensitive to trends than generalised least squares regression in the 20-year time series used. Records of surface chlorophyll are longer than those of subsurface chlorophyll at most locations. An additional barrier to detecting trends in subsurface chlorophyll is that most subsurface observation platforms (such as Argo floats) move over time, and consequently, locations are rarely resampled multiple times, especially over the time scales required to identify trends. Although the spatio-temporal model used in Chapter 3 accounted for this through a spatio-temporal latent effect, it is not guaranteed that a long-term signal could be identified given the high spatio-temporal variability of chlorophyll. Therefore, it may be some time before a subsurface chlorophyll version of the study by Hammond et al. (2020) could be of use, possibly using a functional spatio-temporal framework similar to those developed by Yarger et al. (2022) and Korte-Stapff et al. (2022). Even when this is attempted, it may only be able to make vague inferences about chlorophyll concentration change over large-scale (e.g. latitudinal variation in trends) due to the sparsity of profiling data. Longer-term trends in profile shape (such as those collected by Hawaii Ocean Time-series) could be identified using FDA techniques, although that was beyond the scope of this thesis.

The ongoing monitoring of subsurface chlorophyll on a global scale relies on the continued deployment of autonomous platforms like BGC-Argo floats. There are aims to have a network of 1000 BGC-Argo floats, each carrying all six key biogeochemical variables (Owens et al., 2022; Thierry et al., 2025). A variety of projects deploy the floats, each with slightly differing objectives and regions of interest. For example, the SOCCOM project deployed a high number of floats in the Southern Ocean.

Consequently, climate trends may be more easily identified (or identified earlier) in some regions than others. This is without considering that the time series length required to detect a long-term trend is thought to vary regionally (Henson et al., 2016).

6.4.3 Community composition of phytoplankton

The relative abundances of different phytoplankton groups varies across space and time, including across depths (Uitz et al., 2006; Ward et al., 2014; Sato et al., 2022; Miyares et al., 2024). The chlorophyll data used in this thesis do not provide information about the corresponding phytoplankton species composition. Such datasets are primarily collected through ship-based research, which naturally reduces the spatio-temporal coverage of observations. Vertically-resolved phytoplankton datasets can be constructed through a variety of methods including high-performance liquid chromatography (Uitz et al., 2006; Latasa et al., 2017; Miyares et al., 2024), flow cytometry (Sato et al., 2022) and radiometry (Bracher et al., 2020). Datasets of this type have generally shown that smaller phytoplankton dominate deeper SCMs (Uitz et al., 2006), although Miyares et al. (2024) found that one size class dominated all environments. These datasets could be interesting to view from an FDA perspective as either the absolute biomass for a given taxonomic group, or the proportion of the total across all groups. In particular, it would be interesting to address the following questions. 1) How similar are the profiles of taxonomic groups, and their aggregation, to their combined chlorophyll profile? 2) How does the profile of each taxonomic group relate to light and nutrient fields? The methods from Chapter 5 and Chapter 4 could be used to address these two questions respectively. It is worth noting that Latasa et al. (2017) analysed profiles of phytoplankton biomass as functions of the fraction of surface light intensity, instead of treating profiles as functions of depth. This could be an interesting technique to try using the BGC-Argo float profiles that have an irradiance sensor.

6.4.4 Calibration of sensors across multiple platforms

In Chapter 5, I demonstrated the use of scalar variance and correlation for oceanographic profiles using the approach developed by Urbano-Leon et al. (2023). This has potential to be used to calibrate sensors between multiple platforms. This could take the form of multiple platforms measuring the same watermass concurrently to assess variability across sensors. Alternatively, it could be used to quantify drift of sensors over time, and whether this is dominated by drift at certain depths or is uniform across profiles. A further application could be calibrating time series from multiple platforms observing at different temporal resolutions. For example calibrating measurements between an Argo float sampling the water column

once every 10 days and a nearby glider completing a profile multiple times a day. Given the variety of oceanographic observing techniques and platforms (Chai et al., 2020), it seems reasonable to suggest that it may become common to integrate multiple platforms over a range of spatial and temporal scales and this statistical technique might prove useful. This would be applicable to all oceanographic variables measured in profiles, but chlorophyll would be an interesting example given the variety of profile shapes observed.

6.4.5 BGC-Argo float deployment optimisation

Given the limited (but growing) number of BGC-Argo floats, careful consideration should be given to gaining as much scientific understanding about the global ocean as possible. Several methods have been proposed that aim to maximise the explained variability of an oceanographic variable, by sampling regions with irregularly positioned floats (Mazloff et al., 2018; Chamberlain et al., 2023; Chu et al., 2024). BGC-Argo floats move over time and the method described by (Chamberlain et al., 2023) accounts for this. Measuring variance and autocorrelation length scales of profiles using the method by Urbano-Leon et al. (2023) might be advantageous by combining previous optimisation approaches with the scalar correlation measure presented in this work.

However, in practice, the locations and timings of float deployments are restricted by the routes of research cruises and by the requirements of the project through which they were purchased. Consequently, until recently, the BGC-Argo floats remained quite clustered in space. Projects such as Global Ocean Biogeochemistry array (Matsumoto et al., 2022) aim to deploy floats in regions that have historically been undersampled and could profit from utilising the method demonstrated in Chapter 5.

6.5 Final outlook

This thesis has explored distinct but complementary statistical approaches for analysis of data from BGC-Argo floats, each providing advantages in the study of biogeochemical data. First, spatio-temporal modelling allows for the quantification of large-scale trends and patterns in key variables such as chlorophyll concentration and the structure of subsurface maxima, while accounting for dependencies between nearby observations. In contrast, FDA provides a natural framework for modelling vertical profile data as continuous curves and identifying variability across depth. Although I applied these methods separately in this work, looking ahead, I suggest integrating these, as done by Yarger et al. (2022) and Korte-Stapff et al. (2022), in order to gain a more holistic understanding of the three-dimensional variability in ocean

ecosystems and the effects of different environmental processes. As the quantity and resolution of subsurface biogeochemical observations increases with the deployment of autonomous platforms, the development of statistically robust and ecologically interpretable tools will be vital for conducting effective analysis and producing meaningful scientific insight.

Appendix A

Mapping Global Subsurface Chlorophyll Maxima Characteristics using Argo Floats and Spatio-temporal Models

Additional figures supporting the analyses in Chapter 3 are presented here.

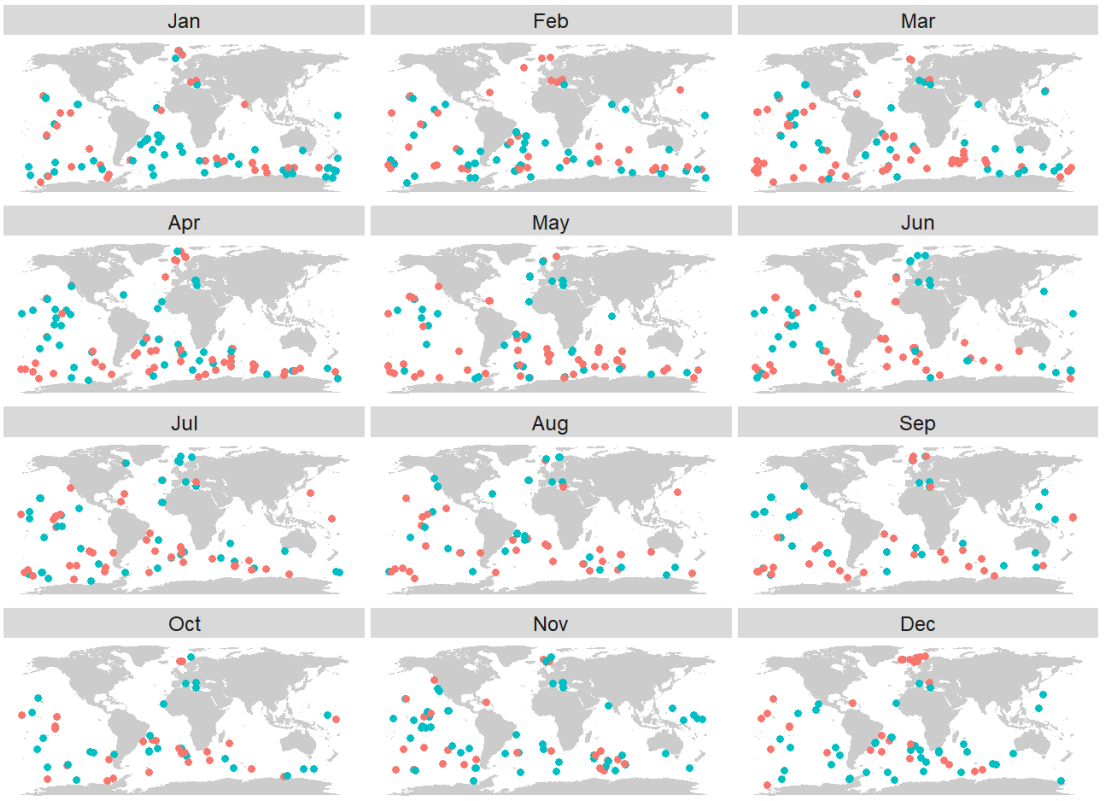

FIGURE A.1: Monthly maps showing which of the observations of z_{SCM} from the validation dataset were successfully within the 95% prediction interval (blue), or not (red) using the spatio-temporal model.

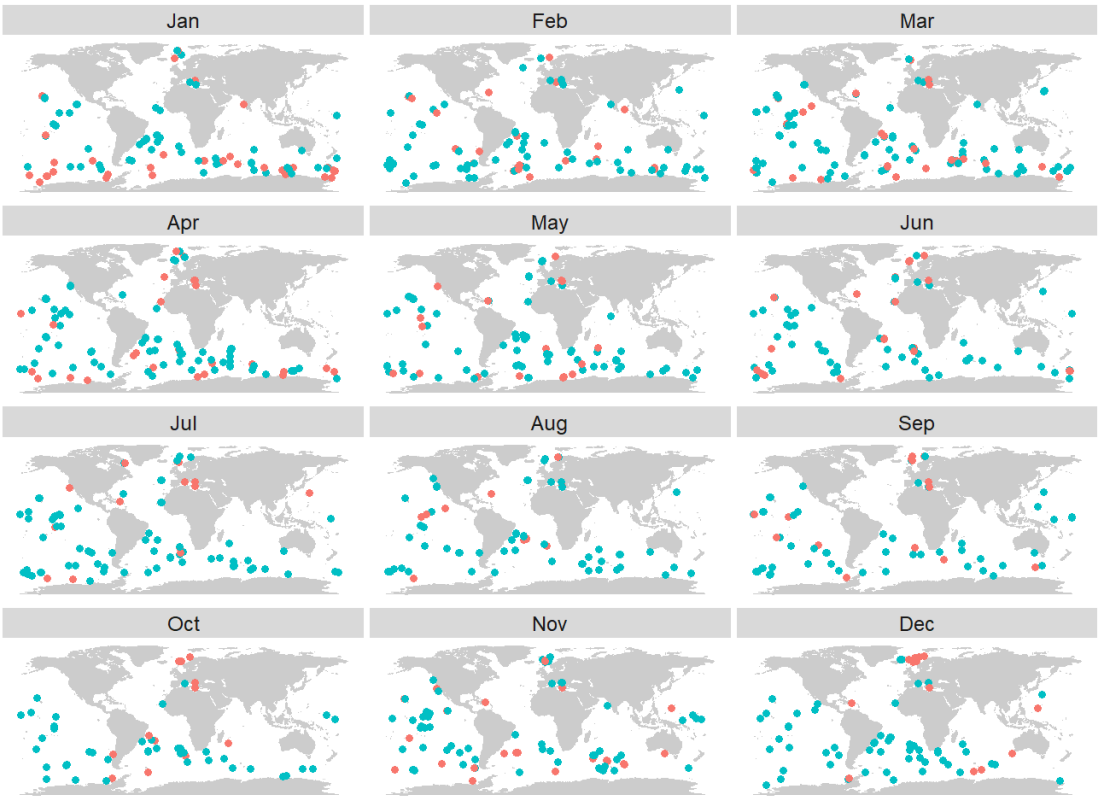

FIGURE A.2: Monthly maps showing which of the observations of Chl_{SCM} from the validation dataset were successfully within the 95% prediction interval (blue), or not (red) using the spatio-temporal model.

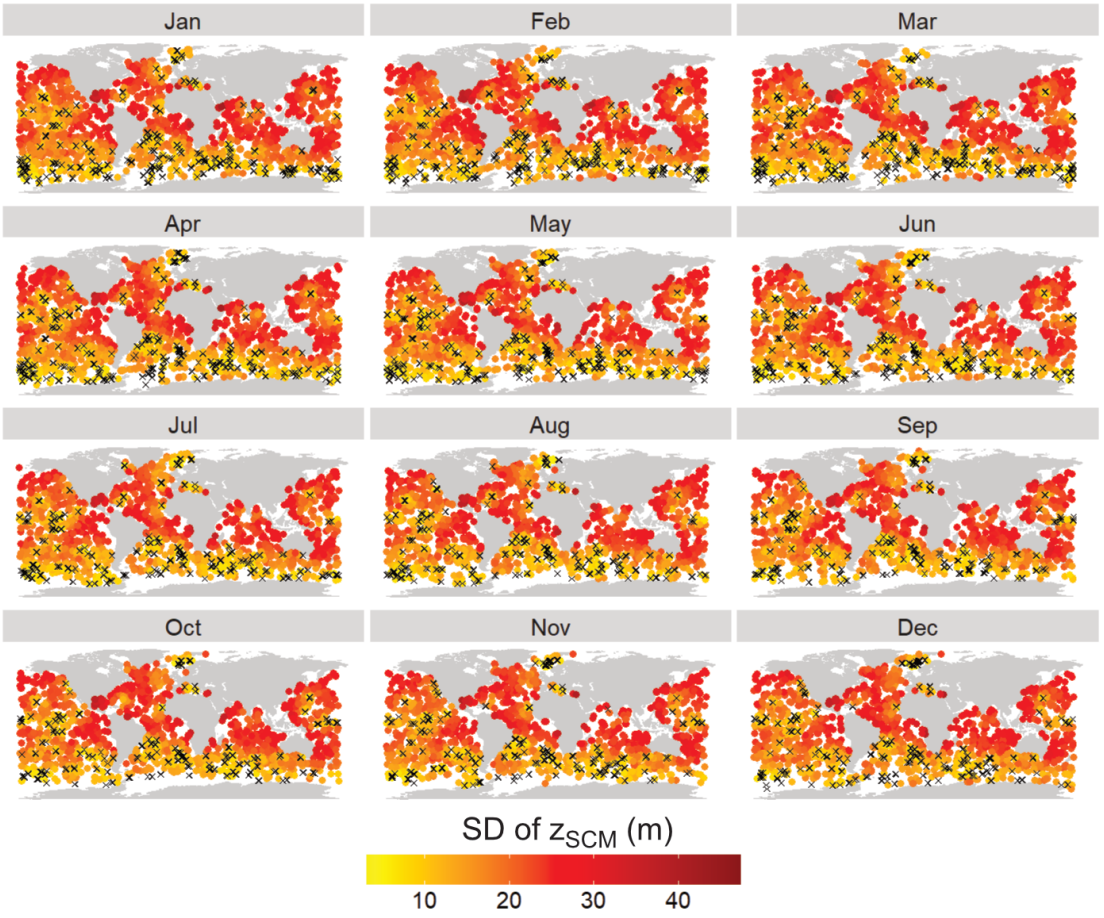

FIGURE A.3: Standard deviation of the z_{SCM} predictions using the spatio-temporal model. Locations of observations are shown as black crosses. Note how the prediction uncertainty typically increases with distance from observations.

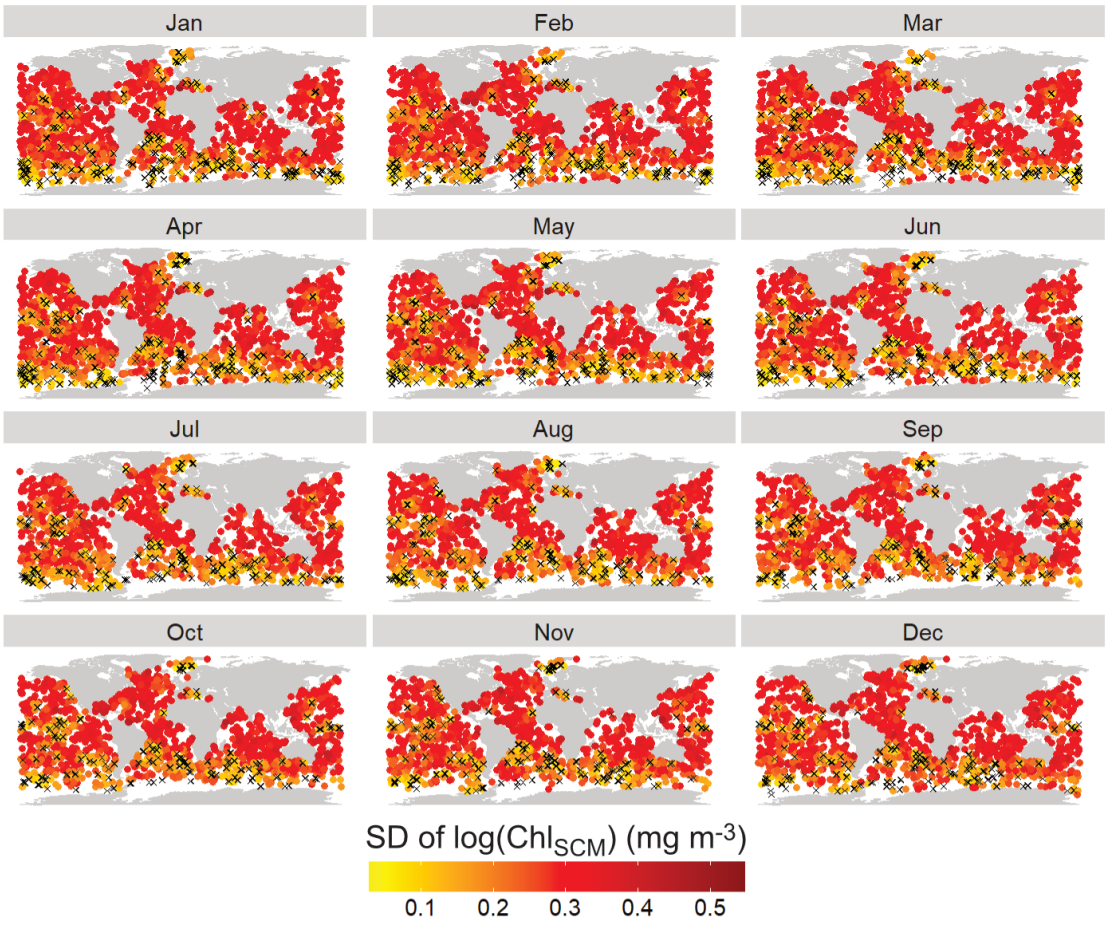

FIGURE A.4: Standard deviation of the Chl_{SCM} predictions using the spatio-temporal model. Locations of observations are shown as black crosses. Note how the prediction uncertainty typically increases with distance from observations.

Appendix B

Assessing Environmental Influences on Subsurface Chlorophyll Maxima with Functional Regression Models

Additional tables and figures supporting the analyses in Chapter 4 are presented here.

Term	Scalar-valued data	Function-valued data
Observation	A single value.	A set of values which are positioned with respect to some other variable. In theory this set is infinite, but in practice it is finite.
Intercept	The mean value of the response variable when all the covariate effects are zero.	The mean function when all of the covariate values are zero.
Linear scalar effect	A value that is multiplied by the covariate value before being added to the intercept value.	A function which is multiplied by a scalar coefficient which added to the intercept function.
Non-linear scalar effect	A value specific to a particular covariate value which is added to the intercept.	A function specific to a scalar covariate value which is added to the intercept function.
Residual	The scalar difference between the observation and the fitted value from a model	A function - the difference between the observed function and a fitted function from a model.

TABLE B.1: A summary of the differences between regression of (univariate) scalar-valued and functional data.

	EDF	Ref. df	F	p-value
Intercept(p)	18.89	19.00	21173.9	$< 2^{-16}$
MLD(p)	66.12	68.85	250.5	$< 2^{-16}$
$z_{\text{ncline}}(p)$	68.17	69.70	791.8	$< 2^{-16}$
$z_{\text{eu}}(p)$	65.20	68.35	566.7	$< 2^{-16}$

TABLE B.2: Summary of functional intercept and non-linear effects for the non-linear model for chlorophyll.

	EDF	Ref. df	F	p-value
Intercept(p)	18.77	19.00	14453.0	$< 2^{-16}$
MLD(p)	64.50	67.90	195.1	$< 2^{-16}$
$z_{\text{ncline}}(p)$	68.37	69.82	199.4	$< 2^{-16}$
$z_{\text{eu}}(p)$	64.87	68.00	450.9	$< 2^{-16}$

TABLE B.3: Summary of functional intercept and non-linear effects for the non-linear model for b_{bp}.

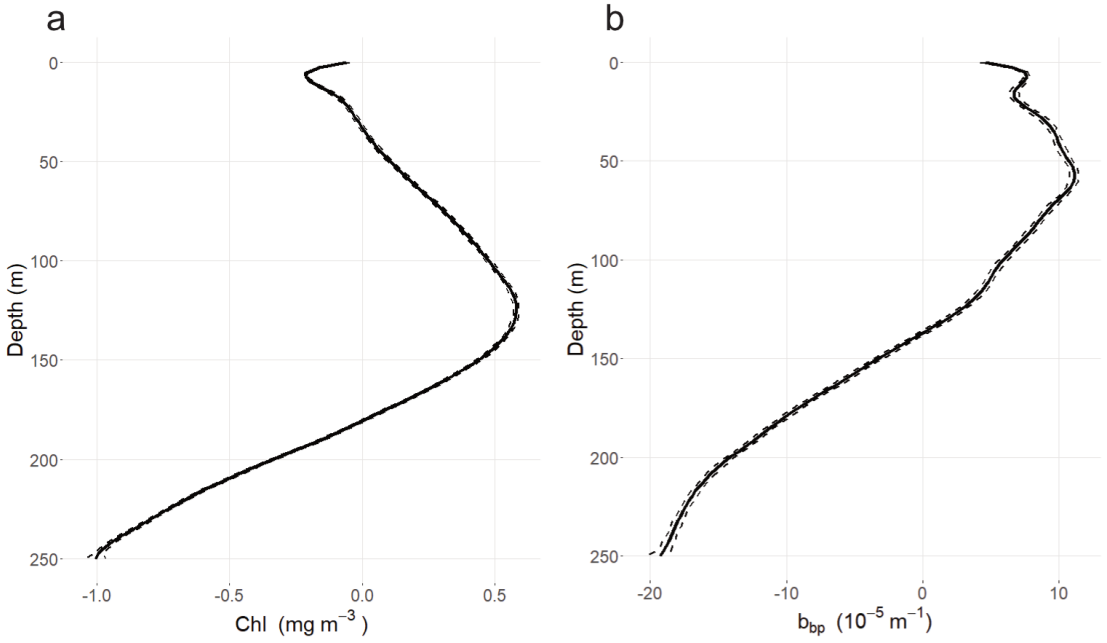

FIGURE B.1: The intercept functions from the non-linear model for (a) Chl and (b) b_{bp}. The upper and lower bounds of the 95% confidence interval are shown as dashed curves.

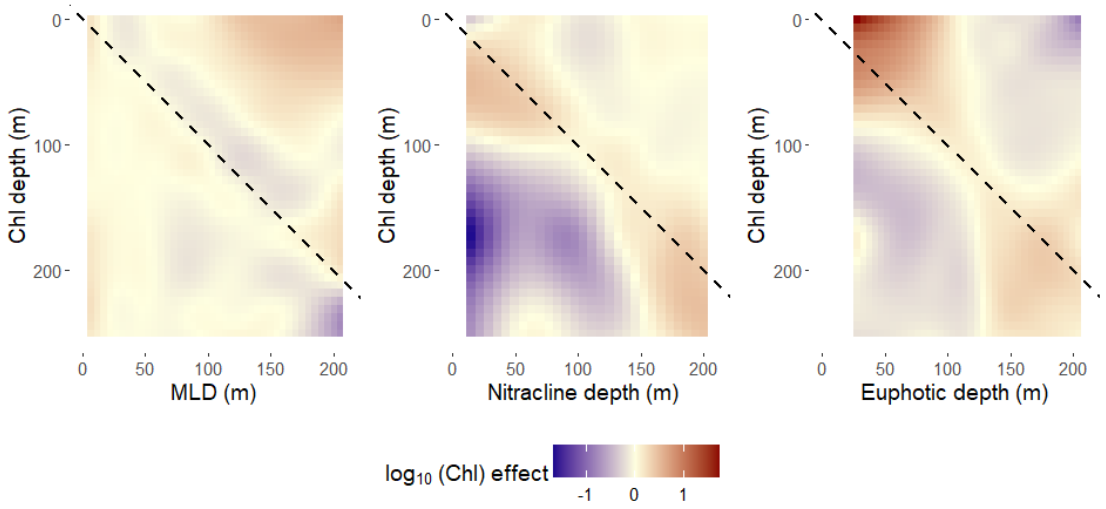

FIGURE B.2: Non-linear effects of z_{eu}, z_{ncline} and MLD on Chl profiles. This is a functional data analogue to a non-linear effect in a GAM. In a GAM, each covariate value has a corresponding scalar effect, whereas here each covariate value has a corresponding function to be added to the intercept function. A vertical line on each panel represents an additive effect for the covariate value (on the x-axis). The diagonal dashed line denotes the value of the covariate on the effect function. Note the $\log_{10}$ scale.

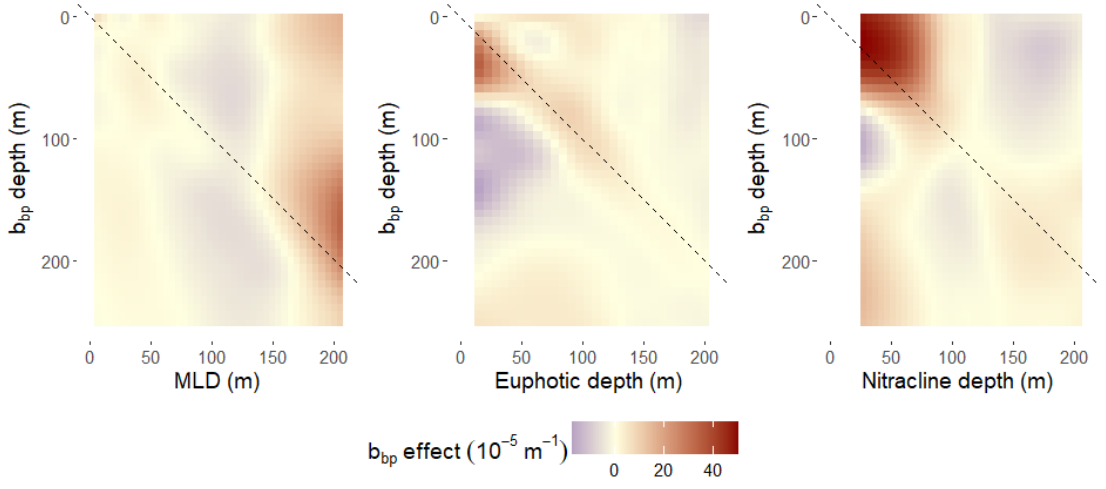

FIGURE B.3: Non-linear effects of z_{eu}, z_{ncline} and MLD on b_{bp} profiles. This is a functional data analogue to a non-linear effect in a GAM. In a GAM, each covariate value has a corresponding scalar effect, whereas here each covariate value has a corresponding function to be added to the intercept function. A vertical line on each panel represents an additive effect for the covariate value (on the x-axis). The diagonal dashed line denotes the value of the covariate on the effect function.

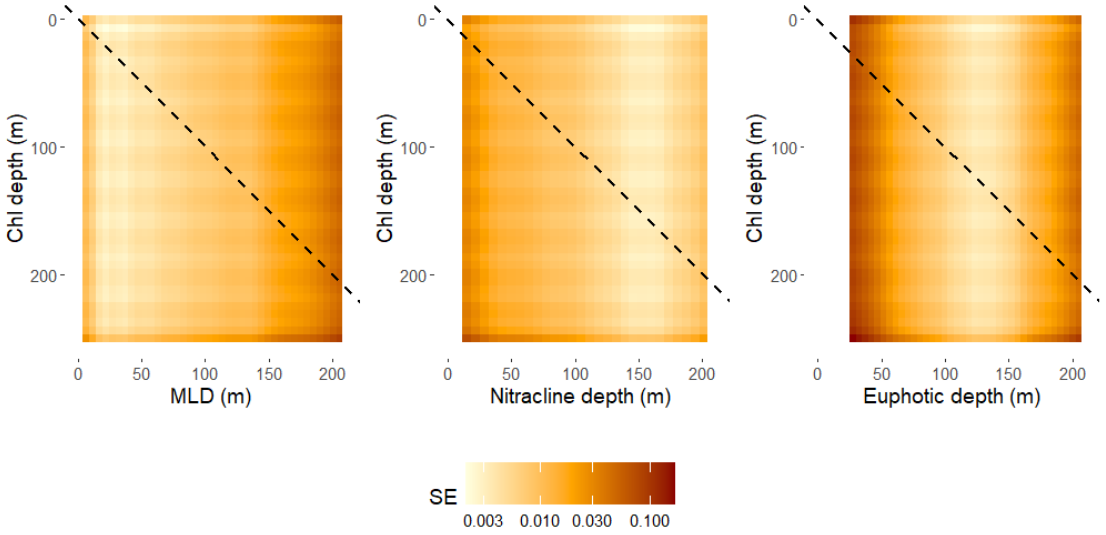

FIGURE B.4: Standard error of the three non-linear model effects for chlorophyll concentration. Note the logarithmic colour scale.

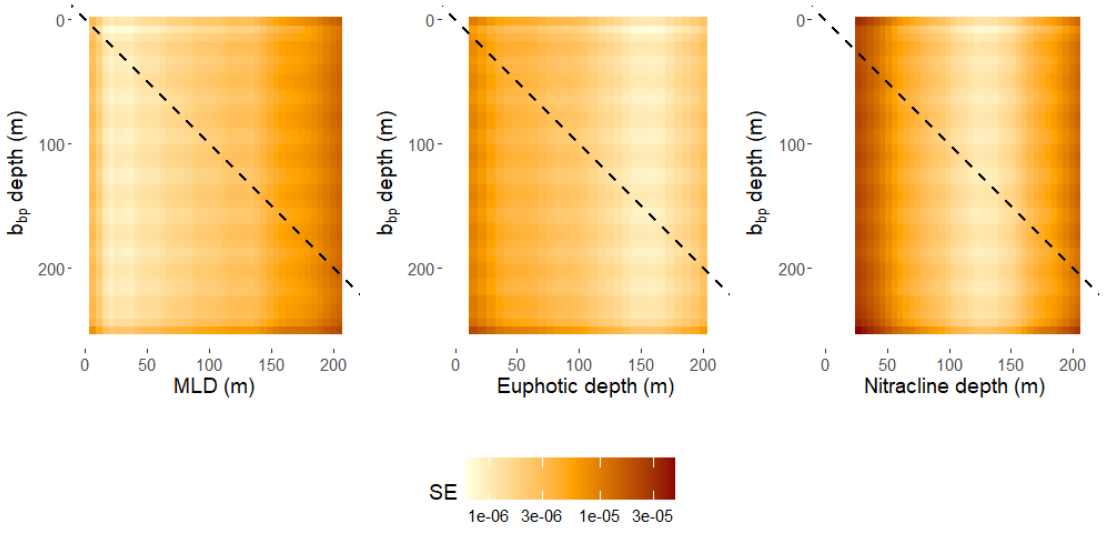

FIGURE B.5: Standard error of the three non-linear effects for b_{bp}. Note the logarithmic colour scale.

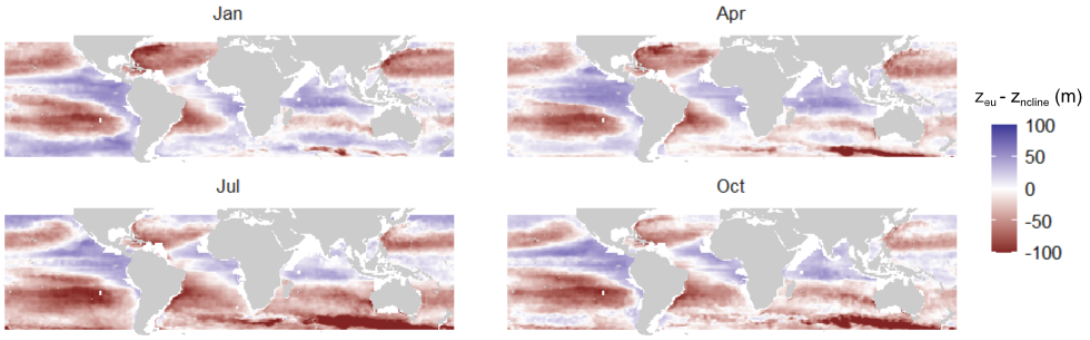

FIGURE B.6: Maps of the difference in covariate values between z_{eu} and z_{ncline} used in the predictions for January, April, July and October.

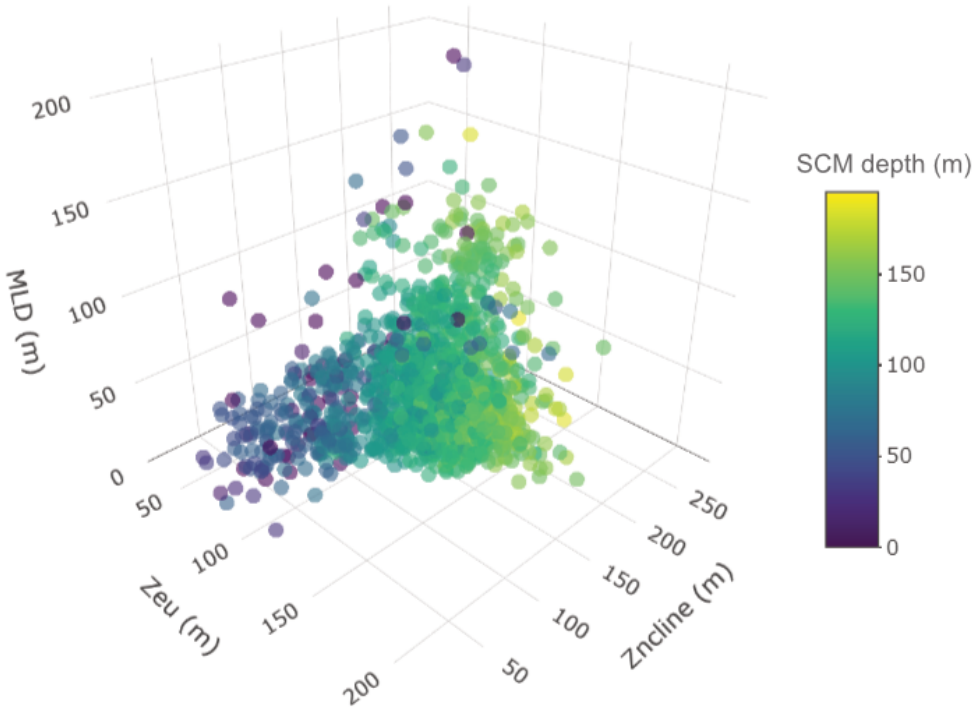

FIGURE B.7: A three-dimensional scatter plot showing the distribution of covariate values in the observed dataset and the corresponding z_{SCM} predictions.

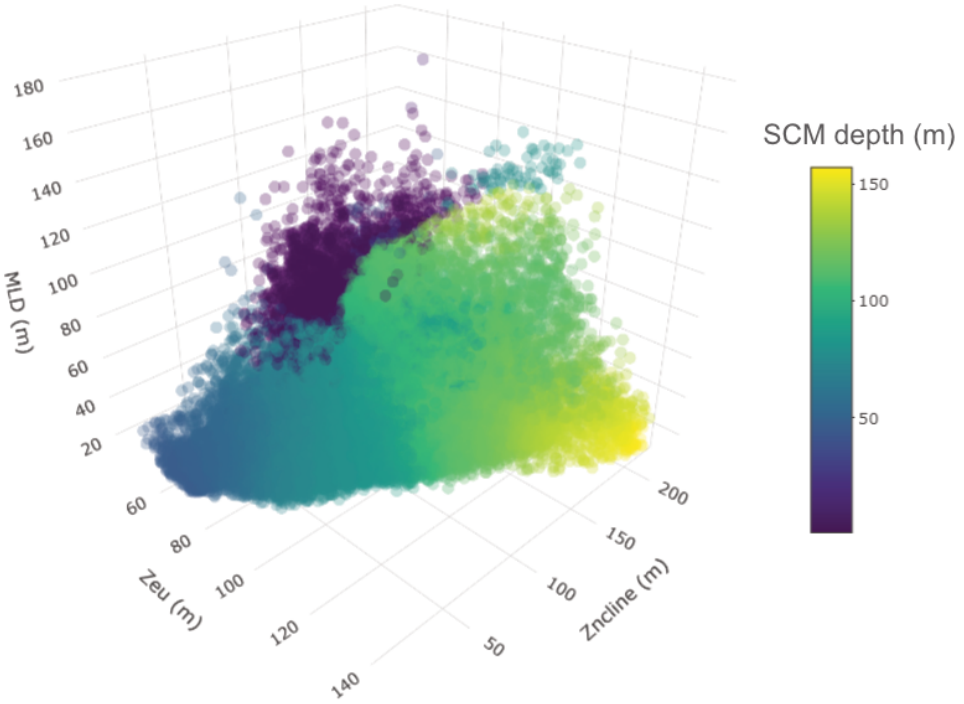

FIGURE B.8: A three-dimensional scatter plot showing the distribution of covariate values in the prediction dataset and the corresponding z_{SCM} predictions.

Appendix C

Scalar Variance and Correlation for Oceanographic Profiles: an Argo Float Application

Additional tables and figures supporting the analyses in Chapter 5 are presented here.

Biome	Season	Variance		Covariance	Correlation
		Chl	Temp		
Global	Winter	15.5	165881	149.1	0.09
	Spring	137.6	156246	-1004.6	-0.21
	Summer	256.4	160994	-2051.5	0.32
	Autumn	60.7	165916	-700.4	-0.22
EQU	Winter	19.8	27951	76.5	0.10
	Spring	17.9	28624	63.1	0.08
	Summer	17.7	25317	18.9	0.03
	Autumn	18.4	28011	60.4	0.08
STPS	Winter	12.1	34123	-164.3	-0.26
	Spring	17.1	33801	-198.8	-0.26
	Summer	18.9	34443	-169.1	-0.21
	Autumn	24.1	33770	-157.2	-0.17
STSS	Winter	16.2	27366	-166.0	-0.25
	Spring	57.0	26255	-225.3	-0.18
	Summer	86.4	30980	-565.7	-0.35
	Autumn	37.2	38549	-454.4	-0.38
SPSS	Winter	12.5	17139	39.3	0.09
	Spring	275.9	19378	-94.1	-0.04
	Summer	281.8	17705	-390.8	-0.18
	Autumn	46.1	19459	-23.5	0.03
ICE	Winter	3.5	1151	7.0	0.11
	Spring	95.5	1006	20.0	0.07
	Summer	594.3	2077	-26.6	-0.02
	Autumn	138.5	2168	57.1	0.10
MED	Winter	10.5	4274	-68.6	-0.32
	Spring	71.6	3607	-169.9	-0.33
	Summer	17.7	4139	-72.3	-0.27
	Autumn	6.4	6330	-53.4	-0.27

TABLE C.1: Summary of the variances of chlorophyll and temperature profiles by biome and meteorological season, and their covariance and correlation.

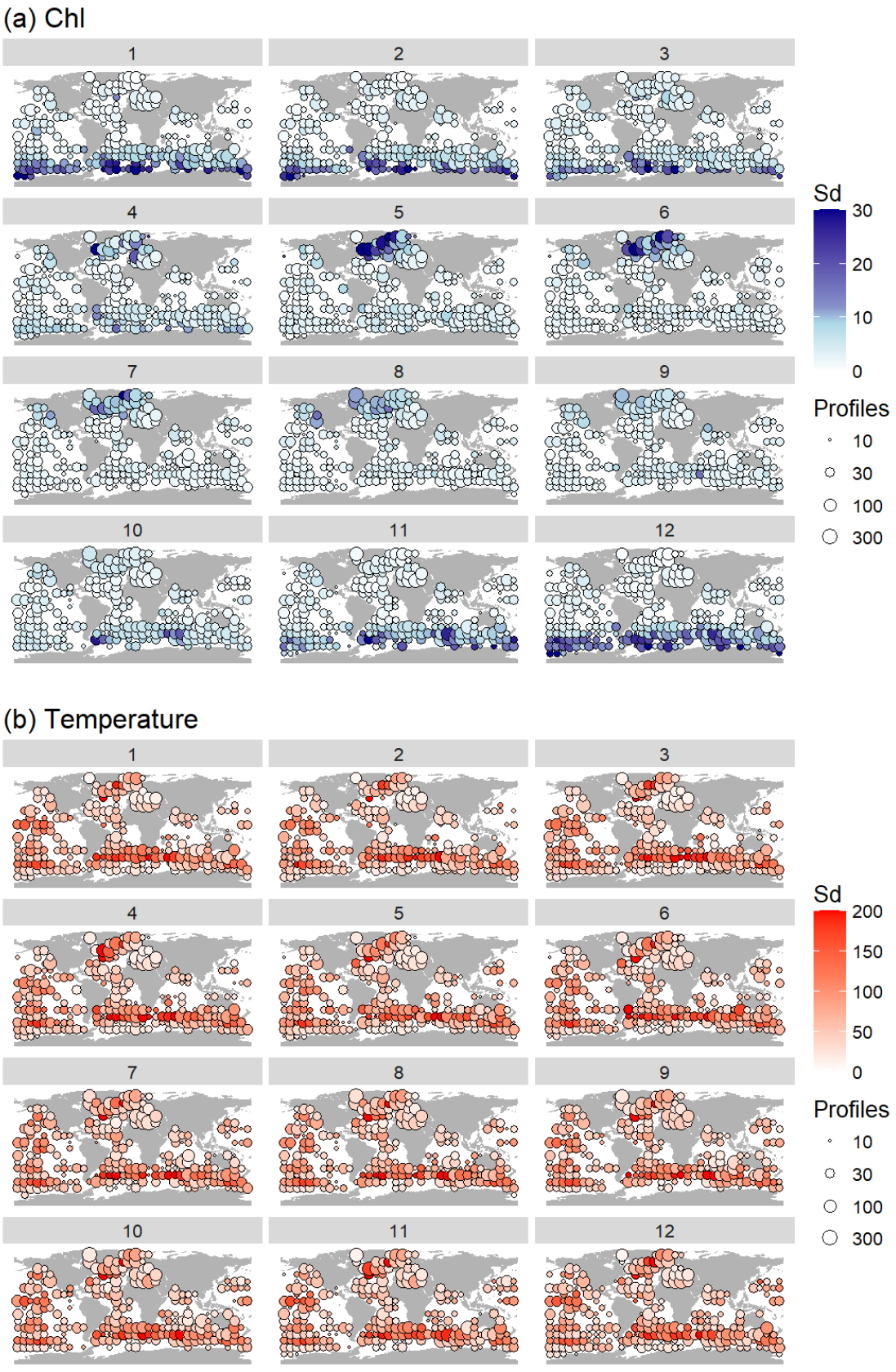

FIGURE C.1: Monthly maps of the standard deviation of (**a**) chlorophyll profiles and (**b**) temperature profiles on a 10° grid.

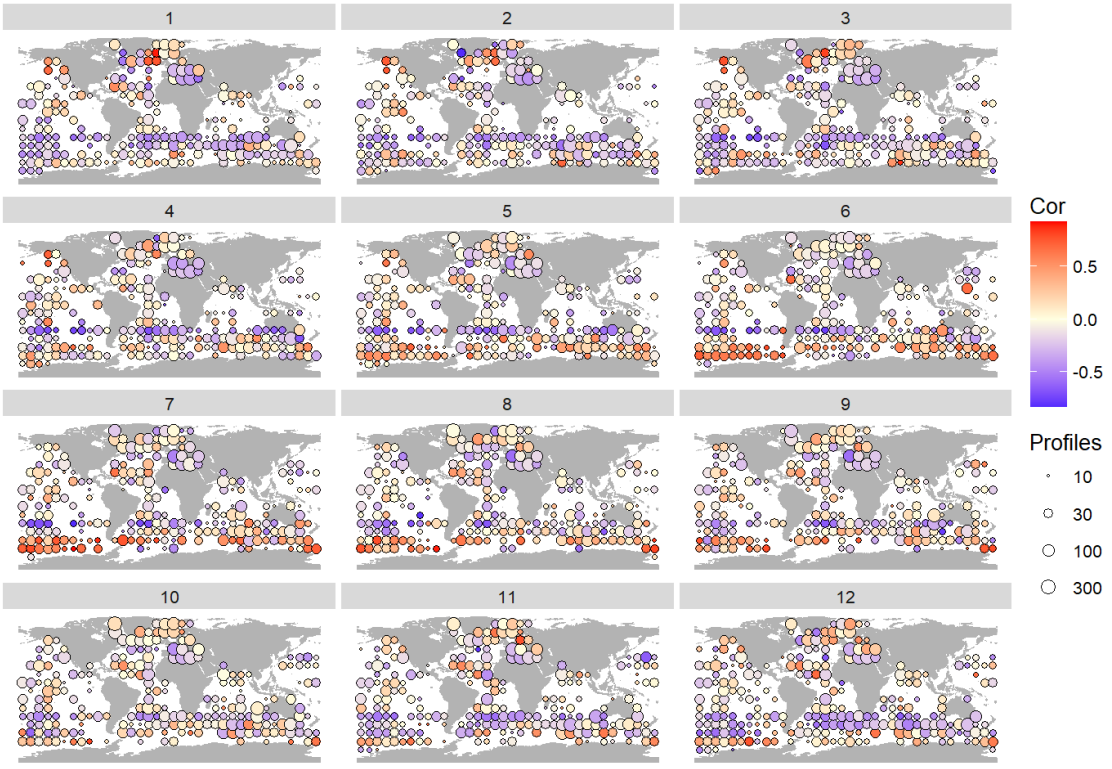

FIGURE C.2: Monthly maps of the correlation between simultaneously measured chlorophyll and temperature profiles on a 10° grid.

References

Susana Agustí and Carlos M Duarte. Phytoplankton chlorophyll a distribution and water column stability in the central Atlantic Ocean. *Oceanologica Acta*, 22(2): 193–203, 1999.

Susana Agusti, Luis M Lubián, Enrique Moreno-Ostos, Marta Estrada, and Carlos M Duarte. Projected changes in photosynthetic picoplankton in a warmer subtropical ocean. *Frontiers in Marine Science*, 5:506, 2019.

Robert F Anderson. GEOTRACES: Accelerating research on the marine biogeochemical cycles of trace elements and their isotopes. *Annual Review of Marine Science*, 12(1):49–85, 2020.

Sean C Anderson, Eric J Ward, Philina A English, and Lewis AK Barnett. sdmTMB: an R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. *bioRxiv*, pages 2022–03, 2022.

M Ardyna, M Babin, M Gosselin, E Devred, S Bélanger, A Matsuoka, and J-É Tremblay. Parameterization of vertical chlorophyll a in the Arctic Ocean: impact of the subsurface chlorophyll maximum on regional, seasonal, and annual primary production estimates. *Biogeosciences*, 10(6):4383–4404, 2013.

Kevin R Arrigo, Patricia A Matrai, and Gert L Van Dijken. Primary productivity in the Arctic Ocean: Impacts of complex optical properties and subsurface chlorophyll maxima on large-scale estimates. *Journal of Geophysical Research: Oceans*, 116(C11), 2011.

Lionel Arteaga, Markus Pahlow, and Andreas Oschlies. Global patterns of phytoplankton nutrient and light colimitation inferred from an optimality-based model. *Global Biogeochemical Cycles*, 28(7):648–661, 2014.

Lionel A Arteaga, Michael J Behrenfeld, Emmanuel Boss, and Toby K Westberry. Vertical Structure in Phytoplankton Growth and Productivity Inferred From Biogeochemical-Argo Floats and the Carbon-Based Productivity Model. *Global Biogeochemical Cycles*, 36(8):e2022GB007389, 2022.

Ramilla V Assunção, Alex C Silva, Amédée Roy, Bernard Bourlès, Carlos Henrique S Silva, Jean-François Ternon, Moacyr Araujo, and Arnaud Bertrand. 3D characterisation of the thermohaline structure in the southwestern tropical Atlantic derived from functional data analysis of in situ profiles. *Progress in Oceanography*, 187:102399, 2020.

Haakon Bakka, Håvard Rue, Geir-Arne Fuglstad, Andrea Riebler, David Bolin, Janine Illian, Elias Krainski, Daniel Simpson, and Finn Lindgren. Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6): e1443, 2018.

Haakon Bakka, Jarno Vanhatalo, Janine B Illian, Daniel Simpson, and Håvard Rue. Non-stationary Gaussian models with physical barriers. *Spatial statistics*, 29: 268–288, 2019.

Kimberlee Baldry, Peter G Strutton, Nicole A Hill, and Philip W Boyd. Subsurface chlorophyll-a maxima in the Southern Ocean. *Frontiers in Marine Science*, 7:671, 2020.

Kimberlee Baldry, Peter G Strutton, Nicole Hill, and Philip W Boyd. Observing subsurface phytoplankton in the Southern Ocean with biogeochemical Argo floats: deep chlorophyll but inconspicuous biomass. *Authorea Preprints*, 2024.

Karl Banse. Should we continue to use the 1% light depth convention for estimating the compensation depth of phytoplankton for another 70 years? *Limnology and Oceanography Bulletin*, 13(3):49–52, 2004.

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.

Aike Beckmann and Inga Hense. Beneath the surface: Characteristics of oceanic ecosystems under weak mixing conditions - A theoretical investigation. *Progress in Oceanography*, 75(4):771–796, 2007.

Charlotte Begouen Demeaux, Emmanuel Boss, Jason R Graff, Michael J Behrenfeld, and Toby K Westberry. Phytoplanktonic photoacclimation under clouds. *Geophysical Research Letters*, 52(6):e2024GL112274, 2025.

Nibedita Behera, Debadatta Swain, and Sourav Sil. Effect of Antarctic sea ice on chlorophyll concentration in the Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, 178:104853, 2020.

Jorgen Bendtsen, Clara Vives Rodriguez, and Katherine Richardson. Evaluating primary production calculated from Argo floats and satellite remote sensing in the North Atlantic. *Frontiers in Marine Science*, 10:118, 2023.

Martin Outzen Berild and Geir-Arne Fuglstad. Spatially Varying Anisotropy for Gaussian Random Fields in Three-Dimensional Space. *arXiv preprint arXiv:2301.01372*, 2023.

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.

Nicholas Bock, Marin Cornec, Hervé Claustre, and Solange Duhamel. Biogeographical Classification of the Global Ocean From BGC-Argo Floats. *Global biogeochemical cycles*, 36(6):e2021GB007233, 2022.

Luis M Bolaños, Lee Karp-Boss, Chang Jae Choi, Alexandra Z Worden, Jason R Graff, Nils Haëntjens, Alison P Chase, Alice Della Penna, Peter Gaube, Françoise Morison, et al. Small phytoplankton dominate western North Atlantic biomass. *The ISME Journal*, 14(7):1663–1674, 2020.

William M Bolstad and James M Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.

Emmanuel Boss and W Scott Pegau. Relationship of light scattering at an angle in the backward direction to the backscattering coefficient. *Applied Optics*, 40(30): 5503–5507, 2001.

Heather A Bouman, Osvaldo Ulloa, David J Scanlan, Katrin Zwirglmaier, William KW Li, Trevor Platt, Venetia Stuart, Ray Barlow, Ole Leth, Lesley Clementson, et al. Oceanographic basis of the global surface distribution of Prochlorococcus ecotypes. *Science*, 312(5775):918–921, 2006.

Heather A Bouman, Thomas Jackson, Shubha Sathyendranath, and Trevor Platt. Vertical structure in chlorophyll profiles: influence on primary production in the Arctic Ocean. *Philosophical Transactions of the Royal Society A*, 378(2181):20190351, 2020.

Daniel G Boyce, Marlon R Lewis, and Boris Worm. Global phytoplankton decline over the past century. *Nature*, 466(7306):591–596, 2010.

Philip W Boyd, David Antoine, Kimberley Baldry, Marin Cornec, Michael Ellwood, Svenja Halfter, Leo Lacour, Pauline Latour, Robert F Strzepek, Thomas W Trull, et al. Controls on polar Southern Ocean deep chlorophyll maxima: viewpoints from multiple observational platforms. *Global Biogeochemical Cycles*, 38(3):e2023GB008033, 2024.

Astrid Bracher, Hongyan Xi, Tilman Dinter, Antoine Mangin, Volker Strass, Wilken-Jon Von Appen, and Sonja Wiegmann. High resolution water column phytoplankton composition across the Atlantic Ocean from ship-towed vertical undulating radiometry. *Frontiers in Marine Science*, 7:235, 2020.

L Brannigan. Intense submesoscale upwelling in anticyclonic eddies. *Geophysical Research Letters*, 43(7):3360–3369, 2016.

Robert JW Brewin, Shubha Sathyendranath, Takafumi Hirata, Samantha J Lavender, Rosa M Barciela, and Nick J Hardman-Mountford. A three-component model of phytoplankton size class for the Atlantic Ocean. *Ecological Modelling*, 221(11): 1472–1483, 2010.

Robert JW Brewin, Shubha Sathyendranath, Trevor Platt, Heather Bouman, Stefano Ciavatta, Giorgio Dall'Olmo, James Dingle, Steve Groom, Bror Jönsson, Tihomir S Kostadinov, et al. Sensing the ocean biological carbon pump from space: A review of capabilities, concepts, research gaps and future developments. *Earth-Science Reviews*, 217:103604, 2021.

Robert JW Brewin, Giorgio Dall'Olmo, John Gittings, Xuerong Sun, Priscila K Lange, Dionysios E Raitsos, Heather A Bouman, Ibrahim Hoteit, Jim Aiken, and Shubha Sathyendranath. A conceptual approach to partitioning a vertical profile of phytoplankton biomass into contributions from two communities. *Journal of Geophysical Research: Oceans*, 127(4):e2021JC018195, 2022.

Thomas J Browning and C Mark Moore. Global analysis of ocean phytoplankton nutrient limitation reveals high prevalence of co-limitation. *Nature Communications*, 14(1):5014, 2023.

Janet W Campbell. The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans*, 100(C7):13237–13254, 1995.

Magdalena M Carranza, Sarah T Gille, Peter JS Franks, Kenneth S Johnson, Robert Pinkel, and James B Girton. When mixed layers are not mixed. Storm-driven mixing and bio-optical vertical gradients in mixed layers of the Southern Ocean. *Journal of Geophysical Research: Oceans*, 123(10):7264–7289, 2018.

Filipa Carvalho, Maxim Y Gorbunov, Matthew J Oliver, Christina Haskins, David Aragon, Josh T Kohut, and Oscar Schofield. FIRe glider: mapping in situ chlorophyll variable fluorescence with autonomous underwater gliders. *Limnology and Oceanography: Methods*, 18(9):531–545, 2020.

Ivona Cetinić, Mary Jane Perry, Nathan T Briggs, Emily Kallin, Eric A D'Asaro, and Craig M Lee. Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment. *Journal of Geophysical Research: Oceans*, 117(C6), 2012.

Fei Chai, Kenneth S Johnson, Hervé Claustre, Xiaogang Xing, Yuntao Wang, Emmanuel Boss, Stephen Riser, Katja Fennel, Oscar Schofield, and Adrienne Sutton. Monitoring ocean biogeochemistry with autonomous platforms. *Nature Reviews Earth & Environment*, 1(6):315–326, 2020.

Paul Chamberlain, Lynne D Talley, Bruce Cornuelle, Matthew Mazloff, and Sarah T Gille. Optimizing the biogeochemical Argo float distribution. *Journal of Atmospheric and Oceanic Technology*, 40(11):1355–1379, 2023.

Paul M Chamberlain, Lynne D Talley, Matthew R Mazloff, Stephen C Riser, Kevin Speer, Alison R Gray, and Armin Schwartzman. Observing the ice-covered Weddell Gyre with profiling floats: Position uncertainties and correlation statistics. *Journal of Geophysical Research: Oceans*, 123(11):8383–8410, 2018.

Christopher C Chapman, Mary-Anne Lea, Amelie Meyer, Jean-Baptiste Sallée, and Mark Hindell. Defining Southern Ocean fronts and their influence on biological and physical processes in a changing climate. *Nature Climate Change*, 10(3):209–219, 2020.

Somnath Chaudhuri, Pablo Juan, Laura Serra Saurina, Diego Varga, and Marc Saez. Modeling spatial dependencies of natural hazards in coastal regions: a nonstationary approach with barriers. *Stochastic Environmental Research and Risk Assessment*, 37(11):4479–4498, 2023.

Dudley B Chelton, Michael G Schlax, and Roger M Samelson. Global observations of nonlinear mesoscale eddies. *Progress in oceanography*, 91(2):167–216, 2011.

Jianqiang Chen, Xun Gong, Xinyu Guo, Xiaogang Xing, Keyu Lu, Huiwang Gao, and Xiang Gong. Improved Perceptron of Subsurface Chlorophyll Maxima by a Deep Neural Network: A Case Study with BGC-Argo Float Data in the Northwestern Pacific Ocean. *Remote Sensing*, 14(3):632, 2022.

Wanfang Chen, Marc G Genton, and Ying Sun. Space-time covariance structures and models. *Annual Review of Statistics and Its Application*, 8(1):191–215, 2021.

Stephen M Chiswell, Paulo HR Calil, and Philip W Boyd. Spring blooms and annual cycles of phytoplankton: a unified perspective. *Journal of Plankton Research*, 37(3): 500–508, 2015.

KM Azam Chowdhury, Wensheng Jiang, Guimei Liu, Md Kawser Ahmed, and Shaila Akhter. Dominant physical-biogeochemical drivers for the seasonal variations in the surface chlorophyll-a and subsurface chlorophyll-a maximum in the Bay of Bengal. *Regional Studies in Marine Science*, 48:102022, 2021.

Winnie U Chu, Matthew R Mazloff, Ariane Verdy, Sarah G Purkey, and Bruce D Cornuelle. Optimizing observational arrays for biogeochemistry in the tropical Pacific by estimating correlation lengths. *Limnology and Oceanography: Methods*, 22 (11):840–852, 2024.

Philippe Ciais, Christopher Sabine, Govindasamy Bala, Laurent Bopp, Victor Brovkin, Joanna Isobel House, et al. Carbon and other biogeochemical cycles. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change Change*, pages 465–570. Cambridge University Press, 2014.

Hervé Claustre, Kenneth S Johnson, and Yuichiro Takeshita. Observing the global ocean with biogeochemical-Argo. *Annual review of marine science*, pages 23–48, 2020.

James E Cloern, Christian Grenz, and Lisa Vidergar-Lucas. An empirical model of the phytoplankton chlorophyll: carbon ratio-the conversion factor between productivity and growth rate. *Limnology and Oceanography*, 40(7):1313–1321, 1995.

Marin Cornec, Hervé Claustre, Alexandre Mignot, Lionel Guidi, Leo Lacour, A Poteau, F d'Ortenzio, Bernard Gentili, and Catherine Schmechtig. Deep chlorophyll maxima in the global ocean: Occurrences, drivers and characteristics. *Global Biogeochemical Cycles*, 35(4):e2020GB006759, 2021a.

Marin Cornec, Rémi Laxenaire, Sabrina Speich, and Hervé Claustre. Impact of mesoscale eddies on deep chlorophyll maxima. *Geophysical research letters*, 48(15): e2021GL093470, 2021b.

Timothy J Cowles, Russell A Desiderio, and Mary-Elena Carr. Small-scale planktonic structure: persistence and trophic consequences. *Oceanography*, 11(1):4–9, 1998.

Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2011.

JJ Cullen and RW Eppley. Chlorophyll maximum layers of the Southern-California Bight and possible mechanisms of their formation and maintenance. *Oceanologica Acta*, 4(1):23–32, 1981.

John J Cullen. Diel vertical migration by dinoflagellates: roles of carbohydrate metabolism and behavioral flexibility. *Contrib. Mar. Sci*, 27:135–152, 1985.

John J Cullen. Subsurface chlorophyll maximum layers: enduring enigma or mystery solved? *Annual Review of Marine Science*, 7:207–239, 2015.

John J Cullen and J Geoffrey MacIntyre. Behavior, physiology and the niche of depth-regulating phytoplankton. *Nato Asi Series G Ecological Sciences*, 41:559–580, 1998.

Unn Dahlén, Johan Lindström, and Marko Scholze. Spatiotemporal reconstructions of global CO_2-fluxes using Gaussian Markov random fields. *Environmetrics*, 31(4): e2610, 2020.

Minhan Dai, Ya-Wei Luo, Eric P Achterberg, Thomas J Browning, Yihua Cai, Zhimian Cao, Fei Chai, Bingzhang Chen, Matthew J Church, Dongjian Ci, et al. Upper ocean biogeochemistry of the oligotrophic North Pacific Subtropical Gyre: From nutrient sources to carbon export. *Reviews of Geophysics*, 61(3):e2022RG000800, 2023.

Giorgio Dall'olmo, Henry Bittig, Emmanuel Boss, Jodi Brewster, Hervé Claustre, Matt Donnelly, David Nicholson, Violetta Paba, Antoine Poteau, Raphaëlle Sauzède, et al. BGC Argo quality control manual for particles backscattering. 2023.

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, 2023.

Clément de Boyer Montégut, Gurvan Madec, Albert S Fischer, Alban Lazar, and Daniele Iudicone. Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *Journal of Geophysical Research: Oceans*, 109(C12), 2004.

Margaret M Dekshenieks, Percy L Donaghay, James M Sullivan, Jan EB Rines, Thomas R Osborn, and Michael S Twardowski. Temporal and spatial occurrence of thin phytoplankton layers in relation to physical processes. *Marine Ecology Progress Series*, 223:61–71, 2001.

Tim DeVries. The ocean carbon cycle. *Annual Review of Environment and Resources*, 47 (1):317–341, 2022.

Tim DeVries and Thomas Weber. The export and fate of organic matter in the ocean: New constraints from combining satellite and oceanographic tracer observations. *Global Biogeochemical Cycles*, 31(3):535–555, 2017.

Zhenhong Du, Sensen Wu, Mei-Po Kwan, Chuanrong Zhang, Feng Zhang, and Renyi Liu. A spatiotemporal regression-kriging model for space-time interpolation: A case study of chlorophyll-a prediction in the coastal areas of Zhejiang, China. *International Journal of Geographical Information Science*, 32(10):1927–1947, 2018.

William M Durham and Roman Stocker. Thin phytoplankton layers: characteristics, mechanisms, and consequences. *Annual review of marine science*, 4:177–207, 2012.

JK Egge and DL Aksnes. Silicate as regulating nutrient in phytoplankton competition. *Marine ecology progress series. Oldendorf*, 83(2):281–289, 1992.

Urs Hofmann Elizondo, Damiano Righetti, Fabio Benedetti, and Meike Vogt. Biome partitioning of the global ocean based on phytoplankton biogeography. *Progress in Oceanography*, 194:102530, 2021.

ML Estapa, ML Feen, and Elly Breves. Direct observations of biological carbon export from profiling floats in the subtropical North Atlantic. *Global Biogeochemical Cycles*, 33(3):282–300, 2019.

Paul Falkowski and Dale A Kiefer. Chlorophyll a fluorescence in phytoplankton: relationship to photosynthesis and biomass. *Journal of Plankton Research*, 7(5): 715–731, 1985.

Paul G Falkowski. The role of phytoplankton photosynthesis in global biogeochemical cycles. *Photosynthesis research*, 39:235–258, 1994.

Paul G Falkowski and John A Raven. *Aquatic photosynthesis*. Princeton University Press, 2013.

AR Fay and GA McKinley. Global open-ocean biomes: mean and temporal variability. *Earth System Science Data*, 6(2):273–284, 2014.

J Feng, J Durant, L Stige, Dag Olav Hessen, Dag Øystein Hjermann, Lin Zhu, Marcos Llope, and Nils C. Stenseth. Contrasting correlation patterns between environmental factors and chlorophyll levels in the global ocean. *Global Biogeochemical Cycles*, 29(12):2095–2107, 2015.

Katja Fennel and Emmanuel Boss. Subsurface maxima of phytoplankton and chlorophyll: Steady-state solutions from a simple model. *Limnology and Oceanography*, 48(4):1521–1534, 2003.

Christopher B Field, Michael J Behrenfeld, James T Randerson, and Paul Falkowski. Primary production of the biosphere: integrating terrestrial and oceanic components. *science*, 281(5374):237–240, 1998.

Guido Fioravanti, Sara Martino, Michela Cameletti, and Andrea Toreti. Interpolating climate variables by using INLA and the SPDE approach. *International Journal of Climatology*, 43(14):6866–6886, 2023.

Mochamad Ramdhan Firdaus, Arief Rachman, Nurul Fitriya, Lady Ayu Sri Wijayanti, Adi Purwandana, Hanif Budi Prayitno, Yustian Rovi Alfiansyah, Oksto Ridho Sianturi, Hagi Yulia Sugeha, et al. Vertical and Horizontal Variability of Chlorophyll-a and Its Relationship with Environmental Parameters in the Waters of Sangihe and Talaud Islands, North Sulawesi, Indonesia. *ILMU KELAUTAN: Indonesian Journal of Marine Sciences*, 29(1), 2024.

Pierre Friedlingstein, Michael O'sullivan, Matthew W Jones, Robbie M Andrew, Judith Hauck, Peter Landschützer, Corinne Le Quéré, Hongmei Li, Ingrid T Luijkx, Are Olsen, et al. Global carbon budget 2024. *Earth System Science Data Discussions*, 2024: 1–133, 2024.

Geir-Arne Fuglstad and Stefano Castruccio. Compression of climate simulations with a nonstationary global SpatioTemporal SPDE model. *The Annals of Applied Statistics*, 14(2):542–559, 2020.

Geir-Arne Fuglstad, Finn Lindgren, Daniel Simpson, and Håvard Rue. Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, pages 115–133, 2015.

Jorge García-Jimenez, Ana Belén Ruescas, Julia Amorós-López, and Raphaëlle Sauzède. Combining BGC-Argo floats and satellite observations for water column estimations of particulate backscattering coefficient. *EGUsphere*, 2025:1–23, 2025.

Shriya Garg, Mangesh Gauns, and Anil K Pratihary. Response of oceanic subsurface chlorophyll maxima to environmental drivers in the Northern Indian Ocean. *Environmental Research*, 240:117528, 2024.

Sarah T Gille and Kathryn A Kelly. Scales of spatial and temporal variability in the Southern Ocean. *Journal of Geophysical Research: Oceans*, 101(C4):8759–8773, 1996.

X Gong, J Shi, HW Gao, and XH Yao. Steady-state solutions for subsurface chlorophyll maximum in stratified water columns with a bell-shaped vertical profile of chlorophyll. *Biogeosciences*, 12(4):905–919, 2015.

Xiang Gong, Wensheng Jiang, Linhui Wang, Huiwang Gao, Emmanuel Boss, Xiaohong Yao, Shuh-Ji Kao, and Jie Shi. Analytical solution of the nitracline with the evolution of subsurface chlorophyll maximum in stratified water columns. *Biogeosciences*, 14(9):2371–2386, 2017.

Jason R Graff, Toby K Westberry, Allen J Milligan, Matthew B Brown, Giorgio Dall'Olmo, Virginie van Dongen-Vogels, Kristen M Reifel, and Michael J Behrenfeld. Analytical phytoplankton carbon measurements spanning diverse ecosystems. *Deep Sea Research Part I: Oceanographic Research Papers*, 102:16–25, 2015.

Robert M Graham, Agatha M De Boer, Erik van Sebille, Karen E Kohfeld, and Christian Schlosser. Inferring source regions and supply mechanisms of iron in the Southern Ocean from satellite chlorophyll data. *Deep Sea Research Part I: Oceanographic Research Papers*, 104:9–25, 2015.

Sonja Greven and Fabian Scheipl. A general framework for functional regression modelling. *Statistical Modelling*, 17(1-2):1–35, 2017.

Nicolas Gruber, Dorothee CE Bakker, Tim DeVries, Luke Gregor, Judith Hauck, Peter Landschützer, Galen A McKinley, and Jens Daniel Müller. Trends and variability in the ocean carbon sink. *Nature Reviews Earth & Environment*, 4(2):119–134, 2023.

Kevin R Gurney, David Baker, Peter Rayner, and Scott Denning. Interannual variations in continental-scale net carbon exchange and sensitivity to observing

networks estimated from atmospheric CO2 inversions for the period 1980 to 2005. *Global Biogeochemical Cycles*, 22(3), 2008.

Mark Hague and Marcello Vichi. Southern Ocean Biogeochemical Argo detect under-ice phytoplankton growth before sea ice retreat. *Biogeosciences*, 18(1):25–38, 2021.

Matthew L Hammond, Claudie Beaulieu, Stephanie A Henson, and Sujit K Sahu. Regional surface chlorophyll trends and uncertainties in the global ocean. *Scientific reports*, 10(1):15273, 2020.

Nicholas J Hawco, Benedetto Barone, Matthew J Church, Lydia Babcock-Adams, Daniel J Repeta, Emma K Wear, Rhea K Foreman, Karin M Björkman, Shavonna Bent, Benjamin AS Van Mooy, et al. Iron depletion in the deep chlorophyll maximum: Mesoscale eddies as natural iron fertilization experiments. *Global Biogeochemical Cycles*, 35(12):e2021GB007112, 2021.

Nicholas J Hawco, Alessandro Tagliabue, and Benjamin S Twining. Manganese limitation of phytoplankton physiology and productivity in the Southern Ocean. *Global Biogeochemical Cycles*, 36(11):e2022GB007382, 2022.

Stephanie A Henson, John P Dunne, and Jorge L Sarmiento. Decadal variability in North Atlantic phytoplankton blooms. *Journal of Geophysical Research: Oceans*, 114 (C4), 2009.

Stephanie A Henson, Jorge Louis Sarmiento, John P Dunne, Laurent Bopp, I Lima, Scott C Doney, J John, and C Beaulieu. Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity. *Biogeosciences*, 7(2):621–640, 2010.

Stephanie A Henson, Claudie Beaulieu, and Richard Lampitt. Observing climate change trends in ocean biogeochemistry: when and where. *Global change biology*, 22 (4):1561–1571, 2016.

Alain Herbland and Bruno Voituriez. Hydrological structure analysis for estimating the primary production in the tropical Atlantic Ocean. *Journal of Marine Research*, 1979.

Anders Hildeman, David Bolin, and Igor Rychlik. Deformed SPDE models with an application to spatial modeling of significant wave height. *Spatial Statistics*, 42: 100449, 2021.

Benjamin A Hodges and Daniel L Rudnick. Simple models of steady deep maxima in chlorophyll and biomass. *Deep Sea Research Part I: Oceanographic Research Papers*, 51 (8):999–1015, 2004.

Osmund Holm-Hansen and Christopher D Hewes. Deep chlorophyll-a maxima (DCMs) in Antarctic waters: I. Relationships between DCMs and the physical, chemical, and optical conditions in the upper water column. *Polar Biology*, 27: 699–710, 2004.

Qiwei Hu, Xiaoyan Chen, Yan Bai, Xianqiang He, Teng Li, and Delu Pan. Reconstruction of 3-D ocean chlorophyll a structure in the northern Indian ocean using satellite and BGC-Argo data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2022a.

Wenjing Hu, Geir-Arne Fuglstad, and Stefano Castruccio. A stochastic locally diffusive model with neural network-based deformations for global sea surface temperature. *Stat*, 11(1):e431, 2022b.

Jie Huang and Fanghua Xu. Observational evidence of subsurface chlorophyll response to mesoscale eddies in the North Pacific. *Geophysical Research Letters*, 45 (16):8462–8470, 2018.

Jef Huisman, Nga N Pham Thi, David M Karl, and Ben Sommeijer. Reduced mixing generates oscillations and chaos in the oceanic deep chlorophyll maximum. *Nature*, 439(7074):322–325, 2006.

Rikke Ingebrigtsen, Finn Lindgren, and Ingelin Steinsland. Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, 8:20–38, 2014.

Sachihiko Itoh, Ichiro Yasuda, Hiroaki Saito, Atsushi Tsuda, and Kosei Komatsu. Mixed layer depth and chlorophyll a: Profiling float observations in the Kuroshio–Oyashio Extension region. *Journal of Marine Systems*, 151:1–14, 2015.

Chiranjivi Jayaram, TVS Udaya Bhaskar, Neethu Chacko, Satya Prakash, and KH Rao. Spatio-temporal variability of chlorophyll in the northern Indian Ocean: A biogeochemical argo data perspective. *Deep Sea Research Part II: Topical Studies in Oceanography*, 183:104928, 2021.

Gabriel Jorda, Núria Marbà, Scott Bennett, Julia Santana-Garcon, Susana Agusti, and Carlos M Duarte. Ocean warming compresses the three-dimensional habitat of marine life. *Nature Ecology & Evolution*, 4(1):109–114, 2020.

Hannah L Joy-Warren, Gert L van Dijken, Anne-Carlijn Alderkamp, Amy Leventer, Kate M Lewis, Virginia Selz, Kate E Lowry, Willem van de Poll, and Kevin R Arrigo. Light is the primary driver of early season phytoplankton production along the western Antarctic Peninsula. *Journal of Geophysical Research: Oceans*, 124(11): 7375–7399, 2019.

Yoba Kande, Ndague Diogoul, Patrice Brehmer, Sophie Dabo-Niang, Papa Ngom, and Yannick Perrot. Demonstrating the relevance of spatial-functional statistical analysis

in marine ecological studies: The case of environmental variations in micronektonic layers. *Ecological Informatics*, 81:102547, 2024.

David M Karl. A sea of change: biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems*, 2:181–214, 1999.

Madhavan Girijakumari Keerthi, Channing J Prend, Olivier Aumont, and Marina Lévy. Annual variations in phytoplankton biomass driven by small-scale physical processes. *Nature Geoscience*, pages 1–7, 2022.

Dan Kelley, Clark Richards, and WG127 SCOR. Package 'gsw', 2021a.

Dan E Kelley, Jaimie Harbin, and Clark Richards. argoFloats: An R package for analyzing Argo data. *Frontiers in Marine Science*, 8:635922, 2021b.

DA Kiefer, RJ Olson, and O Holm-Hansen. Another look at the nitrite and chlorophyll maxima in the central North Pacific. In *Deep Sea Research and Oceanographic Abstracts*, volume 23, pages 1199–1208. Elsevier, 1976.

Moritz Korte-Stapff, Drew Yarger, Stilian Stoev, and Tailen Hsing. A multivariate functional-data mixture model for spatio-temporal data: inference and cokriging. *arXiv preprint arXiv:2211.04012*, 2022.

Elias Krainski, Virgilio Gómez-Rubio, Haakon Bakka, Amanda Lenzi, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren, and Håvard Rue. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC, 2018.

EA Kubryakova, YI Bakueva, and AA Kubryakov. Spatial and Seasonal Variability of the Vertical Distribution of Chlorophyll a Concentration in the Southern Ocean from Bio-Argo Data. *Oceanology*, 65(1):90–103, 2025.

Angela M Kuhn, Matthew Mazloff, Stephanie Dutkiewicz, Oliver Jahn, Sophie Clayton, Tatiana Rynearson, and Andrew D Barton. A global comparison of marine chlorophyll variability observed in Eulerian and Lagrangian perspectives. *Journal of Geophysical Research: Oceans*, 128(7):e2023JC019801, 2023.

Angela M Kuhn, Matthew R Mazloff, Sarah T Gille, and Ariane Verdy. Sensitivity of chlorophyll vertical structure to model parameters in the Biogeochemical Southern Ocean State Estimate (B-SOSE). *Journal of Geophysical Research: Biogeosciences*, 130(1): e2024JG008300, 2025.

Mikael Kuusela and Michael L Stein. Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A*, 474(2220):20180400, 2018.

Lester Kwiatkowski, Olivier Torres, Laurent Bopp, Olivier Aumont, Matthew Chamberlain, James R Christian, John P Dunne, Marion Gehlen, Tatiana Ilyina, Jasmin G John, et al. Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections. *Biogeosciences*, 17(13):3439–3470, 2020.

Mikel Latasa, Andrés Gutiérrez-Rodríguez, Ana María Cabello, and Renate Scharek. Influence of light and nutrients on the vertical distribution of marine phytoplankton groups in the deep chlorophyll maximum. *Scientia Marina 80S1*, pages 57–62, 2016.

Mikel Latasa, Ana María Cabello, Xosé Anxelu G Morán, Ramon Massana, and Renate Scharek. Distribution of phytoplankton groups within the deep chlorophyll maximum. *Limnology and Oceanography*, 62(2):665–685, 2017.

Mikel Latasa, Francisco Rodríguez, Susana Agusti, and Marta Estrada. Distribution patterns of phytoplankton groups along isoirradiance layers in oligotrophic tropical and subtropical oceans. *Progress in Oceanography*, 217:103098, 2023.

Márcio P Laurini. A spatio-temporal approach to estimate patterns of climate change. *Environmetrics*, 30(1):e2542, 2019.

H Lavigne, F D'ortenzio, M Ribera D'Alcalà, Hervé Claustre, R Sauzède, and M Gacic. On the vertical distribution of the chlorophyll a concentration in the Mediterranean Sea: a basin-scale and seasonal approach. *Biogeosciences*, 12(16):5021–5039, 2015.

Loïc Le Ster, Hervé Claustre, Francesco d'Ovidio, David Nerini, Baptiste Picard, and Christophe Guinet. Improved accuracy and spatial resolution for bio-logging-derived chlorophyll a fluorescence measurements in the Southern Ocean. *Frontiers in Marine Science*, 10:1122822, 2023.

Ricardo M Letelier, David M Karl, Mark R Abbott, and Robert R Bidigare. Light driven seasonal patterns of chlorophyll and nitrate in the lower euphotic zone of the North Pacific Subtropical Gyre. *Limnology and Oceanography*, 49(2):508–519, 2004.

Marina Levy, D Couespel, C Haeck, MG Keerthi, I Mangolte, and CJ Prend. Impact of finescale currents on biogeochemical cycles in a changing ocean. *Annual Review of Marine Science*, 2023.

Marlon R Lewis, John J Cullen, and Trevor Platt. Phytoplankton and thermal structure in the upper ocean: Consequences of nonuniformity in chlorophyll profile. *Journal of Geophysical Research: Oceans*, 88(C4):2565–2570, 1983.

Qian P Li and Dennis A Hansell. Mechanisms controlling vertical variability of subsurface chlorophyll maxima in a mode-water eddy. *Journal of Marine Research*, 2016.

Senjie Lin, Richard Wayne Litaker, and William G Sunda. Phosphorus physiological ecology and molecular mechanisms in marine phytoplankton. *Journal of Phycology*, 52(1):10–36, 2016.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Hubert Loisel and André Morel. Light scattering and chlorophyll concentration in case 1 waters: A reexamination. *Limnology and Oceanography*, 43(5):847–858, 1998.

Alan Longhurst, Shubha Sathyendranath, Trevor Platt, and Carla Caverhill. An estimate of global primary production in the ocean from satellite radiometer data. *Journal of plankton Research*, 17(6):1245–1271, 1995.

Alan R Longhurst. Interactions between zooplankton and phytoplankton profiles in the eastern tropical Pacific Ocean. In *Deep Sea Research and Oceanographic Abstracts*, volume 23, pages 729–754. Elsevier, 1976.

Carl J Lorenzen. A method for the continuous measurement of in vivo chlorophyll concentration. In *Deep Sea Research and Oceanographic Abstracts*, volume 13, pages 223–227. Elsevier, 1966.

Amala Mahadevan. The impact of submesoscale physics on primary productivity of plankton. *Annual review of marine science*, 8:161–184, 2016.

Johannie Martin, Jean-Éric Tremblay, Jonathan Gagnon, Geneviève Tremblay, Amandine Lapoussière, Caroline Jose, Michel Poulin, Michel Gosselin, Yves Gratton, and Christine Michel. Prevalence, structure and properties of subsurface chlorophyll maxima in Canadian Arctic waters. *Marine Ecology Progress Series*, 412: 69–84, 2010.

Elodie Martinez, David Antoine, Fabrizio d'Ortenzio, and Clément de Boyer Montégut. Phytoplankton spring and fall blooms in the North Atlantic in the 1980s and 2000s. *Journal of Geophysical Research: Oceans*, 116(C11), 2011.

Yoshio Masuda, Yasuhiro Yamanaka, Sherwood Lan Smith, Takafumi Hirata, Hideyuki Nakano, Akira Oka, and Hiroshi Sumata. Photoacclimation by phytoplankton determines the distribution of global subsurface chlorophyll maxima in the ocean. *Communications Earth & Environment*, 2(1):1–8, 2021.

Jorge Mateu and Ramon Giraldo. *Geostatistical functional data analysis*. John Wiley & Sons, 2021.

George I Matsumoto, Kenneth S Johnson, Steve Riser, Lynne Talley, Susan Wijffels, and Roberta Hotinski. The Global Ocean Biogeochemistry (GO-BGC) Array of

Profiling Floats to Observe Changing Ocean Chemistry and Biology. *Marine Technology Society Journal*, 56(3):122–123, 2022.

MR Mazloff, BD Cornuelle, ST Gille, and A Verdy. Correlation lengths for estimating the large-scale carbon and heat content of the Southern Ocean. *Journal of Geophysical Research: Oceans*, 123(2):883–901, 2018.

Dennis J McGillicuddy Jr. Mechanisms of physical-biological-biogeochemical interaction at the oceanic mesoscale. *Annual Review of Marine Science*, 8:125–159, 2016.

Darren C McKee, Scott C Doney, Alice Della Penna, Emmanuel S Boss, Peter Gaube, and Michael J Behrenfeld. Biophysical dynamics at ocean fronts revealed by Bio-Argo floats. *Journal of Geophysical Research: Oceans*, page e2022JC019226, 2023.

Darren Craig McKee, Scott C Doney, Alice Della Penna, Emmanuel S Boss, Peter Gaube, Michael J Behrenfeld, and David M Glover. Lagrangian-Eulerian statistics of mesoscale ocean chlorophyll from Bio-argo floats and satellites. *Biogeosciences Discussions*, pages 1–36, 2022.

Lingsheng Meng, Chi Yan, Wei Zhuang, Weiwei Zhang, and Xiao-Hai Yan. Reconstruction of three-dimensional temperature and salinity fields from satellite observations. *Journal of Geophysical Research: Oceans*, 126(11):e2021JC017605, 2021.

A Mignot, Hervé Claustre, F d'Ortenzio, X Xing, A Poteau, and J Ras. From the shape of the vertical profile of in vivo fluorescence to Chlorophyll-a concentration. *Biogeosciences*, 8(8):2391–2406, 2011.

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.

Alexandre Mignot, Hervé Claustre, Gianpiero Cossarini, Fabrizio d'Ortenzio, Elodie Gutknecht, Julien Lamouroux, Paolo Lazzari, Coralie Perruche, Stefano Salon, Raphaelle Sauzède, et al. Defining BGC-Argo-based metrics of ocean health and biogeochemical functioning for the evaluation of global ocean models. *Biogeosciences Discussions*, pages 1–66, 2021.

Harvey J Miller. Tobler's first law and spatial analysis. *Annals of the association of American geographers*, 94(2):284–289, 2004.

Isabelle Mirouze, Edward W Blockley, Daniel J Lea, Matthew J Martin, and Michael J Bell. A multiple length scale correlation operator for ocean data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, 68(1):29744, 2016.

Marta Estrada Miyares, Mikel Latasa, Ana M Cabello, Patricia de la Fuente, Carles Guallar, Patricija Mozetič, Max Riera-Lorente, Montserrat Vidal, and Dolors Blasco. Relationships between the deep chlorophyll maximum and hydrographic characteristics across the Atlantic, Indian and Pacific oceans. *Scientia Marina*, 88(4): e092–e092, 2024.

Holly V Moeller, Charlotte Laufkötter, Edward M Sweeney, and Matthew D Johnson. Light-dependent grazing can drive formation and deepening of deep chlorophyll maxima. *Nature communications*, 10(1):1–8, 2019.

CM Moore, MM Mills, KR Arrigo, I Berman-Frank, L Bopp, PW Boyd, ED Galbraith, RJ Geider, C Guieu, SL Jaccard, et al. Processes and patterns of oceanic nutrient limitation. *Nature geoscience*, 6(9):701–710, 2013.

André Morel and Jean-François Berthon. Surface pigments, algal biomass profiles, and potential production of the euphotic layer: Relationships reinvestigated in view of remote-sensing applications. *Limnology and oceanography*, 34(8):1545–1562, 1989.

André Morel and Bernard Gentili. Diffuse reflectance of oceanic waters: its dependence on Sun angle as influenced by the molecular scattering contribution. *Applied optics*, 30(30):4427–4438, 1991.

André Morel and Stéphane Maritorena. Bio-optical properties of oceanic waters: A reappraisal. *Journal of Geophysical Research: Oceans*, 106(C4):7163–7180, 2001.

Mauricio Muñoz-Anderson, Roberto Millán-Núñez, Rafael Hernández-Walls, Adriana González-Silvera, Eduardo Santamaría-del Ángel, Evaristo Rojas-Mayoral, and Salvador Galindo-Bect. Fitting vertical chlorophyll profiles in the California Current using two Gaussian curves. *Limnology and Oceanography: Methods*, 13(8):416–424, 2015.

W Brechner Owens, Nathalie Zilberman, Ken S Johnson, Hervé Claustre, Megan Scanderbeg, Susan Wijffels, and Toshio Suga. OneArgo: A New Paradigm for Observing the Global Ocean. *Marine Technology Society Journal*, 56(3), 2022.

P Parekh, S Dutkiewicz, MJ Follows, and T Ito. Atmospheric carbon dioxide in a less dusty world. *Geophysical research letters*, 33(3), 2006.

Mike Pereira, Nicolas Desassis, and Denis Allard. Geostatistics for large datasets on Riemannian manifolds: a matrix-free approach. *arXiv preprint arXiv:2208.12501*, 2022.

Flavien Petit, Julia Uitz, Catherine Schmechtig, Céline Dimier, Joséphine Ras, Antoine Poteau, Melek Golbol, Vincenzo Vellucci, and Hervé Claustre. Influence of the phytoplankton community composition on the in situ fluorescence signal: Implication for an improved estimation of the chlorophyll-a concentration from BioGeoChemical-Argo profiling floats. *Frontiers in Marine Science*, 9:959131, 2022.

Alberto Pilati and Wayne A Wurtsbaugh. Importance of zooplankton for the persistence of a deep chlorophyll layer: a limnocorral experiment. *Limnology and Oceanography*, 48(1):249–260, 2003.

Channing J Prend, Madhavan Girijakumari Keerthi, Marina Lévy, Olivier Aumont, Sarah T Gille, and Lynne D Talley. Sub-Seasonal Forcing Drives Year-To-Year Variations of Southern Ocean Primary Productivity. *Global Biogeochemical Cycles*, 36 (7):e2022GB007329, 2022.

AE Friederike Prowe, Markus Pahlow, Stephanie Dutkiewicz, Michael Follows, and Andreas Oschlies. Top-down control of marine phytoplankton diversity in a global ecosystem model. *Progress in Oceanography*, 101(1):1–13, 2012.

Graham D Quartly, Jim Aiken, Robert JW Brewin, and Andrew Yool. The link between surface and sub-surface chlorophyll-a in the centre of the Atlantic subtropical gyres: a comparison of observations and models. *Frontiers in Marine Science*, 10:1197753, 2023.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

James Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, 2005.

James O Ramsay and CJ1125714 Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 53(3):539–561, 1991.

Pannimpullath Remanan Renosh, Jie Zhang, Raphaëlle Sauzède, and Hervé Claustre. Vertically Resolved Global Ocean Light Models Using Machine Learning. *Remote Sensing*, 15(24):5663, 2023.

Gabriel Reygondeau, Alan Longhurst, Elodie Martinez, Gregory Beaugrand, David Antoine, and Olivier Maury. Dynamic biogeochemical provinces in the global ocean. *Global Biogeochemical Cycles*, 27(4):1046–1058, 2013.

Katherine Richardson and Jørgen Bendtsen. Vertical distribution of phytoplankton and primary production in relation to nutricline depth in the open ocean. *Marine Ecology Progress Series*, 620:33–46, 2019.

Gordon Arthur Riley. Quantitative ecology of the plankton of the western North Atlantic. *Bull. Bingham Oceanogr. Collection*, 12:1–169, 1949.

JOSIE Robinson, EE Popova, MA Srokosz, and A Yool. A tale of three islands: Downstream natural iron fertilization in the Southern Ocean. *Journal of Geophysical Research: Oceans*, 121(5):3350–3371, 2016.

C Carl Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64 (1):38–40, 1957.

Susana Rodríguez-Gálvez, Diego Macías, Laura Prieto, and Javier Ruiz. Top-down and bottom-up control of phytoplankton in a mid-latitude continental shelf ecosystem. *Progress in Oceanography*, 217:103083, 2023.

Dean Roemmich and John Gilson. The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo program. *Progress in oceanography*, 82(2):81–100, 2009.

Collin Roesler, Julia Uitz, Hervé Claustre, Emmanuel Boss, Xiaogang Xing, Emanuele Organelli, Nathan Briggs, Annick Bricaud, Catherine Schmechtig, Antoine Poteau, et al. Recommendations for obtaining unbiased chlorophyll estimates from in situ chlorophyll fluorometers: A global analysis of WET Labs ECO sensors. *Limnology and Oceanography: Methods*, 15(6):572–585, 2017.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

John H Ryther. Photosynthesis in the Ocean as a Function of Light Intensity 1. *Limnology and Oceanography*, 1(1):61–70, 1956.

Sujit K Sahu. *Bayesian modeling of spatio-temporal data with R*. Chapman and Hall/CRC, 2022.

Sujit K Sahu and Peter Challenor. A space-time model for joint modeling of ocean temperature and salinity levels as measured by Argo floats. *Environmetrics: The official journal of the International Environmetrics Society*, 19(5):509–528, 2008.

Jorge L Sarmiento, R Slater, R Barber, L Bopp, SC Doney, AC Hirst, J Kleypas, R Matear, Uwe Mikolajewicz, P Monfray, et al. Response of ocean ecosystems to climate warming. *Global Biogeochemical Cycles*, 18(3), 2004.

Jorge L Sarmiento, Kenneth S Johnson, Lionel A Arteaga, Seth M Bushinsky, Heidi M Cullen, Alison R Gray, Roberta M Hotinski, Tanya L Maurer, Matthew R Mazloff, Stephen C Riser, et al. The Southern Ocean carbon and climate observations and modeling (SOCCOM) project: A review. *Progress in Oceanography*, page 103130, 2023.

Mitsuhide Sato, Takuhei Shiozaki, Fuminori Hashihama, Taketoshi Kodama, Hiroshi Ogawa, Hiroaki Saito, Atsushi Tsuda, Shigenobu Takeda, and Ken Furuya. Relative depths of the subsurface peaks of phytoplankton abundance conserved over ocean provinces. *Limnology and Oceanography*, 2022.

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.

Raphaëlle Sauzède, Hervé Claustre, C Jamet, Julia Uitz, Josephine Ras, A Mignot, and F d'Ortenzio. Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications. *Journal of Geophysical Research: Oceans*, 120(1):451–470, 2015.

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.

Catherine Schmechtig, Herve Claustre, Antoine Poteau, Fabrizio D'Ortenzio, Christina Schallenberg, Thomas Trull, and Xiaogang Xing. Biogeochemical-Argo quality control manual for chlorophyll-a concentration and chl-fluorescence, Version 3.0. 2023.

Katrin Schroeder, Toste Tanhua, Jacopo Chiggiato, Dimitris Velaoras, Simon A Josey, Jesús García Lafuente, and Manuel Vargas-Yáñez. The forcings of the Mediterranean Sea and the physical properties of its water masses. In *Oceanography of the Mediterranean Sea*, pages 93–123. Elsevier, 2023.

John McN Sieburth, Victor Smetacek, and Jürgen Lenz. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions. *Limnology and oceanography*, 23(6):1256–1263, 1978.

David A Siegel, Timothy DeVries, Ivona Cetinić, and Kelsey M Bisson. Quantifying the ocean's biological pump and its carbon cycle impacts on global scales. *Annual review of marine science*, 15(1):329–356, 2023.

Greg M Silsbe and Sairah Y Malkin. Where light and nutrients collide: The global distribution and activity of subsurface chlorophyll maximum layers. *Aquatic microbial ecology and biogeochemistry: A dual perspective*, pages 141–152, 2016.

Shawn R Smith, Gaël Alory, Axel Andersson, William Asher, Alex Baker, David I Berry, Kyla Drushka, Darin Figurskey, Eric Freeman, Paul Holthus, et al. Ship-based contributions to global ocean, weather, and climate observing systems. *Frontiers in Marine Science*, 6:434, 2019.

Serguei Sokolov and Stephen R Rintoul. On the relationship between fronts of the Antarctic Circumpolar Current and surface chlorophyll concentrations in the Southern Ocean. *Journal of Geophysical Research: Oceans*, 112(C7), 2007.

Tao Song, Wei Wei, Fan Meng, Jiarong Wang, Runsheng Han, and Danya Xu. Inversion of ocean subsurface temperature and salinity fields based on spatio-temporal correlation. *Remote Sensing*, 14(11):2587, 2022.

JH Steele. A study of production in the Gulf of Mexico. *Journal of Marine Research*, 1964.

JH Steele and CS Yentsch. The vertical distribution of chlorophyll. *Journal of the Marine Biological Association of the United Kingdom*, 39(2):217–226, 1960.

Michael L Stein. Some statistical issues in climate science. *Statistical Science*, 2020.

Deborah K Steinberg and Michael R Landry. Zooplankton and the ocean carbon cycle. *Annual review of marine science*, 9(1):413–444, 2017.

Andrea Storto, Paolo Oddo, Andrea Cipollone, Isabelle Mirouze, and Benedicte Lemieux-Dudon. Extending an oceanographic variational scheme to allow for affordable hybrid and four-dimensional data assimilation. *Ocean Modelling*, 128: 67–86, 2018.

John DH Strickland. A comparison of profiles of nutrient and chlorophyll concentrations taken from discrete depths and by continuous recording. *Limnology and Oceanography*, 13(2):388–391, 1968.

Peter G Strutton, Thomas W Trull, Helen E Phillips, Earl R Duran, and Sylvia Pump. Biogeochemical Argo floats reveal the evolution of subsurface chlorophyll and particulate organic carbon in southeast Indian Ocean eddies. *Journal of Geophysical Research: Oceans*, page e2022JC018984, 2023.

Jiaoyang Su, Peter G Strutton, and Christina Schallenberg. The subsurface biological structure of Southern Ocean eddies revealed by BGC-Argo floats. *Journal of Marine Systems*, 220:103569, 2021.

Jing Tan, Robert Frouin, Dominique Jolivet, Mathieu Compiègne, and Didier Ramon. Evaluation of the NASA OBPG MERIS ocean surface PAR product in clear sky conditions. *Optics Express*, 28(22):33157–33175, 2020.

Weiyi Tang, Joan Llort, Jakob Weis, Morgane MG Perron, Sara Basart, Zuchuan Li, Shubha Sathyendranath, Thomas Jackson, Estrella Sanz Rodriguez, Bernadette C Proemse, et al. Widespread phytoplankton blooms triggered by 2019–2020 Australian wildfires. *Nature*, 597(7876):370–375, 2021.

Pierre Testor, Brad De Young, Daniel L Rudnick, Scott Glenn, Daniel Hayes, Craig M Lee, Charitha Pattiaratchi, Katherine Hill, Emma Heslop, Victor Turpin, et al. OceanGliders: a component of the integrated GOOS. *Frontiers in Marine Science*, 6: 422, 2019.

Virginie Thierry, Hervé Claustre, Orens Pasqueron de Fommervault, Nathalie Zilberman, Kenneth S Johnson, Brian A King, Susan E Wijffels, Udaya TVS Bhaskar, Magdalena Alonso Balmaseda, Mathieu Belbeoch, et al. Advancing ocean monitoring and knowledge for societal benefit: the urgency to expand Argo to OneArgo by 2030. *Frontiers in Marine Science*, 12:1593904, 2025.

Sandy J Thomalla, A Gilbert Ogunkoya, Marcello Vichi, and Sebastiaan Swart. Using optical sensors on gliders to estimate phytoplankton carbon concentrations and chlorophyll-to-carbon ratios in the Southern Ocean. *Frontiers in Marine Science*, 4:34, 2017.

Jerry F Tjiputra, Damien Couespel, and Richard Sanders. Marine ecosystem role in setting up preindustrial and future climate. *Nature Communications*, 16(1):2206, 2025.

Matteo Tomasetto, Eleonora Arnone, and Laura M Sangalli. Modeling Anisotropy and Non-Stationarity Through Physics-Informed Spatial Regression. *Environmetrics*, 35 (8):e2889, 2024.

SC Tripathy, Sini Pavithran, Prabhakaran Sabu, Honey UK Pillai, Dipti RG Dessai, and Narayanapillai Anilkumar. Deep chlorophyll maximum and primary productivity in Indian Ocean sector of the Southern Ocean: Case study in the Subtropical and Polar Front during austral summer 2011. *Deep Sea Research Part II: Topical Studies in Oceanography*, 118:240–249, 2015.

Takaya Uchida, Dhruv Balwada, Ryan Abernathey, Channing J Prend, Emmanuel Boss, and Sarah T Gille. Southern Ocean phytoplankton blooms observed by biogeochemical floats. *Journal of Geophysical Research: Oceans*, 124(11):7328–7343, 2019.

Julia Uitz, Hervé Claustre, André Morel, and Stanford B Hooker. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *Journal of Geophysical Research: Oceans*, 111(C8), 2006.

Cristhian Leonardo Urbano-Leon, Manuel Escabias, Diana Paola Ovalle-Muñoz, and Javier Olaya-Ochoa. Scalar Variance and Scalar Correlation for Functional Data. *Mathematics*, 11(6):1317, 2023.

Sergio M Vallina, Michael J Follows, Stephanie Dutkiewicz, José M Montoya, Pedro Cermeño, and Michel Loreau. Global relationship between phytoplankton diversity and productivity in the ocean. *Nature communications*, 5(1):4299, 2014.

Ramiro A Varela, Antonio Cruzado, and Joaquín Tintoré. Modelling the deep-chlorophyll maximum: A coupled physical-biological approach. *Journal of Marine Research*, 50:441–463, 1992.

Hugh Venables and C Mark Moore. Phytoplankton and light limitation in the Southern Ocean: Learning from high-nutrient, high-chlorophyll areas. *Journal of Geophysical Research: Oceans*, 115(C2), 2010.

Clara R Vives, Christina Schallenberg, Peter G Strutton, Jørgen Bendtsen, Katherine Richardson, and Philip W Boyd. An inter-comparison of Deep Chlorophyll Maxima characteristics from 30S to 74S and their contribution to Net Primary Production. 2024a.

Clara R Vives, Christina Schallenberg, Peter G Strutton, and Philip W Boyd. Biogeochemical-argo floats show that chlorophyll increases before carbon in the high-latitude southern ocean spring bloom. *Limnology and Oceanography Letters*, 9(3): 172–182, 2024b.

Changjie Wang and Fenfen Liu. Influence of oceanic mesoscale eddies on the deep chlorophyll maxima. *Science of The Total Environment*, 917:170510, 2024.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295, 2016.

Shanlin Wang, D Bailey, Keith Lindsay, JK Moore, and Marika Holland. Impact of sea ice on the marine iron cycle and phytoplankton productivity. *Biogeosciences*, 11(17): 4713–4731, 2014.

Yan Wang, Jie Yang, and Ge Chen. Euphotic Zone Depth Anomaly in Global Mesoscale Eddies by Multi-Mission Fusion Data. *Remote Sensing*, 15(4):1062, 2023.

Ben A Ward, Stephanie Dutkiewicz, and Michael J Follows. Modelling spatial and temporal patterns in size-structured marine plankton communities: top–down and bottom–up controls. *Journal of Plankton Research*, 36(1):31–47, 2014.

Jakob Weis, Christina Schallenberg, Zanna Chase, Andrew R Bowie, Bozena Wojtasiewicz, Morgane MG Perron, Marc D Mallet, and Peter G Strutton. Southern Ocean phytoplankton stimulated by wildfire emissions and sustained by iron recycling. *Geophysical Research Letters*, 49(11):e2021GL097538, 2022.

Christopher Whitt, Jay Pearlman, Brian Polagye, Frank Caimi, Frank Muller-Karger, Andrea Copping, Heather Spence, Shyam Madhusudhana, William Kirkwood, Ludovic Grosjean, et al. Future vision for autonomous ocean observations. *Frontiers in Marine Science*, 7:697, 2020.

Juliane U Wihsgott, Jonathan Sharples, Joanne E Hopkins, E Malcolm S Woodward, Tom Hull, Naomi Greenwood, and David B Sivyer. Observations of vertical mixing in autumn and its effect on the autumn phytoplankton bloom. *Progress in Oceanography*, 177:102059, 2019.

Franziska M Willems, Johannes Fredericus Scheepens, and Oliver Bossdorf. Forest wildflowers bloom earlier as Europe warms: lessons from herbaria and spatial modelling. *New Phytologist*, 235(1):52–65, 2022.

Kai Wirtz, S Lan Smith, Moritz Mathis, and Jan Taucher. Vertically migrating phytoplankton fuel high oceanic primary production. *Nature Climate Change*, 12(8): 750–756, 2022.

Bożena Wojtasiewicz, Thomas W Trull, TVS Udaya Bhaskar, Mangesh Gauns, Satya Prakash, M Ravichandran, Damodar M Shenoy, Dirk Slawinski, and Nick J Hardman-Mountford. Autonomous profiling float observations reveal the dynamics of deep biomass distributions in the denitrifying oxygen minimum zone of the Arabian Sea. *Journal of Marine Systems*, 207:103103, 2020.

Annie Wong, Robert Keeley, Thierry Carval, and the Argo Data Management Team. *Argo Quality Control Manual for CTD and Trajectory Data*, 2020.

Annie Wong, Robert Keeley, and Thierry Carval. Argo quality control manual for CTD and trajectory data. 2025.

Annie PS Wong and Stephen C Riser. Profiling float observations of the upper ocean under sea ice off the Wilkes Land coast of Antarctica. *Journal of Physical Oceanography*, 41(6):1102–1115, 2011.

Jinghui Wu, Zhongping Lee, Yuyuan Xie, Joaquim Goes, Shaoling Shang, John F Marra, Gong Lin, Lei Yang, and Bangqin Huang. Reconciling between optical and biological determinants of the euphotic zone depth. *Journal of Geophysical Research: Oceans*, 126(5):e2020JC016874, 2021.

Xiaogang Xing, Hervé Claustre, Stéphane Blain, Fabrizio d'Ortenzio, David Antoine, Josephine Ras, and Christophe Guinet. Quenching correction for in vivo chlorophyll fluorescence acquired by autonomous platforms: A case study with instrumented elephant seals in the Kerguelen region (Southern Ocean). *Limnology and Oceanography: Methods*, 10(7):483–495, 2012.

Xiaogang Xing, Hervé Claustre, Emmanuel Boss, Collin Roesler, Emanuele Organelli, Antoine Poteau, Marie Barbieux, and Fabrizio d'Ortenzio. Correction of profiles of in-situ chlorophyll fluorometry for the contribution of fluorescence originating from non-algal matter. *Limnology and Oceanography: Methods*, 15(1):80–93, 2017.

Xiaogang Xing, Nathan Briggs, Emmanuel Boss, and Hervé Claustre. Improved correction for non-photochemical quenching of in situ chlorophyll fluorescence based on a synchronous irradiance profile. *Optics express*, 26(19):24734–24751, 2018.

Xiaogang Xing, Peng Xiu, Edward A Laws, Guo Yang, Xin Liu, and Fei Chai. Light-Driven and Nutrient-Driven Displacements of Subsurface Chlorophyll

Maximum Depth in Subtropical Gyres. *Geophysical Research Letters*, 50(22): e2023GL104510, 2023.

Dan Xu, Tao Wang, Xiaogang Xing, and Changwei Bian. The relationship between Nitrate and Potential Density in the Ocean South of 30° S. *Journal of Geophysical Research: Oceans*, page e2022JC018948, 2022a.

Wenlong Xu, Guifen Wang, Xuhua Cheng, Long Jiang, Wen Zhou, and Wenxi Cao. Characteristics of subsurface chlorophyll maxima during the boreal summer in the South China Sea with respect to environmental properties. *Science of The Total Environment*, 820:153243, 2022b.

Drew Yarger, Stilian Stoev, and Tailen Hsing. A functional-data approach to the Argo data. *The Annals of Applied Statistics*, 16(1):216–246, 2022.

Sayaka Yasunaka, Tsuneo Ono, Kosei Sasaoka, and Kanako Sato. Global distribution and variability of subsurface chlorophyll a concentration. *Ocean Science Discussions*, 2021:1–22, 2021.

Yongjun Yu, Baoxiang Huang, Milena Radenkovic, Tingting Wang, and Ge Chen. Intelligent Sparse2Dense Profile Reconstruction for Predicting Global Subsurface Chlorophyll Maxima. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Arianna Zampollo, Thomas Cornulier, Rory O'Hara Murray, Jacqueline Fiona Tweddle, James Dunning, and Beth E Scott. The bottom mixed layer depth as an indicator of subsurface Chlorophyll a distribution. *Biogeosciences*, 20(16):3593–3611, 2023.

Mimi Zhang and Andrew Parnell. Review of clustering methods for functional data. *ACM Transactions on Knowledge Discovery from Data*, 17(7):1–34, 2023.